Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche


# RIMOZIONE DELLA COMPONENTE DI VARIABILITÀ CAUSATA DA FATTORI NON BIOLOGICI: UN CASO STUDIO SU DATI DI TUMORE ALL'OVAIO

Relatore: Prof.ssa Chiara Romualdi
Dipartimento di Biologia

Laureando: Marco Pegoraro
Matricola N. 1013326

Anno Accademico 2013/2014

# Indice

# Capitolo 1

## 1. Introduzione

Questa tesi si pone un duplice obiettivo: il primo è lo studio di un fenomeno particolarmente noto nell'ambiente della biostatistica e definito "batch effect", il secondo è lo studio dei profili di espressione di campioni biologici su delle pazienti affette da tumore all'ovaio. Il batch effect consiste nell'introduzione di variabilità non biologica in un esperimento, la quale comporta una non confrontabilità tra i campioni di "batch" diversi. I campioni che si vogliono utilizzare non sono infatti stati tutti ricavati impiegando un'unica piattaforma, ma derivano da due piattaforme Affymetrix differenti. E' proprio questa una delle possibili cause del batch effect e, dunque, i dati dovranno necessariamente essere preprocessati per stimare ed eliminare la parte di variabilità non legata ai fattori biologici di interesse e, dunque, non voluta.

Prima di iniziare la trattazione del fenomeno del batch effect è opportuno dedicare qualche paragrafo all'illustrazione di alcuni concetti fondamentali per la comprensione della tesi per chi non conosce l'ambito della biostatistica e della bioinformatica.

## 1.1. Bioinformatica e biostatistica

La bioinformatica è una disciplina nata alla fine degli anni '70 grazie allo sviluppo delle prime tecniche per il sequenziamento degli acidi nucleici. E' evidente come il termine derivi dall'unione delle due parole "biologia" e "informatica" e questo perché con lo sviluppo delle suddette tecniche nasceva l'esigenza di avere un supporto informatico alla biologia. Con il passare del tempo la quantità di dati prodotti nel campo della biologia molecolare, soprattutto nei progetti di sequenziamento di interi genomi, è aumentata a dismisura rendendo la

bioinformatica una materia a sé stante. Si può affermare, quindi, che la bioinformatica è l'uso dell'informatica e della statistica per la gestione e l'analisi dei dati biologici. E' evidente come la figura del bioinformatico debba quindi avere competenze interdisciplinari: biologia molecolare, informatica, matematica e statistica sono di fondamentale importanza.

## 1.2. Introduzione alla biologia molecolare

La biologia molecolare è un ramo della biologia che studia gli esseri viventi a livello molecolare, in particolare nelle relazioni tra le loro macromolecole, ovvero proteine e acidi nucleici (DNA, RNA).

DNA è l'acronimo per acido desossiribonucleico; è un polimero costituito da catene di basi, le quali possono essere Adenina (A), Citosina (C), Guanina (G) e Timina (T). Ogni base è parte di una molecola più grande, detta nucleotide.

Il DNA è una doppia elica destrorsa in cui ogni elica è fatta da una successione di nucleotidi aventi tra loro dei legami, detti fosfodiesterici, e le due eliche si uniscono grazie a legami, detti a idrogeno, che si formano tra le coppie di basi. Questi ultimi legami si formano solo tra A e T e tra C e G. Le eliche di DNA hanno un orientamento, dal 5' al 3', facendo riferimento agli atomi di carbonio del desossiribosio.

Il DNA è contenuto nel nucleo di ogni cellula e al suo interno sono codificate le informazioni per tutte le funzioni che devono essere svolte dalla cellula stessa, ad esempio quelle per la codifica delle proteine. Oltre a queste regioni ce ne sono altre nelle quali sono contenute le funzioni di regolazione della cellula.

Le regioni del DNA codificanti per le proteine sono dette geni. La fase di copiatura viene chiamata trascrizione ed il suo prodotto è un filamento di RNA messaggero (mRNA) che, sostanzialmente, contiene una sequenza di basi. Un gene, però, non contiene solamente parti effettivamente utili per la codifica delle proteine, ma contiene anche altre regioni. Si è soliti distinguere, infatti, tra esoni, introni e UTRs. Le regioni effettivamente utili per la codifica delle proteine sono gli esoni, per cui una fase importante è quella dello splicing, durante la quale vengono mantenuti solo gli esoni. Al termine di questa procedura il frammento di

mRNA viene detto maturo. Il passo successivo per la codifica delle proteine è la traduzione, nella quale l'mRNA viene interpretato secondo il principio per cui tre basi (codone) codificano per uno specifico amminoacido. In questa fase l'mRNA esce dal nucleo della cellula e si unisce ai ribosomi per creare la catena di amminoacidi che forma la proteina.

## 1.3. Lo studio dell'espressione genica

Per studio dell'espressione genica si intende la determinazione della quantità di RNA messaggero trascritto dalle cellule di un organismo in una certa condizione sperimentale. L'utilità di ciò sta nella possibilità di effettuare un confronto tra il livello di espressione genica in tipi cellulari diversi o in condizioni patologiche diverse, per determinare il ruolo che i geni hanno in queste. Si definiscono geni differenzialmente espressi (DEG) quei geni che in due condizioni differenti (es. tessuti sani e tessuti cancerogeni) hanno un livello di espressione significativamente diverso. Rispetto ad una situazione di riferimento, i DEG possono dunque essere sovra o sotto espressi. Naturalmente la significatività deve essere verificata con l'impiego di opportune tecniche statistiche e ciò apre la strada ad una moltitudine di problemi che sorgono dall'utilizzo di dati di questo tipo.

Esistono due differenti tecnologie che permettono di svolgere studi di espressione genica a livello globale: il primo approccio è basato sull'ibridizzazione e la corrispettiva tecnologia è detta "microarray"; il secondo è basato sul sequenziamento ed in questo caso la tecnologia è definita "RNA-seq". I dati che si utilizzeranno in questa tesi sono di microarray.

# Capitolo 2

## 2. DNA microarray

Una tecnologia che permette di effettuare studi di espressione genica è la tecnologia DNA microarray. Esistono due tipi di DNA microarray:

1. a singolo canale, i quali permettono di rilevare l'espressione assoluta di migliaia di geni in una certa condizione (es. sano o malato);
2. a doppio canale, i quali permettono di confrontare migliaia di geni in due condizioni diverse (es. sano vs malato) studiando l'espressione relativa, ossia il rapporto tra le espressioni nelle due diverse condizioni.

Per poter spiegare cosa sono i DNA microarray è necessario fare una puntualizzazione. Si consideri il caso di un microarray a singolo canale: lo scopo è quello di misurare contemporaneamente la quantità di trascritti relativi a migliaia di geni in una certa condizione, ad esempio in un certo tessuto. Per quanto detto in precedenza ciò che adeguatamente misura i trascritti è l'mRNA ma in realtà ciò che viene utilizzato è una copia inversa dell'mRNA, detta DNA complementare (cDNA). Il motivo di ciò è che, mentre l'RNA è particolarmente instabile, il cDNA ha una maggiore stabilità.

I microarray sono dei vetrini, o altre superfici solide definite anche chip o semplicemente array, ai quali vengono attaccati dei frammenti di DNA a singola elica corrispondenti a diversi geni, in modo che questi fungano da sonda per intercettare i cDNAs sintetizzati a partire da mRNA estratto da un tessuto. Le sonde per ogni gene sono molteplici ed ogni insieme di sonde relativo ad un gene è definito probe. I probe sono disposti come una griglia, in modo che uno scanner possa rilevarne esattamente la posizione. Alla base di questa tecnica sta il fatto che un'elica di DNA si lega in modo univoco alla sua complementare grazie ai legami tra Adenina e Timina e tra Guanina e Citosina.

Per la quantificazione di un valore di intensità nei microarray a singolo canale è necessario estrarre mRNA da un tessuto e convertirlo in cDNA, incorporando nel frammento anche un fluoroforo; a questo punto tutto il cDNA marcato a fluorescenza viene ibridato nel chip: in questa fase ogni frammento di cDNA si appaierà alla sua corrispondente sonda. Dato che ogni spot contiene molte sonde, ce ne saranno alcuni con molto cDNA legato alle sonde (geni molto espressi) e altri con poco cDNA (geni poco espressi). Per ottenere un valore di intensità il chip viene letto da uno scanner ottico, il quale eccita i fluorofori e ottiene delle immagini ad altissima definizione di ogni spot. Di queste viene analizzato ogni singolo pixel, ottenendo dei valori sulla base della gradazione del colore. Viene poi costruita una curva della distribuzione del segnale: la mediana della distribuzione è il valore assegnato al segnale. E' prassi comune scannerizzare le immagini diverse volte e utilizzare tecniche per l'integrazione delle immagini o dei valori ottenuti dalle scansioni in modo da ridurre la variabilità.

Nel caso di microarray a doppio canale il procedimento è del tutto simile; la differenza sta nel fatto che il materiale biologico è estratto da due tessuti in condizioni diverse e, per ogni spot, la misura di espressione è data dal rapporto tra i valori nelle due condizioni. Naturalmente, proprio perché ci sono tessuti derivanti da due diverse condizioni, si devono utilizzare due fluorofori.

Come si avrà modo di approfondire più avanti, ogni singolo esperimento di microarray è influenzato da molteplici fattori esterni e non riguardanti il fattore biologico di interesse. Nel singolo canale, quando si vanno a confrontare valori di espressione derivanti da diversi esperimenti, ci possono essere delle differenze nelle intensità, ad esempio a causa di parametri diversi nella scansione. Ciò implica la necessità di correggere i valori tra gli array per renderli direttamente confrontabili. La tecnologia a doppio canale permette di ridurre l'impatto di fattori esterni perché confronta le due condizioni in uno stesso esperimento; d'altra parte, però, è noto che i due fluorofori utilizzati hanno efficienze diverse e, dunque, i valori ricavati devono essere comunque in qualche modo aggiustati. Il processo di aggiustamento dei valori è detto *normalizzazione* ed è trattato nel §2.3.

In definitiva, uno studio effettuato con microarray a doppio canale permette di dimezzare il numero di campioni necessari rispetto al singolo canale e, dunque,

risulta essere meno costoso. D'altra parte il singolo canale ha maggiore riproducibilità, permette di ottenere valori assoluti e di effettuare confronti tra più di due condizioni; inoltre, come sarà discusso nel terzo capitolo, si stanno studiando metodi che permettano di correggere variazioni sistematiche tra array ricavati da studi diversi e rendere dunque confrontabili esperimenti di *batch* diversi. Per tutti questi motivi il singolo canale sta avendo un'espansione sempre più grande. Visti gli obiettivi di questa tesi, d'ora in avanti si farà riferimento solamente alla tecnologia a singolo canale.

## 2.1. Il chip Affymetrix

Dal punto di vista della manifattura, esistono diverse tecnologie per la produzione dei microarray. Una tra le più note è quella sfruttata dai chip Affymetrix. Questo chip è un supporto di vetro con un'area di 1,28 $cm^2$ e, grazie ad una speciale tecnica detta fotolitografia, la sintesi degli oligonucleotidi sonda avviene direttamente nel chip, rendendo possibile la creazione di sonde per migliaia di geni contemporaneamente in uno spazio ridotto. Il supporto è diviso in molte aree, ognuna della dimensione di 0,005 $cm^2$, contenenti circa $40x10^7$ oligonucleotidi identici lunghi 25 basi che costituiscono una parte della sonda per identificare un determinato gene. La sonda completa per un gene è costituita da 20 aree distinte, così ogni gene è identificato da 20 sonde. Inoltre, per evitare l'ibridazione aspecifica, ogni area è affiancata da un'area di controllo; anche queste contengono $40x10^7$ oligonucleotidi sonda, ma questi ultimi differiscono da quelli "veri" per la sola base centrale della sequenza.

Utilizzando termini più tecnici si può definire il chip Affymetrix come un vetrino composto da migliaia di *probeset* (uno per ogni gene) ed ognuno di questi è fatto di 20 *probe pairs*, cioè 20 coppie di *probe cell*. All'interno di queste coppie, un *probe cell* è di *perfect match* (PM) e l'altro è di *mismatch* (MM). Ogni singolo *probe cell* ha al suo interno $40x10^7$ *probe*, ossia sequenze di nucleotidi lunghe 25 basi. La Figura 2.1 mostra il dettaglio del chip Affymetrix.

In realtà il numero di *probe pairs* non è fissato a 20 per tutti i tipi di array Affymetrix, ma con il passare del tempo si è assistito ad un decremento di questo numero e, ad oggi, questo è un numero compreso tra gli 11 e i 20.

Come detto, i MM servono a misurare l'ibridazione aspecifica, ossia danno una misura di quanto, in uno spot, le sonde catturano anche frammenti di cDNA non perfettamente complementari.

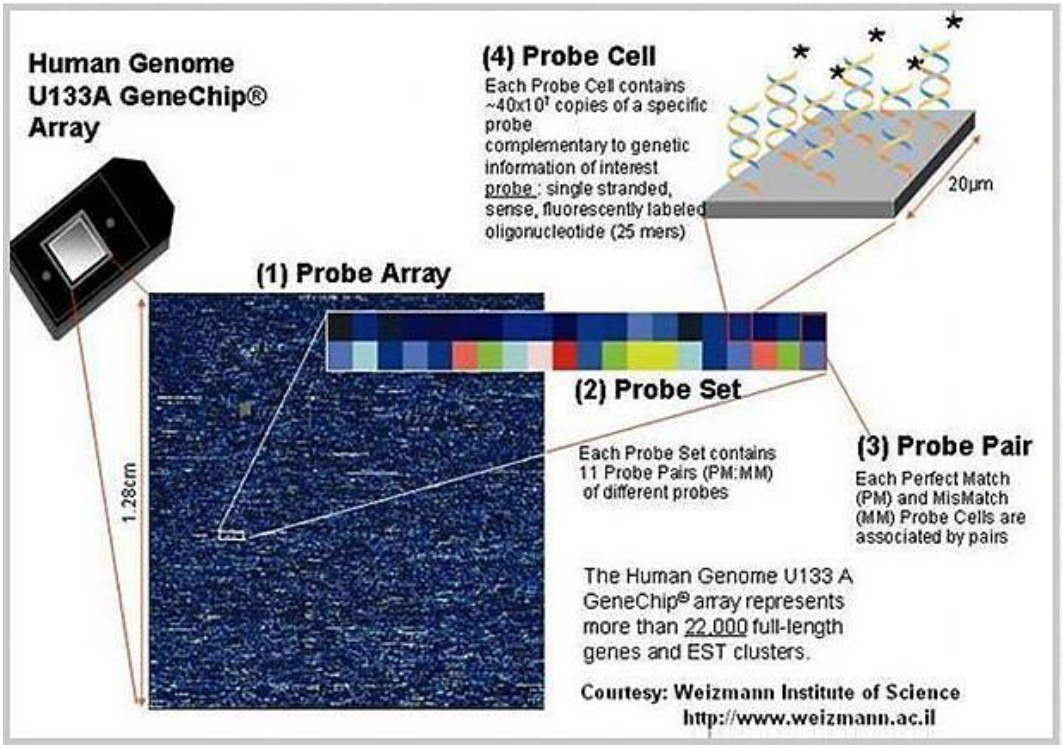



**Figura 2.1: Struttura del chip Affymetrix. Il vetrino è composto da migliaia di probeset, ognuno lungo 20 probe pairs. La singola coppia di probe pairs è composta da un probe cell di perfect match (PM) e da un probe cell di mismatch (MM). Ogni probecell contiene circa $40x10^7$ probe, ossia sequenze di nucleotidi lunghe 25 basi.**

## 2.2. La matrice dei dati

Si ritiene utile dedicare un breve paragrafo alla spiegazione della struttura della matrice dei dati osservati, in quanto dati derivanti da esperimenti di microarray hanno caratteristiche che rendono questo contesto molto differente da altre applicazioni statistiche. Come primo punto, si noti che le unità statistiche sono i chip stessi, dato che sono loro le unità elementari sulle quali si vuole

rilevare un insieme di caratteri. Questi caratteri, o variabili, sono invece i geni, dei quali si vogliono misurare i valori di espressione.

Per quanto detto in precedenza, un chip Affymetrix è in grado di valutare i livelli di espressione di diverse migliaia di geni. Per fare un esempio, uno dei due tipi di chip che verranno utilizzati nell'analisi dei dati di tumore all'ovaio nella seconda parte della tesi è il *GeneChip Human Genome U133A Array*. Questo chip è in grado di sondare i valori di espressione di circa 14.500 geni. Ciò significa, in termini statistici, che si hanno a disposizione circa 14.500 variabili in un solo esperimento. Naturalmente si possono effettuare diversi esperimenti ed ottenere diversi chip, ma si tratta comunque di un processo costoso in termini economici per cui, generalmente, si è costretti a lavorare con una numerosità campionaria pari ad alcune decine di array. La matrice dei dati, dunque, è come quella della Tabella 2.1.

In generale, è possibile identificare ciascuno dei valori di intensità definendolo come $y_{jg}$. Il deponente $j$ sta ad indicare il campione ($j$ = 1,…,$n$), $g$ indica il gene ($g$ = 1,…,$G$).

| | **Condizione 1** | | | **Condizione 2** | | |
|---|---|---|---|---|---|---|
| **Gene 1** | $y_{1,1}$ | … | $y_{1,n_1}$ | $y_{1,n_1+1}$ | … | $y_{1,n}$ |
| **…** | | | | | | |
| **Gene g** | $y_{g,1}$ | … | $y_{g,n_1}$ | $y_{g,n_1+1}$ | … | $y_{g,n}$ |
| **…** | | | | | | |
| **Gene G** | $y_{G,1}$ | … | $y_{G,n_1}$ | $y_{G,n_1+1}$ | … | $y_{G,n}$ |

**Tabella 2.1: Matrice dei dati in caso si disponga di $n_1$ array per la condizione 1 e di $n_2$ =n- $n_1$ array per la condizione 2. Il numero di geni $G$ rappresenta il numero di variabili esplicative, mentre il numero totale di campioni è $n$.**

# 2.3. Normalizzazione

In riferimento a microarray a singolo canale, i valori di intensità ottenuti sono assoluti e devono solitamente essere confrontati con quelli ottenuti in esperimenti diversi. Come accennato all'inizio di questo capitolo, in questo tipo di esperimenti

ci possono essere delle differenze tra i valori di intensità di array diversi dovute a molte cause (Hartemink, et al., 2003). Si tratta di fonti di variabilità introdotte durante la preparazione del campione, la creazione dell'array e la conduzione dell'esperimento (ibridazione, scansione). Cause di ciò possono essere, per esempio, diversi parametri di scansione o diversa potenza del laser utilizzato. In questi casi si rende necessario normalizzare i dati tra gli array, ossia correggere i valori di espressione in modo da eliminare gli effetti causati da errori sistematici negli esperimenti e renderli, dunque, confrontabili.

In realtà esistono anche altre fonti di variabilità, le quali comportano effetti più forti di quelle appena introdotte, ma non sono necessariamente presenti in tutti gli esperimenti di microarray. Per queste la normalizzazione non è sufficiente a rendere confrontabili tra loro gli esperimenti e, dunque, si rende necessario l'uso di tecniche di aggiustamento specifiche. Il terzo capitolo di questo elaborato sarà interamente dedicato alla trattazione di alcune di queste tecniche.

La normalizzazione dei dati è basata su specifiche assunzioni:

- pochi geni differenzialmente espressi: ci si aspetta che il numero di geni differenzialmente espressi tra le due condizioni sia abbastanza basso; solitamente intorno al 10%;
- simmetria tra sovra e sotto espressi: i geni differenzialmente espressi devono essere equamente ripartiti tra sovra e sotto espressi;
- la differenziale espressione non deve dipendere dalla media del segnale.

Prima di passare in rassegna i principali e più utilizzati metodi per normalizzare i valori di espressione tra array è utile introdurre un importante strumento grafico chiamato *grafico MA.* Questo pone in ascissa la media delle log-intensità, indicata con *A* (Amplitude), e in ordinata la differenza delle log-intensità, indicata con *M* (Magnitude). Detti $y_{jg}$ il valore di espressione per il gene $g$ nel campione $j$ e $y_{j'g}$ il valore di espressione per lo stesso gene $g$ nel campione $j'$, si può individuare il grafico MA:

$$\begin{cases} A_{g,(j,j')} = \frac{1}{2}\left[log_2\, y_{jg} + log_2\, y_{j'g}\right] \\ M_{g,(j,j')} = log_2\, y_{jg} - log_2\, y_{j'g} \end{cases} \tag{2.1}$$

dove il pedice $g,(j,j')$ sta ad indicare che il valore è relativo al gene $g$ derivante dal confronto tra il campione $j$ e il campione $j'$.

E' stato dimostrato che il grafico MA ha la capacità di evidenziare differenze sistematiche di espressione lungo i livelli medi di intensità (Dudoit, et al., 2002).

### 2.3.1. Quantile

Una delle più semplici ed utilizzate normalizzazioni è la *Quantile*. L'obiettivo è quello di rendere identiche le distribuzioni empiriche di tutti gli array. Il metodo è motivato dall'idea che, dati due vettori di dati, la loro distribuzione è perfettamente identica solo se il grafico quantile-quantile è esattamente una linea diagonale. Questo concetto può essere esteso a *n* dimensioni, dove *n* è il numero di esperimenti effettuati, per cui tutti gli *n* vettori hanno identica distribuzione solo se il grafico quantile-quantile *n*-dimensionale è esattamente la diagonale dell'ipercubo di lato *n*, ossia il vettore unitario del piano *n*-dimensionale $\left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$. Questo suggerisce la possibilità di rendere uguali le distribuzioni dei dati proiettando i punti del grafico quantile-quantile osservato sulla diagonale.

Detti $q_k = (q_{k1}, \dots, q_{kn})$, per $k = 1,\dots,p$ il vettore dei *k*-esimi quantili di tutti gli *n* array e $d = \left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$ la diagonale unitaria, si può considerare la proiezione di *q* su *d*:

$$proj_d q_k = \left(\frac{1}{n}\sum_{j=1}^{n} q_{kj}, \dots, \frac{1}{n}\sum_{j=1}^{n} q_{kj}\right) \quad (2.2)$$

Per cui si può fare in modo che tutti gli array abbiano la stessa distribuzione prendendo come riferimento il quantile medio e sostituendo con questo i valori nel dataset. In forma di algoritmo, la normalizzazione Quantile prevede:

1. sia $X$ la matrice di dimensione $G$ x $n$ dove ogni esperimento è una colonna e ogni gene è una riga;
2. ordina le colonne di $X$ in modo da ottenere $X_{ord}$;
3. calcola le medie per riga di $X_{ord}$ e assegna questi valori medi ad ogni elemento delle righe in modo da ottenere $X'_{ord}$;
4. ottieni $X_{norm}$ riordinando ogni colonna di $X'_{ord}$ in modo che abbia lo stesso ordine della matrice originale $X$.

## 2.3.2. VSN

La normalizzazione VSN (Variance Stabilization and Normalization) ha, come dice il nome stesso, l'obiettivo di normalizzare i dati stabilizzando la varianza. I dati di microarray hanno infatti la caratteristica di essere eteroschedastici rispetto al livello medio di espressione: a livelli di espressione bassi la varianza è bassa, mentre per livelli alti di espressione la varianza è alta (Huber, et al., 2002). La relazione tra le due quantità è però complessa e gli autori della tecnica hanno derivato una famiglia di trasformazioni delle misure di intensità che renda l'intensità media indipendente dalla varianza.

L'assunzione alla base della normalizzazione è che i dati siano generati secondo un modello del tipo:

$$Y_g = \alpha_g + \mu_g e^{\eta_g} + \varepsilon_g \quad \text{con} \quad \eta_g \sim N\left(0, \sigma_\eta^2\right), \quad \varepsilon_g \sim N(0, \sigma_\varepsilon^2) \tag{2.3}$$

dove $Y_g$ rappresenta il livello di espressione del gene $g$, $\alpha_g$ è un rumore medio di background e $\mu_g$ è il vero valore di espressione del gene. $\eta_g$ ed $\varepsilon_g$ sono termini di errore moltiplicativo e additivo a media zero e varianza costante. Si assume inoltre l'indipendenza dei due termini d'errore.

Si noti che per livelli di espressione bassi ($\mu_g \to 0$) il valore di espressione è approssimabile da:

$$Y_g \approx \alpha_g + \varepsilon_g \qquad \text{per cui} \quad Y_g \approx N\left(\alpha_g, \sigma_\varepsilon^2\right) \tag{2.4}$$

Per livelli alti di espressione ($\mu_g \to \infty$), invece, si ha che:

$$Y_g \approx \mu_g e^{\eta_g} \qquad \text{per cui} \quad Y_g \approx logN\left(log\mu_g, \sigma_\eta^2\right) \tag{2.5}$$

dato che il secondo e il terzo termine sono trascurabili.

Quando il livello di espressione assume valori intermedi, la misura di espressione si distribuisce come una combinazione lineare di una Normale e di una Log-Normale. Si ha dunque che:

$$Var\left(Y_g\right) = Var\left(\alpha_g + \mu_g e^{\eta_g} + \varepsilon_g\right) = \mu_g^2 Var(e^{\eta_g}) + Var\left(\varepsilon_g\right) = \\ = \mu_g^2 \cdot S_\eta^2 + \sigma_\varepsilon^2 \tag{2.6}$$

dove $S_\eta^2 = e^{\sigma_\eta^2}\left(e^{\sigma_\eta^2} - 1\right)$. Per cui è evidente che la varianza dipende dalla media di espressione $\mu_g$. Si cerca dunque una trasformazione dei dati che stabilizzi la varianza asintotica. E' stato dimostrato che questa trasformazione è la Generalized Logarithm Transformation (GLOG):

$$f(Y_g) = \ln\left(Y_g - \alpha_g + \sqrt{(Y_g - \alpha_g)^2 + c}\right), \qquad c = \sigma_\varepsilon^2 / S_\eta^2 \tag{2.7}$$

Questa funzione ha le seguenti proprietà:

- è monotona crescente per tutti i valori di $Y_g$;
- quando $\mu_g \to 0$ è approssimativamente lineare;
- per valori elevati di $\mu_g$ assume una distribuzione logaritmica;
- rende costante e pari a $S_\eta^2$ la varianza asintotica.

## 2.4. Misure di espressione

Come spiegato nel §2.1, la tecnologia Affymetrix permette di ottenere due valori in ogni probe cell: i *PM*, che misurano l'ibridazione "corretta", e i *MM*, che misurano l'ibridazione aspecifica. I probe pairs del probeset sono solitamente tra gli 11 e i 20 e, dunque, per un singolo spot (gene) si hanno a disposizione tra i ventidue e i quaranta valori. Ora il problema è come riassumere questi in un unico valore di espressione.

Nel tempo sono stati proposti diversi metodi per ottenere una sola misura di espressione. Uno dei primi e più utilizzati era AvDiff, sostanzialmente basato su una media delle differenze $PM_i - MM_i$, dove l'indice $i = 1,\ldots,I$ è il numero del probe pair.

Successivamente si è osservato che misure come AvDiff non sono ottimali a causa di effetti specifici dei singoli probe che la semplice media di differenze non teneva in considerazione. Li e Wong (2001) hanno quindi proposto una misura alternativa basata sull'utilizzo di un modello del tipo $PM_{ji} - MM_{ji} = \theta_j \phi_i + \epsilon_{ji}$, nel quale $\phi_i$ rappresenta l'effetto specifico del probe $i$ e $\epsilon_{ji}$ è un termine d'errore indipendente e normalmente distribuito. In questo modello la stima del livello di espressione è data da $\hat{\theta}_j$, stima di massima verosimiglianza di $\theta_j$.

La stessa Affymetrix, resasi conto che la sua precedente misura AvDiff non era ottimale, ha proposto un algoritmo, detto MAS 5.0, per il calcolo di una misura riassuntiva di espressione. Questa è definita come $Tukey\ Biweight\{\log(PM_i - CT_i)\}$. E' basata sull'utilizzo di uno stimatore

robusto (Tukey's biweight), mentre le quantità $CT_i$ sono derivate dai valori *MM* in modo da non essere mai maggiori dei corrispondenti *PM*.

Tutte le misure sovraesposte sono basate sulla differenza tra *PM* e *MM*, ma in realtà è stato dimostrato che i *MM* non colgono solo l'ibridazione aspecifica, ma anche il segnale vero e proprio, per cui alcuni autori hanno preferito concentrarsi su misure che coinvolgessero solo i *PM* per l'ottenimento di una misura di espressione. Tra queste, una delle più note è il metodo RMA, descritto nel prossimo paragrafo.

### 2.4.1. Il metodo RMA

Come dimostrato da Irizarry et al. (2003), la sottrazione dei valori *MM* dai *PM* non è in grado di eliminare l'effetto specifico del probe e, inoltre, i *MM* colgono una parte del segnale, per cui ci possono essere dei problemi legati al loro utilizzo.

Il metodo RMA (Robust Multi-array Average) prevede la costruzione di un modello per ricavare un valore di espressione per ogni probeset:

$$Y_{jig} = \mu_{jg} + \alpha_{ig} + \epsilon_{jig} \tag{2.10}$$

dove l'indice $j$ è riferito al campione ($j$ = 1,…,$n$), $i$ è l'indice del probe pair ($i$ = 1,…,$I$) e $g$ indica il probeset (gene). $Y$ identifica solamente i *PM*, considerati nella trasformata logaritmica (log 2), quindi la misura non tiene conto dei *MM*. $\mu_{jg}$ rappresenta il fattore campione, $\alpha_{ig}$ un effetto specifico del probe e $\epsilon_{jig}$ rappresenta un termine d'errore a media zero indipendente e identicamente distribuito. La stima di $\mu_{jg}$ dà il valore di espressione, su scala logaritmica, per il probeset $g$ dell'array $j$.

In realtà, $Y_{jig}$ non sono semplicemente i valori dei *PM* considerati in logaritmo, ma si assume che, prima, ci sia stata la sottrazione del background e la normalizzazione Quantile. La rimozione del background consiste nel considerare i valori dei *PM* come derivanti da un modello del tipo $PM_{jig} = bg_{jig} + s_{jig}$. In pratica, si considera che ogni realizzazione dei *PM* deriva da una somma tra una componente di background, la quale non è biologica, e una componente di segnale, che rappresenta l'ibridazione avvenuta del cDNA. La componente di

background, dunque, comprende "rumore" derivante dalla scansione (optical noise) e anche dall'ibridazione non specifica. La correzione consiste nel considerare la trasformazione $B(PM_{jig}) = E[s_{jig} | PM_{jig}]$ e assumere che $s_{jig}$ abbia una distribuzione esponenziale e $bg_{jig}$ abbia distribuzione normale.

Per ricapitolare, il metodo RMA prevede (i) la correzione dei valori di espressione *PM* tramite la trasformazione $B(\cdot)$, (ii) la normalizzazione degli array utilizzando la Quantile, e (iii) la stima di un modello lineare, per ogni probeset $g$, per i valori corretti, normalizzati e trasformati in logaritmo (base 2). La stima di $\mu_{jg}$ è il valore di espressione su scala logaritmica e viene ottenuta utilizzando uno stimatore robusto per proteggersi da possibili outliers.

## 2.5. Identificazione dei geni differenzialmente espressi

La fase successiva alla normalizzazione è l'identificazione dei geni differenzialmente espressi, definendo in questo modo i geni che hanno un valore di espressione significativamente diverso tra due o più condizioni. Si tratta di un problema di verifica d'ipotesi di dimensioni molto grandi, dato che si dovrà condurre un test per ognuno delle migliaia di geni rilevati.
Si passeranno in rassegna i principali test statistici utilizzati per l'individuazione dei geni differenzialmente espressi, ponendo poi attenzione al problema della significatività dei test multipli e di come questo venga risolto.

### 2.5.1. t-Test

Se il problema che si vuole risolvere è il confronto tra due condizioni, il primo test parametrico che viene in mente è il test *t*. Si tratta di un test semplice, che risponde al problema di verificare se le medie di due popolazioni possono essere considerate uguali. Supponendo di disporre dei valori di espressione nella prima e nella seconda condizione sperimentale, l'ipotesi statistica da verificare è la seguente:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$$

dove si sta supponendo che $Y_1$ sia la variabile casuale che descrive il valore di espressione dei geni nella prima condizione, mentre $Y_2$ quella che descrive i valori di espressione nella seconda condizione. Si supponga, inoltre, che $n_1$ e $n_2$ siano le rispettive numerosità campionarie. Il test *t* assume che le osservazioni siano i.i.d. e provengono da una distribuzione normale, ossia $Y_1 \sim N(\mu_1, \sigma^2)$ e $Y_2 \sim N(\mu_2, \sigma^2)$. Si noti che si sta assumendo anche l'omoschedasticità, ossia la stessa varianza nelle due popolazioni.

La statistica utilizzata è la seguente:

$$T_g = \frac{\bar{y}_{1g} - \bar{y}_{2g}}{\sqrt{s_g^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{2.11}$$

dove:

$$s_g^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2} \tag{2.12}$$

è detta varianza *pooled*.

E' noto che $T_g \sim t_{n_1+n_2-2}$, per cui la decisione sull'accettazione o il rifiuto di $H_0$ può essere presa calcolando il corrispondente p-value osservato.

## 2.5.2. Test SAM

Uno dei maggiori problemi legati all'utilizzo del test *t* in dati di microarray è che i livelli di espressione vanno da valori prossimi allo zero a valori estremamente elevati, con spesso un numero di repliche per condizione sperimentale molto basso. Capita quindi che geni con valori di espressione bassi abbiano anche valori di varianza molto bassi e questo provoca l'esplosione dei valori del test verso valori alti, tanto da portare al rifiuto di $H_0$. Per questo motivo sono state proposte delle statistiche alternative, definite in questo contesto *test t moderati*, il cui scopo è ridurre questo fenomeno. Una di queste è il test SAM (Signifiance Analysis Microarray), il quale modifica la statistica del *t*-test aggiungendo al denominatore una costante chiamata *fudge factor*, per evitare che i geni poco espressi dominino l'analisi.

L'ipotesi che si vuole verificare è identica a quella descritta per il test *t*:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$$

mentre la statistica test utilizzata è la seguente:

$$T_g = \frac{\bar{y}_{1g} - \bar{y}_{2g}}{\sqrt{s_g^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right) + s_0}} \tag{2.13}$$

$s_0$ ha il compito di fare in modo che geni con valori di espressione troppo bassi, e dunque varianze basse, non dominino l'analisi. Ciò che si desidera è che la distribuzione dei valori di $T_g$ sia indipendente dal livello di espressione dei geni, ossia che $s_0$ renda il coefficiente di variazione di $T_g$ costante, indipendentemente dai valori $s_g$. Per questo motivo la scelta del *fudge factor* è specifica per ogni dataset. Alcuni autori hanno suggerito di porlo pari al novantesimo percentile della distribuzione dei $s_g$, altri hanno proposto algoritmi automatici per la sua scelta.

Diversamente dal test *t*, in questo caso non c'è l'assunzione di normalità delle due popolazioni, dunque la statistica test $T_g$ non ha distribuzione *t* di Student sotto $H_0$. Perciò è necessario utilizzare un approccio permutazionale. Considerando la matrice di espressione così come definita al §2.2, definito *n* il numero totale di campioni (colonne) e *p* il numero totale di geni (righe):

1) Calcolare i valori osservati della statistica $T_g$ per ogni gene, $g = 1,\ldots,G$;
2) Ordinare in modo crescente i valori osservati:
$$T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(G)} ;$$
3) Effettuare *K* permutazioni per colonna dei dati osservati e, per ogni permutazione, ottenere:
$$T_g^{*k} = \frac{\bar{y}_{1g}^{*k} - \bar{y}_{2g}^{*k}}{s_g^{*k} + s_0^{*k}} \tag{2.14}$$
In questo modo si ottengono *K* valori della statistica per ogni gene;
4) Ordinare in modo crescente i valori appena ottenuti:
$$T_{(1)}^{*k} \leq T_{(2)}^{*k} \leq \cdots \leq T_{(G)}^{*k} ;$$
5) Per ogni gene calcolare la quantità media:
$$\bar{T}_{(g)} = \frac{1}{K} \sum_{k=1}^{K} T_{(g)}^{*k} \tag{2.15}$$
Si ottengono così *G* valori medi.

Per identificare i geni differenzialmente espressi è necessario confrontare le statistiche originali $T_{(g)}$ con i valori medi attesi sotto $H_0$, $\bar{T}_{(g)}$, appena calcolati.

Se il valore originale si discosta molto rispetto al valore medio, allora si rifiuta l'ipotesi di uguaglianza di espressione. In pratica si definiscono due soglie, $t_1$ e $t_2$, dipendenti da Δ:

$$\left|T_{(g)} - \bar{T}_{(g)}\right| > \Delta \tag{2.16}$$

Graficamente si ha:

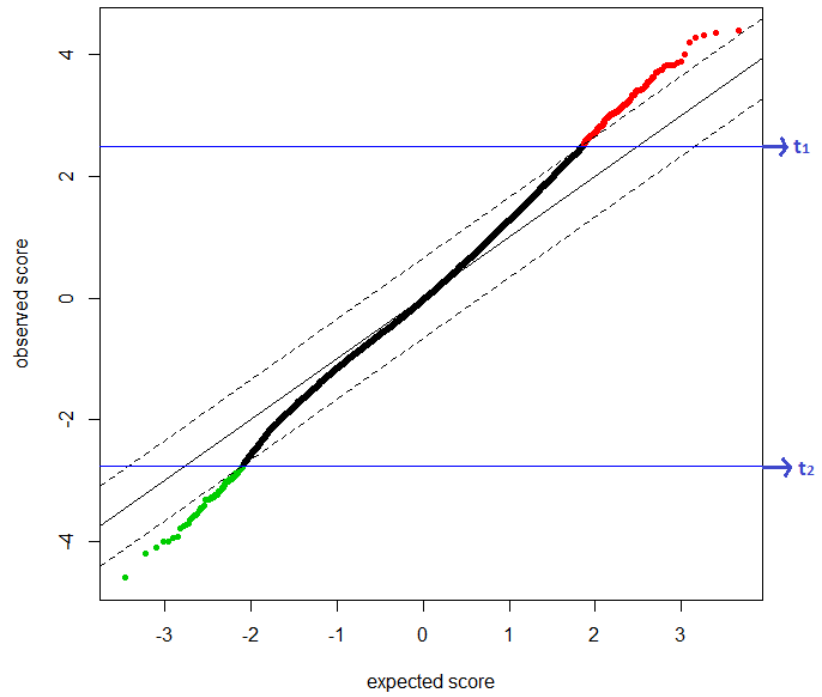



**Figura 2.2: Grafico test SAM. In ascissa sono riportati i valori medi $\bar{T}_{(g)}$ e in ordinata quelli delle statistiche originali $T_{(g)}$. Le due fasce parallele tratteggiate hanno ampiezza pari a Δ, mentre le due linee orizzontali definiscono le due soglie $t_1$ e $t_2$.**

Nodo centrale è la scelta di Δ, che regola il numero di geni definiti come differenzialmente espressi. Questo valore è scelto in modo da controllare la presenza di falsi positivi.

### 2.5.3. Test Ebayes

Il test Ebayes (Empirical Bayes) utilizza un approccio Bayesiano empirico per il calcolo delle probabilità a posteriori di ogni gene di essere differenzialmente espresso. Questo metodo offre la possibilità di utilizzare informazioni raccolte da

tutto l'insieme di geni per sfruttarle nell'inferenza di ogni singolo gene; in altre parole sfrutta la struttura parallela dell'analisi.

Per ogni gene si assume un modello lineare del tipo:

$$Y_g = X\alpha_g + \varepsilon_g \tag{2.17}$$

dove $Y_g$ è il vettore dei valori di espressione, $X$ è la matrice del disegno, $\alpha_g$ è il vettore dei parametri e $\varepsilon_g$ è un termine d'errore non necessariamente normale a media zero. Si avrà, dunque, che:

$$E[Y_g] = X\alpha_g \tag{2.18}$$

e

$$VAR(Y_g) = W_g\sigma_g^2 \tag{2.19}$$

dove $W_g$ è una matrice di pesi, nota, definita non negativa.

Nel caso di microarray a singolo canale, la matrice del disegno è uguale a quella dei modelli lineari classici. Se si suppone di avere, ad esempio, tre array nella condizione 1 e tre nella condizione 2 la matrice sarà del tipo:

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \tag{2.20}$$

Alcuni contrasti sono di interesse biologico e si possono definire come:

$$\beta_g = C^T\alpha_g \tag{2.21}$$

Si assume che sia di interesse verificare l'ipotesi che il singolo valore del contrasto sia pari a zero, ossia $\beta_{gj} = 0$.

Definite:

- $\hat{\alpha}_g$ stima di $\alpha_g$
- $s_g^2$ stima di $\sigma_g^2$
- $VAR(\hat{\alpha}_g) = V_g s_g^2$
- $\hat{\beta}_g = C^T\hat{\alpha}_g$
- $VAR(\hat{\beta}_g) = C^T V_g C s_g^2$

le stime ottenute, le assunzioni di distribuzione sui parametri sono le seguenti:

$$\hat{\beta}_{gj}|\beta_{gj}, \sigma_g^2 \sim N(\beta_{gj}, v_{gj}\sigma_g^2) \tag{2.22}$$

$$s_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi_{d_g}^2 \tag{2.23}$$

dove $d_g$ è il numero di gradi di libertà residui del modello lineare per il gene $g$.

La statistica test che ne deriva è allora:

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g\sqrt{v_{gj}}} \sim t_{d_g} \tag{2.24}$$

Finora, però, non si è sfruttata la struttura parallela dell'analisi che si sta svolgendo. Ciò viene fatto imponendo una gerarchia sui parametri:

- $\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2}\chi_{d_0}^2$ (2.25)
- $P(\beta_{gj} \neq 0) = p_j \; \forall j$ (2.26)
- $\beta_{gj}|\sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{0j}\sigma_g^2)$ (2.27)

Il primo punto spiega come varia la varianza tra i geni; il secondo indica la vera percentuale di geni differenzialmente espressi; il terzo descrive la distribuzione dei *fold change* per i geni che sono differenzialmente espressi.
Utilizzando un modello di questo tipo la media a posteriori di $\sigma^2|s_g^2$ è:

$$s_g^{2*} = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g} \tag{2.28}$$

La statistica test moderata risulta dunque:

$$t_{gj}^* = \frac{\hat{\beta}_{gj}}{s_g^*\sqrt{v_{gj}}} \sim t_{d_g + d_0} \tag{2.29}$$

E' stato dimostrato che $t^*$ e $s^2$ sono indipendenti in distribuzione. La differenza nei gradi di libertà rispetto al classico *t*-test, ovvero la quantità $d_0$, riflette l'informazione aggiuntiva data dall'utilizzo di tutto l'insieme di geni nel modo sopra descritto.

Resta in sospeso la scelta dei parametri $s_0$ e $d_0$. Questi parametri vengono stimati dai dati rilevati ed è dunque per questo motivo che si parla di approccio Bayesiano empirico. La stima avviene attraverso il metodo dei momenti, eguagliando i primi due momenti teorici della variabile $\log(s_g^2)$ con quelli empirici.

Come detto in precedenza, il metodo Bayesiano empirico calcola le probabilità a posteriori, per ogni gene, di essere sregolato. Questo viene fatto semplicemente calcolando l'ODD a posteriori, in relazione al contrasto $\beta_{gj}$:

$$O_{gj} = \frac{P(\beta_{gj}\neq 0|t_{gj}^*, s_g^2)}{P(\beta_{gj}=0|t_{gj}^*, s_g^2)} = \frac{P(\beta_{gj}\neq 0, t_{gj}^*, s_g^2)}{P(\beta_{gj}=0, t_{gj}^*, s_g^2)} = \frac{p_j}{1-p_j}\frac{P(t_{gj}^*|\beta_{gj}\neq 0)}{P(t_{gj}^*|\beta_{gj}=0)} \tag{2.30}$$

E' utile considerare la seguente trasformazione dell'ODD:

$$B_{gj} = \log(O_{gj}) \tag{2.31}$$

che assume valore pari a zero nel caso di maggiore incertezza, ossia quando l'ODD è pari a uno: ciò significa che le due probabilità *p* e *1-p* sono uguali.

### 2.5.4. Test multipli, FWER e FDR

Esiste un problema statistico molto noto che si presenta quando si effettuano molti test sullo stesso insieme di dati: la significatività globale dei test svolti non è pari alla significatività di ogni singolo test condotto. A questo scopo si definisca $\alpha$ il livello di significatività del test: ciò significa che $\alpha$ è la probabilità di rifiutare $H_0$ quando è vera. Questo è definito errore di primo tipo. L'errore di secondo tipo, invece, consiste nell'accettare $H_0$ quando è falsa e la sua probabilità è $\beta$. Si definisce potenza del test la probabilità $1 - \beta$.

Si supponga di effettuare $N$ verifiche d'ipotesi, ognuna con livello significatività $\alpha$. In generale, se i test sono tra loro indipendenti, la probabilità di commettere almeno un errore è pari a $1 - (1 - \alpha)^N$. Se $N = 10$, questa probabilità è circa $0.4013$. Se $N = 200$ è $0.99996$, cioè è praticamente certo commettere almeno un errore.

Quando i test sono tra loro dipendenti quanto detto prima non vale e la significatività globale può essere sia maggiore che minore. L'unica affermazione che può essere fatta è sul limite superiore di questa probabilità. E' noto, infatti, che, se $\alpha_i$ è il livello di significatività dell'i-esimo test, si ha che $\alpha_t \leq \sum_{i=1}^{n} \alpha_i$, dove $\alpha_t$ è il livello di copertura totale. Per un maggiore approfondimento si veda Wiley (2006).

In ambito genomico, dato che si deve effettuare un test per ogni gene, il numero di test che si devono verificare è dell'ordine di migliaia di geni; per le piattaforme di ultima generazione si parla anche di 50.000 test. Ciò significa che, se 5% è il livello di significatività di ogni test, è atteso che il 5% dei geni sarà un *falso positivo*, ossia sarà stata rifiutata l'ipotesi $H_0$ di uguale espressione sbagliando. Dunque, con ad esempio 10.000 geni, 500 saranno attesi falsi positivi. Tutto questo se si considerano i test indipendenti, ma è noto che questi non lo sono. Da questo ne deriva che, in particolar modo in ambito genomico, è

necessario adottare delle strategie per tenere sotto controllo l'errore globale commesso.

La maggior parte dei metodi di correzione dei livelli di significatività è basata sul controllo della quantità chiamata Family Wise Error Rate (*FWER*). Questa quantità è definita come la probabilità di avere almeno un *falso positivo* tra i test fatti.

In relazione alla matrice di confusione rappresentata in Tabella 2.2 si ha:

$$FWER = \Pr(V \geq 1) \tag{2.32}$$

| | | **Test** | | ***Totale*** |
|---|---|---|---|---|
| | | *Accetto $H_0$ (-)* | *Rifiuto $H_0$ (+)* | |
| **Realtà** | *$H_0$ vera (-)* | $U$ | $V$ | $m_0$ |
| | *$H_0$ falsa (+)* | $T$ | $S$ | $m - m_0$ |
| ***Totale*** | | $m - R$ | $R$ | $m$ |

**Tabella 2.2: Eventi possibili in un test statistico. U è l'insieme dei geni identificati come egualmente espressi (EE) e che lo sono veramente (veri negativi); V è l'insieme dei geni identificati come differenzialmente espressi (DE) ma che non lo sono nella realtà (falsi positivi); T è l'insieme dei geni identificati come EE ma che in realtà non lo sono; S è l'insieme dei geni identificati come differenzialmente espressi (DE) e che lo sono veramente.**

Le più note correzioni che controllano il *FWER* sono quella di Bonferroni, quella di Holm e quella di Holm-Sidak.

Un'alternativa migliore è quella di controllare il False Discovery Rate (*FDR*). Questa quantità è definita come la frazione attesa di falsi positivi nella lista di geni differenzialmente espressi, ossia la percentuale di geni all'interno della lista degli identificati come differenzialmente espressi, che in realtà non lo sono. La misura della significatività tramite il controllo del *FDR* è detta *q-value*.

Si ritiene utile sottolineare la grande differenza tra l'utilizzo del p-value e del q-value (FDR). Utilizzando un p-value del 5% ci si aspetta, in media, che il 5% dei geni che *nella realtà* non sono differenzialmente espressi vengano identificati come differenzialmente espressi; con un FDR del 5% ci si aspetta, in media, che il 5% dei geni *della lista* degli identificati come differenzialmente espressi in realtà non lo siano.

In relazione alla precedente Tabella 2.2, il FDR è definito come segue:

$$FDR = E\left[\frac{V}{V+S}|V+S>0\right]\cdot \Pr(V+S>0) =$$
$$= E\left[\frac{V}{R}|R>0\right]\cdot \Pr(R>0) \quad (2.33)$$

Quando il numero di test eseguiti è elevato $(m \to \infty)$ la quantità $\Pr(R>0)$ tende a uno. Per questo motivo si può scrivere:

$$FDR \approx E\left[\frac{V}{R}|R>0\right] \approx \frac{E[V]}{E[R]} \quad (2.34)$$

Naturalmente le quantità $V$, $U$, $T$, $S$ dipendono dalla soglia *t* scelta, infatti si ha:

$$V(t) = \#\{p_i \le t; i = 1,\dots,m;\ H_0\ vera\} \quad (2.35)$$

$$R(t) = \#\{p_i \le t; i = 1,\dots,m\} \quad (2.36)$$

e dunque, per maggiore precisione, si ha:

$$FDR = \frac{E[V(t)]}{E[R(t)]} \quad (2.37)$$

Mentre $R(t)$ è noto, $V(t)$ non lo è ed è dunque necessario stimarlo. Ciò che si sa è che, sotto $H_0$, la distribuzione dei p-value è uniforme, quindi:

$$E[V(t)] = m_0 \cdot t \quad (2.38)$$

Bisogna allora stimare $m_0$ o, equivalentemente, la proporzione $\pi_0 = m_0/m$ la quale, in riferimento alla Tabella 2.2, rappresenta la proporzione di veri negativi.

Ottenuta una stima di $\pi_0$, $\hat{\pi}_0$, si può ottenere una stima del FDR come:

$$\widehat{FDR}(t) = \frac{\hat{\pi}_0 \cdot m \cdot t}{R(t)} = \frac{\hat{\pi}_0 \cdot m \cdot t}{\#\{p_i \le t; i=1,\dots,m\}} \quad (2.39)$$

A questo punto, il q-value può essere definito come:

$$\hat{q}(p_i) = \min_{t \ge p_i} \widehat{FDR}(t) \quad (2.40)$$

ossia come il minimo *FDR* che si ottiene quando si definisce significativo un test.

In realtà si dimostra che, dopo aver ordinato i p-value in ordine crescente, il q-value relativo all'*i*-esimo p-value ordinato è:

$$\hat{q}\left(p_{(i)}\right) = \min_{t \ge p_{(i)}} \frac{\hat{\pi}_0 \cdot m \cdot t}{\#\{p_j \le t\}} = \min\left(\frac{\hat{\pi}_0 \cdot m \cdot p_{(i)}}{i}, \hat{q}\left(p_{(i+1)}\right)\right) \quad (2.41)$$

per *i* = 1, …, *m-1*, mentre per l'm-esimo p-value il relativo q-value è semplicemente $\hat{q}\left(p_{(i)}\right) = \hat{\pi}_0 \cdot p_{(m)}$.

I primi due autori che affrontarono il problema della stima del FDR furono Benjamini e Hochberg (1995), i quali decisero di stimarlo considerando il caso più conservativo possibile, ossia quando $\hat{\pi}_0 = 1$, che equivale a dichiarare che

nessun gene è differenzialmente espresso. Il p-value osservato per una singola verifica d'ipotesi viene dunque corretto calcolando $P_{(j)}^{BH} = \min_{j \le i}(P_i \cdot m/i)$.

Successivamente, i due autori Storey e Tibshirani (2003) hanno proposto un metodo per stimare $\pi_0$. Questo prevede di analizzare la distribuzione dei p-value e cercare la soglia oltre la quale la distribuzione è uniforme, ossia stimare $\pi_0$ con:

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \dots, m\}}{m(1-\lambda)} \tag{2.42}$$

I q-value si ottengono, dunque, inserendo la stima appena trovata nella (2.41).

## 2.6. Gene set analysis

La gene set analysis è la fase successiva a quella di identificazione dei geni differenzialmente espressi. L'obiettivo di questa è il raggruppamento dei geni in classi funzionali (gene set) per l'interpretazione dei risultati ottenuti. In pratica, piuttosto che studiare i geni singolarmente, si ricercano gruppi di geni tra loro collegati che abbiano livelli di espressione differenti in condizioni biologiche diverse, in modo da comprendere i processi cellulari coinvolti nel problema biologico che si sta analizzando. E' evidente che, da un punto di vista statistico, queste analisi si concretizzano in una verifica d'ipotesi. Esistono svariati metodi per effettuare una gene set analysis; in particolare si è soliti distinguere tra *metodi competitivi* e *metodi indipendenti* e la differenza tra i due sta nella diversa ipotesi nulla che viene considerata. Nei primi, infatti, si vuole verificare l'ipotesi che un certo gene set abbia lo stesso livello di associazione con il fenotipo che il resto dei geni; nei secondi, invece, l'ipotesi nulla è che nessun gene di un gene set sia associato con il fenotipo considerato. I metodi indipendenti, dunque, considerano solo i geni di uno specifico gene set e non quelli di altri. E' evidente che l'ipotesi nulla nei metodi indipendenti è più restrittiva rispetto a quella nei metodi competitivi ed è per questo che i primi sono più potenti dei secondi e danno risultati significativi per un numero maggiore di gene set. I metodi competitivi, inoltre, dipendono fortemente dalla soglia scelta per l'identificazione dei geni differenzialmente espressi. Accade così che geni con cambiamenti moderati nei valori di espressione non vengano identificati come differenzialmente espressi a

causa del valore di cut-off troppo stringente e questo porta ad una ulteriore riduzione della potenza statistica.

Altra distinzione che può essere effettuata è quella tra metodi *topologici* e *non topologici*. Per comprendere questa suddivisione è necessario capire cosa si intende per pathway biologico. Questo non è una mera lista di geni ma riproduce le relazioni biologiche tra le macromolecole di una cellula. Queste ultime sono rappresentate tramite grafi nei quali ogni nodo è un gene ed ogni arco è una relazione tra i geni. I metodi topologici, dunque, sono quelli che sono in grado di sfruttare il grafo, guadagnando potenza da tutte le informazioni che questo contiene. I metodi non topologici, invece, trattano un pathway come una semplice lista di geni, senza sfruttare le informazioni sulle loro relazioni. E' necessario sottolineare che, dato che i nodi dei pathway biologici non sono geni ma entità più grandi (es. proteine o famiglie di geni), i pathway devono essere convertiti in reti di geni per essere utili allo scopo di queste analisi. La difficoltà di far ciò è uno dei motivi per cui, ad oggi, i metodi maggiormente utilizzati sono quelli non topologici. Altro motivo di ciò è la difficoltà di integrare lo sfruttamento della topologia del grafo nei modelli statistici.

Uno dei database più utilizzati per la conoscenza dei pathway è la Kyoto Encyclopedia of Genes and Genomes (KEGG), la quale contiene i pathway di geni con funzioni regolatrici e metaboliche conosciuti. Altri database dello stesso tipo sono Biocarta, Reactome e WikiPathways.

### 2.6.1. Analisi di arricchimento

L'analisi di arricchimento classica è un metodo competitivo e non topologico. E' basata sostanzialmente sul test esatto di Fisher e stima la probabilità di osservare casualmente un certo numero di geni di uno specifico pathway nella lista di geni identificati come differenzialmente espressi. In particolare, per ogni pathway (gene set) si considera la tabella di contingenza riportata in Tabella 2.3, dove *EEG* sta per geni egualmente espressi, *N* è il numero totale di geni della piattaforma utilizzata, *G* è il pathway considerato e $G^C$ è il suo complementare.

La probabilità di osservare casualmente almeno $n_{G,DEG}$ geni della categoria funzionale descritta da *G* nella lista dei geni differenzialmente espressi è data da:

$$P\left(N_{G,DEG} \geq n_{G,DEG}\right) = \sum_{i=n_{G,DEG}}^{N_{DEG}} \frac{\binom{N_G}{i}\binom{N-N_G}{N_{DEG}-i}}{\binom{N}{N_{DEG}}} \quad (2.43)$$

Dato che si deve effettuare un test per ogni pathway, le probabilità devono essere aggiustate per i noti problemi legati ai test multipli. Questo può essere fatto, ad esempio, con il metodo di Benjamini e Hochberg illustrato nel §2.5.4.

| | ***DEG*** | ***EEG*** | ***Tot*** |
|---|---|---|---|
| ***G*** | $n_{G,DEG}$ | $n_{G,EEG}$ | $N_G$ |
| $\boldsymbol{G^C}$ | $n_{G^C,DEG}$ | $n_{G^C,EEG}$ | $N_{G^C}$ |
| ***Tot*** | $N_{DEG}$ | $N_{EEG}$ | $N$ |

**Tabella 2.3: Tabella di contingenza relativa ad un'analisi di arricchimento.**

## 2.6.2. Global test

Il global test è un metodo indipendente e non topologico basato su un modello di regressione logistica. In questo modello la variabile dipendente è la classe biologica di appartenenza, mentre le covariate sono i profili di espressione dei geni appartenenti ad un pathway. In particolare, trattandosi di un modello lineare generalizzato (GLM), si ha:

$$h(E[Y_i|\beta]) = \alpha + \sum_{j=1}^{m} \beta_j x_{ij} \quad (2.44)$$

dove $\beta_j$ è il coefficiente di regressione del gene *j*. Per verificare l'ipotesi di assenza di associazione dei geni di un pathway con il fenotipo è sufficiente verificare l'ipotesi di nullità di tutti i coefficienti di regressione, ossia:

$$H_0: \beta_1 = \cdots = \beta_m = 0 \quad (2.45)$$

Il numero di parametri del modello è uguale al numero di geni del pathway ma tipicamente, in questo genere di esperimenti, il numero di repliche a disposizione è minore del numero di geni e, dunque, la verifica d'ipotesi non può essere fatta nella maniera classica. E' possibile, però, assumere che tutti i $\beta_j$ derivino da una distribuzione comune con valore atteso pari a zero e varianza $\tau^2$. Questo parametro, ignoto, determina quanto i coefficienti sono liberi di deviare

dal loro valore atteso, cioè da zero. L'ipotesi nulla precedente può allora essere riformulata come:

$$H_0: \tau^2 = 0 \tag{2.46}$$

Questo modello può essere visto in diversi modi. Il primo è considerare gli assunti distributivi di $\beta$ come una distribuzione a priori con una forma ignota e varianza dipendente da un solo parametro $\tau^2$. Visto in questo modo si ha un modello Bayesiano empirico. Una seconda interpretazione è vedere il modello come un modello di regressione penalizzato, dove i coefficienti sono compressi verso una media comune. La log-verosimiglianza di *Y*, infatti, può essere riscritta come:

$$l(Y, \beta) = l(Y|\beta) + l(\beta) \tag{2.47}$$

dove il primo termine è la log-verosimiglianza del modello lineare generalizzato, mentre il secondo termine è una penalità. La stima dei parametri, dunque, può essere fatta utilizzando un modello di regressione penalizzato come la regressione Ridge o il Lasso (Goeman, et al., 2004).

### 2.6.3. Signaling Pathway Impact Analysis

Il metodo Signaling Pathway Impact Analysis (SPIA) è un metodo topologico e misto, ossia in parte competitivo e in parte indipendente. E' basato sul calcolo di un punteggio ottenuto combinando diversi aspetti dei dati: il fold change dei geni differenzialmente espressi, un punteggio di arricchimento del pathway e la sua topologia. In particolare, SPIA aumenta l'impatto di un pathway se i geni differenzialmente espressi tendono ad essere vicini ai punti di entrata dello stesso e lo riduce se, invece, i geni differenzialmente espressi si trovano in punti terminali.

SPIA combina l'evidenza che si ottiene dalla classica analisi di arricchimento con un altro tipo di evidenza, la quale misura la vera perturbazione che subisce un pathway in una certa condizione (Tarca, et al., 2009). Gli autori hanno dimostrato che queste evidenze sono tra loro indipendenti e, dunque, può essere definito un p-value globale come combinazione dei due p-value derivanti dall'analisi di arricchimento e da quella di perturbazione del pathway. Nello specifico SPIA

calcola: (i) il p-value derivante dall'analisi di arricchimento classica, *pNDE;* (ii) un fattore di perturbazione come funzione lineare dei fattori di perturbazione provocati da ognuno dei geni del pathway, la cui significatività è ottenuta tramite bootstrap, *pPERT*; (iii) una combinazione dei due p-value precedenti, chiamata *pG*. Questi ultimi vengono poi aggiustati per controllare l'errore dovuto ai test multipli.

Il calcolo del fattore di perturbazione è fatto nel modo seguente:

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \cdot \frac{PF(g_j)}{N_{ds}(g_j)} \tag{2.48}$$

dove il primo termine, $\Delta E(g_i)$, rappresenta il cambiamento nell'espressione del gene $g_i$ (il log fold change nel caso di due condizioni sperimentali); il secondo termine è la somma dei fattori di perturbazione dei geni $g_j$ "a monte" del gene $g_i$ nella rete del pathway, normalizzata per il numero di geni "a valle" di ciascuno di essi, $N_{ds}(g_j)$. Il valore assoluto di $\beta_{ij}$ quantifica la forza dell'interazione tra il gene $g_i$ e il gene $g_j$. Questi pesi permettono di catturare le proprietà descritte dal pathway. Si noti che i pesi non sono stimati dai dati, ma fissati pari a $\pm 1$: hanno segno positivo se l'interazione tra i geni è un'attivazione, negativo se si tratta di inibizione. Il peso è posto pari a zero per i geni non direttamente collegati.

### 2.6.4. Pathway analysis through Gaussian Graphical Models

Il metodo che si vuole introdurre in questo paragrafo fa uso dei modelli grafici Gaussiani. Per comprendere a fondo il suo funzionamento, dunque, si farà una breve introduzione a questo tipo di modelli grafici. Per maggiori approfondimenti si veda (Massa, et al., 2010) e i suoi riferimenti.

#### *Modelli grafici Gaussiani*

Un grafo $G$ è una coppia $G = (V, E)$, dove $V$ è un insieme di vertici (nodi) ed $E$ è un insieme di archi ($E \subseteq V \times V$) composto da coppie ordinate di vertici diversi. Se entrambi $(u, v) \in E$ e $(v, u) \in E$, l'arco $(u, v)$ è detto non orientato. Se $(u, v) \in E$ ma $(v, u) \notin E$, l'arco $(u, v)$ è detto orientato. Un grafo con archi

non orientati è detto non orientato; se gli archi sono, invece, orientati, anche il grafo è detto orientato. In un grafo non orientato, se c'è un arco tra $u$ e $v$, i due nodi sono detti adiacenti; in un grafo orientato, se $u \to v$, $u$ è detto padre di $v$ e $v$ è detto figlio di $u$. Un path è una sequenza di vertici per cui ogni vertice ha un ha un arco che lo collega a quello successivo della sequenza. Un ciclo è un path tale per cui il primo vertice è anche l'ultimo del path. Un grafo orientato aciclico (DAG) è un grafo orientato senza cicli. Dato un DAG $D$, un grafo morale $D^m$ è un grafo non orientato ottenuto da $D$ aggiungendo archi non orientati tra tutte le coppie di vertici che hanno figli in comune (se non sono già presenti) e trasformando tutti gli archi in non orientati. Un grafo è completo se $E$ contiene tutte le coppie di elementi distinti di $V$, ossia se ogni vertice è collegato a tutti i vertici rimanenti. Si definisce clique un sotto-grafo completo non contenuto in nessun altro sotto-grafo completo. Un grafo triangolato è un grafo non orientato con la proprietà che ogni ciclo di lunghezza $n \geq 4$ possiede una corda, ossia c'è un arco tra due nodi non consecutivi del ciclo.

I pathway vengono rappresentati come grafi composti da nodi e archi. Come detto in precedenza, però, i nodi non sono sempre singoli geni, ma anche prodotti dei geni come proteine, RNA e altre entità complesse. Le relazioni tra questi, descritte con archi, possono essere di diversi tipi, tra cui dirette o indirette. Per essere utili alle analisi che si vogliono svolgere, i pathway devono essere processati e tradotti in reti di geni. La prima fase è quella di conversione del pathway in un grafo (DAG) $D$, rendendo tutti gli archi di esso orientati, facendo ricorso anche a informazioni provenienti da altri database come Biocarta. In caso di nodi che rappresentino entità complesse, si considera come espressione la prima componente principale, ossia la combinazione lineare dei valori di espressione di tutti i geni coinvolti la quale cattura la maggior parte della variabilità dei dati.

La seconda fase consiste nella conversione di $D$ in un grafo morale, $D^m$, ottenendo un grafo non orientato. Detto $G$ questo grafico, assumendo di modellare i dati dello stesso pathway in due diverse condizioni come realizzazioni di un modello grafico Gaussiano si ha:

$$M_1(G) = \left\{Y_1 \sim N_p(\mu_1, \Sigma_1), \Sigma_1^{-1} \in S^+(G)\right\} \tag{2.49}$$

$$M_2(G) = \{Y_2 \sim N_p(\mu_2, \Sigma_2), \Sigma_2^{-1} \in S^+(G)\} \quad (2.50)$$

dove $p$ è il numero di geni (vertici del grafo) e $S^+(G)$ è l'insieme delle matrici simmetriche definite positive con elementi nulli corrispondenti agli archi mancanti in $G$. L'assunto di normalità è motivato dal fatto che i valori di espressione sono approssimativamente normali nella scala logaritmica. In pratica i nodi vengono considerati come realizzazioni di una variabile casuale Gaussiana multivariata con una specifica matrice di concentrazione (inversa della matrice di covarianza) che riflette le dipendenze presenti tra le variabili. Perché la matrice di concentrazione sia rappresentativa delle specifiche relazioni presenti tra i nodi è necessario che ci siano dei vincoli su alcuni dei suoi valori, in particolare su quelli che devono rappresentare l'assenza di collegamento tra due nodi. Questi valori sono vincolati ad essere pari a zero.

In applicazioni reali non sono noti $\mu_1$, $\Sigma_1$, $\mu_2$ e $\Sigma_2$, ma si conoscono le posizioni degli zeri nelle matrici di concentrazione, $\Sigma_1^{-1}$ e $\Sigma_2^{-1}$, definiti dal grafo. E' necessario, dunque, stimare i suddetti parametri dai dati. La stima delle matrici di covarianza viene fatta tramite un algoritmo detto Iterative Proportional Scaling algorithm (*IPS*) che si applica alle matrici di covarianza stimate dai dati. Questo algoritmo garantisce che le matrici di concentrazione create siano definite positive e con elementi nulli in corrispondenza degli archi mancanti nel grafo.

### *CliPPER*

Le dimensioni dei pathway sono spesso molto grandi e ci si aspetta, da un punto di vista biologico, che solo alcune porzioni di esso siano veramente coinvolte nel causare differenze tra condizioni biologiche diverse, specialmente per pathway grandi. Tra tutti i metodi topologici, nessuno tenta di individuare le parti dei pathway maggiormente coinvolte nel problema biologico. In questa prospettiva è stato sviluppato CliPPER, un metodo indipendente e topologico basato su modelli grafici Gaussiani che (i) seleziona i pathway con matrici di covarianza o media significativamente diversa tra le condizioni sperimentali e, in questi, (ii) identifica le porzioni del pathway, dette *signal path*, che sono maggiormente associate con il fenotipo.

Solitamente, negli studi di espressione genica, le diverse condizioni sperimentali sono confrontate in termini di differenza di espressione media. Tali

differenze, però, non necessariamente si concretizzano in una modifica della forza delle interazioni tra i geni. Per esempio, un incremento proporzionale nell'espressione dei geni $A$ e $B$ in una delle due condizioni si tradurrà in una differenza di espressione media, ma la forza della correlazione tra $A$ e $B$ non cambia. In questo caso, dunque, si avranno pathway con una significativa alterazione nei livelli medi di espressione ma interazioni biologiche inalterate. Se, al contrario, i rapporti di espressione dei due geni cambiano in maniera non proporzionale, ci si aspetta un'alterazione significativa non solo nei livelli medi di espressione, ma anche nella forza delle loro correlazioni. Ciò corrisponde ad un cambiamento nell'attività biologica che può essere catturato controllando la covarianza dell'espressione. CliPPER, quindi, ricerca i pathway fortemente coinvolti in un processo biologico richiedendo che le medie o le covarianze dei livelli di espressione risultino significativamente alterate tra le due condizioni.

Vista in un contesto di modelli grafici Gaussiani, la traduzione di quanto detto sta nel compiere la seguente serie di passi: (i) identificazione del DAG associato al pathway, (ii) moralizzazione del DAG, (iii) triangolazione, (iv) identificazione delle clique e (v) costruzione del junction tree. Quest'ultima fase consiste nella creazione di un albero avente come nodi le clique identificate e che soddisfi la cosiddetta *running intersection property*, ossia per ogni clique $C_i$ e $C_j$ dell'albero, ogni clique del path che connette $C_i$ e $C_j$ deve contenere $C_i \cap C_j$.

Fatti questi passaggi si procede alla selezione dei pathway con medie o matrici di covarianza significativamente diverse nelle due condizioni effettuando, appunto, due tipi di verifica d'ipotesi: una sull'uguaglianza delle due matrici di concentrazione (inverse delle matrici di covarianza) che descrivono le relazioni tra i geni $(H_0: \Sigma_1^{-1} = \Sigma_2^{-1})$ e una sull'uguaglianza dell'espressione media tra le due condizioni sperimentali considerate $(H_0: \mu_1 = \mu_2)$. Questa seconda verifica d'ipotesi viene fatta in modo diverso a seconda dei risultati ottenuti nella prima e, cioè, a seconda che si possa assumere o meno l'omoschedasticità dei due modelli grafici Gaussiani. Come spiegato in precedenza, la stima delle matrici di covarianza viene fatta tramite un algoritmo, *IPS*, partendo dalle matrici di covarianza empiriche. Dato che accade frequentemente che il numero di campioni disponibili sia minore del numero di nodi del pathway, Martini et. al. (2013)

hanno implementato in CliPPER un metodo di shrinking per la stima della matrice di covarianza.

Riguardo, poi, l'identificazione delle porzioni di pathway maggiormente associate con il fenotipo, CliPPER parte dall'identificazione dei path e dei corrispettivi sub-path dal junction tree costruito e calcola le rilevanze dei sub-path nel modo indicato in Martini et. al. (2013).

# Capitolo 3

# 3. Variazioni non biologiche e batch effect

In questo capitolo si discuterà di un problema molto noto in studi di microarray, ovvero del problema dell'introduzione negli esperimenti di variabilità non rilevante dal punto di vista del fattore biologico che si sta considerando, ma dovuta ad altri fattori. Si presenteranno dunque due metodi per stimarla e rimuoverla, in modo da rendere maggiormente confrontabili gli esperimenti. Infine verranno proposte delle tecniche per la valutazione della bontà di un metodo di rimozione di questa forma di variabilità.

Come appena accennato, negli studi di espressione genica con l'utilizzo di microarray a singolo canale si assiste alla presenza di *unwanted variation*, letteralmente variazione non voluta. Oltre ai fattori biologici di interesse per gli sperimentatori, infatti, ci sono solitamente altri fattori, non biologici, che influenzano i valori di espressione. Questi possono essere, ad esempio, diversi parametri di scansione, diversa potenza del laser o diversi reagenti. Per questo motivo, prima di eseguire qualsiasi analisi sui dati, questi ultimi devono essere normalizzati, allo scopo di rimuovere le differenze sistematiche tra gli array che rappresentano varianza non biologica, dovuta al processo con il quale si ottengono i valori stessi. Si confronti il §2.3 per maggiori dettagli sulla normalizzazione degli array. La conoscenza di questo fenomeno è sostanzialmente dovuta alle analisi effettuate su array di repliche tecniche, nelle quali è nota l'assenza di differenziale espressione.

Un tipo molto più forte di variazione non voluta, per la quale non è sufficiente normalizzare i dati, è quella che si ottiene quando si utilizzano array ottenuti in condizioni sperimentali diverse. Questa è generalmente definita *batch effect* e si ha, ad esempio, quando i dati sono stati ricavati da due laboratori diversi, oppure con tecnologie di microarray diverse, o ancora semplicemente

quando i campioni vengono ricavati in giorni diversi. Il batch effect può essere causato addirittura dal diverso momento del giorno nel quale si fanno le repliche (mattina/pomeriggio) e dal livello di ozono nell'atmosfera (Fare, et al., 2003). Per campioni generati in uno stesso batch si intende, infatti, microarray processati nello stesso posto, in un breve periodo di tempo e utilizzando la stessa piattaforma (Chen, et al., 2011).

Come già detto, la normalizzazione non è in grado di rimuovere le variazioni dovute all'effetto batch. Questa metodologia, infatti, fa parte dei metodi di aggiustamento globali (global adjustment), i quali producono un dataset modificato rispetto a quello di partenza, dal quale è stata rimossa la varianza indesiderata. Per rimuovere il batch effect è necessario, invece, utilizzare metodi cosiddetti "application specific", i quali integrano il metodo di aggiustamento direttamente durante l'analisi di espressione. Un metodo "application specific" in uno studio di differenziale espressione è, ad esempio, l'aggiunta di un termine per modellare l'effetto batch in un modello lineare (Gagnon-Bartsch & Speed, 2012).

Generalmente è da considerarsi sbagliato combinare dati provenienti da batch diversi senza prima aggiustarli eliminando l'effetto batch. Sono stati proposti diversi metodi per correggere questo tipo di effetto, molti dei quali richiedono un numero di campioni per batch abbastanza alto, in genere più alto di 25. Nella pratica, però, molte analisi sono condotte con un numero di campione per batch più basso.

A questo scopo si vogliono analizzare due metodi: uno è un metodo Bayesiano empirico, il quale ha portato alla creazione di *COMBAT*, uno script appositamente creato per svolgere l'analisi con il software statistico R; l'altro è un metodo basato sull'utilizzo dei cosiddetti geni di controllo ed è denominato *RUV*. Di questo ne sono state proposte due versioni: la prima detta *RUV-2* (Remove Unwanted Variance, 2 steps) e la seconda, che rappresenta una modifica rispetto alla prima, detta *RUV-4*, la quale prevede un numero di passi più alto rispetto alla precedente.

# 3.1. Il metodo Bayesiano Empirico

Il metodo *COMBAT* è stato proposto da Evan Johnson e Cheng Li (2007). Si tratta di un metodo pensato per essere robusto nell'aggiustamento del batch effect in esperimenti con un numero di campioni per batch basso e si basa sulla stima di un modello per ognuno dei geni che si hanno a disposizione. Questo modello è dello stesso tipo di quelli utilizzati dai metodi che sfruttano aggiustamenti di posizione e scala (Location/Scale adjustments).

*COMBAT* può essere descritto in tre passi. Si assuma che i valori di espressione siano normalizzati; inoltre che siano stati eliminati i geni con valori di espressione nulla, non significativamente maggiore del background. La matrice dei valori di espressione contiene, quindi, *G* geni ($g$ = 1,…,*G*) divisi in *m* batch, ciascuno con $n_i$ campioni, *i* = 1,…,*m*.

Un modello L/S per i valori di espressione, che tenga conto dell'effetto batch, è il seguente:

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg} \quad (3.1)$$

dove $Y_{ijg}$ rappresenta il valore di espressione per il gene $g$ nel campione $j$ del batch *i*; $\alpha_g$ è il livello di espressione globale; $X$ è la matrice del disegno, la quale rappresenta il fattore biologico di interesse (es. malati vs sani), e $\beta_g$ è il vettore dei coefficienti di regressione corrispondenti a $X$. $\varepsilon_{ijg}$ è un termine di errore per il quale si può assumere una distribuzione Normale di media zero e varianza $\sigma_g^2$. I parametri $\gamma_{ig}$ e $\delta_{ig}$ rappresentano gli effetti, rispettivamente, additivo e moltiplicativo del batch *i* per il gene $g$.

E' evidente fin da subito come in questo modello si stia supponendo di conoscere esattamente il numero di batch presenti.

I tre passi necessari per l'aggiustamento dei dati sono qui sotto riportati.

***1) Standardizzazione dei valori di espressione***

La grandezza del valore di espressione può differire tra i geni a causa del vero livello di espressione e per la sensibilità della sonda (probe). Per questo i coefficienti $\alpha$, $\beta$, $\gamma$ e $\sigma^2$ differiscono tra i geni. Se non si tiene in considerazione che i parametri differiscono tra i geni, le stime Bayesiane empiriche della distribuzione a priori del batch effect vengono distorte e ciò riduce la quantità di

informazioni sull'effetto batch che si può stimare dai geni. Per evitare questo problema si standardizzano i dati in modo che i geni abbiano la stessa media e varianza. Si possono, quindi, individuare le stime dei parametri $\hat{\alpha}_g$, $\hat{\beta}_g$ e $\hat{\gamma}_{ig}$ per $i$ = 1,…,$m$ e $g$ = 1,…,$G$ con il metodo dei minimi quadrati, imponendo il vincolo $\sum_i n_i \hat{\gamma}_{ig} = 0$ $(g = 1, \dots, G)$ per assicurare l'identificabilità della soluzione, e stimare $\hat{\sigma}_g^2 = \frac{1}{N}\sum_{ij}\left(Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g - \hat{\gamma}_{ig}\right)^2$, dove $N$ è il numero totale di campioni. I valori di espressione standardizzati sono dunque:

$$Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g}{\hat{\sigma}_g} \tag{3.2}$$

***2) Stima dei parametri del batch effect utilizzando distribuzioni a priori empiriche***

Riguardo le nuove variabili standardizzate, è possibile affermare che $Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2)$ , dove i γ qui non sono gli stessi di quelli definiti in precedenza (infatti sarebbero $\gamma_{ig}/\sigma_g$). Si assumono le seguenti forme parametriche per le distribuzioni a priori:

$$\gamma_{ig} \sim N(\gamma_i, \tau_i^2) \quad \text{e} \quad \delta_{ig}^2 \sim Inv - Gamma(\lambda_i, \theta_i) \tag{3.3}$$

dove i parametri $\gamma_i$, $\tau_i^2$, $\lambda_i$e $\theta_i$ possono essere stimati empiricamente dai dati utilizzando il metodo dei momenti.

Le distribuzioni a priori sono state scelte perché coniugate alla distribuzione normale dei dati standardizzati e la loro adeguatezza deve perciò essere verificata. Nel caso le distribuzioni ipotizzate non fossero adeguate è stato sviluppato anche un metodo non parametrico.

Per ricavare le stime degli iper-parametri si parte dal ricavare la stima di $\gamma_{ig}$, ossia il valore atteso della variabile $Z_{ijg}$ indicante i valori di espressione standardizzati. Questa è semplicemente ottenibile come media dei valori di espressione per il $g$-esimo gene nell'$i$-esimo batch:

$$\hat{\gamma}_{ig} = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ijg} \tag{3.4}$$

Le stime degli iper-parametri $\gamma_i$ e $\tau_i^2$ della distribuzione a priori Normale per la media dei valori di espressione standardizzati dell'$i$-esimo batch vengono ricavate con il metodo dei momenti e sono:

$$\bar{\gamma}_i = \frac{1}{G}\sum_{g=1}^{G}\hat{\gamma}_{ig} \tag{3.5}$$

$$\bar{\tau}_i^2 = \frac{1}{G-1}\sum_{g=1}^{G}\left(\hat{\gamma}_{ig} - \bar{\gamma}_i\right)^2 \tag{3.6}$$

In pratica, la media dei valori di espressione standardizzati nell'*i*-esimo batch ($\gamma_{ig}$) segue una distribuzione Normale la cui media è stimata con il valore medio di espressione di tutti i geni degli array dello stesso batch e la cui varianza stimata è pari alla varianza campionaria corretta dei valori di espressione di quel batch.

Per ricavare le stime degli altri due iper-parametri si deve prima ricavare la stima di $\delta_{ig}^2$. Questa è:

$$\hat{\delta}_{ig}^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(Z_{ijg} - \hat{\gamma}_{ig}\right)^2 \tag{3.7}$$

ossia la varianza campionaria corretta del *g*-esimo gene nel batch *i*.

Similmente a quanto fatto in precedenza è possibile definire la media e la varianza campionaria della distribuzione a priori di $\hat{\delta}_{ig}^2$ come:

$$\bar{V}_i = \frac{1}{G}\sum_{g=1}^{G}\hat{\delta}_{ig}^2 \tag{3.8}$$

$$\bar{S}_i^2 = \frac{1}{G-1}\sum_{g=1}^{G}\left(\hat{\delta}_{ig}^2 - \bar{V}_i\right)^2 \tag{3.9}$$

Dato che, però, la distribuzione a priori per $\hat{\delta}_{ig}^2$ è Gamma Inversa, anziché Normale come nel caso precedente, bisogna eguagliare i momenti teorici di questa distribuzione con quelli empirici appena ricavati. Ciò porta alle stime:

$$\bar{\lambda}_i = \frac{\bar{V}_i + 2\cdot\bar{S}_i^2}{\bar{S}_i^2} \tag{3.10}$$

$$\bar{\theta}_i = \frac{\bar{V}_i^3 + \bar{V}_i\cdot\bar{S}_i^2}{\bar{S}_i^2} \tag{3.11}$$

Per trovare le stime a posteriori Empirical Bayes per i parametri relativi al batch effect è necessario trovare le distribuzioni a posteriori di $\gamma_{ig}$ e di $\delta_{ig}^2$. Riguardo la prima, si dimostra che:

$$\pi\left(\gamma_{ig}|Z_{ig},\delta_{ig}^2\right) \propto exp\left\{-\frac{1}{2}\left(\frac{n_i\tau_i^2+\delta_{ig}^2}{\delta_{ig}^2\tau_i^2}\right)\left[\gamma_{ig}^2 - 2\left(\frac{\tau_i^2\sum_j Z_{ijg}+\delta_{ig}^2\gamma_i}{n_i\tau_i^2+\delta_{ig}^2}\right)\gamma_{ig}\right]\right\} \tag{3.12}$$

e, notando che si tratta del kernel di una distribuzione Normale, si ricava:

$$\gamma_{ig}^* = \frac{n_i\bar{\tau}_i^2\hat{\gamma}_{ig} + \delta_{ig}^{2*}\bar{\gamma}_i}{n_i\bar{\tau}_i^2 + \delta_{ig}^{2*}} \tag{3.13}$$

Per la seconda stima, quella di $\delta_{ig}^2$, si dimostra che:

$$\pi\left(\delta_{ig}^2|Z_{ig},\gamma_{ig}\right) \propto \left(\delta_{ig}^2\right)^{-\left(\frac{n_i}{2}+\lambda_i\right)-1} exp\left\{-\frac{\theta_i + \frac{1}{2}\sum_j\left(Z_{ijg}-\gamma_{ig}\right)^2}{\delta_{ig}^2}\right\} \tag{3.14}$$

che è il kernel di una distribuzione Gamma Inversa. Per cui si ottiene la stima a posteriori:

$$\delta_{ig}^{2*} = \frac{\bar{\theta}_i + \frac{1}{2}\sum_j \left(z_{ijg} - \gamma_{ig}^*\right)^2}{\frac{n_j}{2} + \bar{\lambda}_i - 1} \tag{3.15}$$

Si noti che le soluzioni trovate, $\gamma_{ig}^*$ e $\delta_{ig}^{2*}$, dipendono ognuna dall'altra e non è possibile ottenerle in maniera indipendente. Per questo motivo le stime dovranno essere ottenute iterativamente.

***3) Aggiustamento dei dati per il batch effect***

Una volta calcolate le stime a posteriori dei parametri relativi al batch effect, $\gamma_{ig}^*$ e $\delta_{ig}^{2*}$, si possono aggiustare i dati.

$$Y_{ijg}^* = \hat{\sigma}_g \left(\frac{z_{ijg} - \gamma_{ig}^*}{\hat{\delta}_{ig}^*}\right) + \hat{\alpha}_g + X\hat{\beta}_g \tag{3.16}$$

# 3.2. Il metodo RUV

Il principale problema del metodo COMBAT è che richiede di essere a conoscenza di quali sono i batch all'interno dei campioni e, di conseguenza, di quale fattore li ha provocati. Non sempre, però, in applicazioni reali, si dispone di tali informazioni. Molto spesso non si è nemmeno a conoscenza dell'esistenza di un effetto batch. Per questo motivo sono stati proposti alcuni metodi che presumono ignota l'origine del batch effect, la maggior parte dei quali utilizza il metodo dell'analisi fattoriale (cfr. §A.1) per individuare la variabilità non biologica, incorporando nel modello i fattori relativi a effetti non biologici. Il problema principale nel far ciò deriva dalla difficoltà di separare la variabilità biologica da quella non biologica. Il metodo in esame è definito RUV, del quale esistono due versioni: la prima, detta RUV-2, differisce dalla seconda, RUV-4, per il numero di passi necessario per l'eliminazione delle variazioni non volute. Il punto focale di questo metodo, a prescindere della versione che si adopera, è l'utilizzo dei *geni di controllo* (control genes). Questi geni possono essere controlli positivi o negativi: *controlli negativi* sono quei geni per i quali è nota a priori l'assenza di associazione con il fattore di interesse; *controlli positivi* sono

invece i geni che notoriamente hanno un'associazione con il fattore in esame. Il metodo RUV utilizza i geni di controllo per identificare i fattori non di interesse, per poi includerli, come covariate, nella matrice del disegno di un modello di regressione.

Si assuma di disporre di *n* array, ognuno con *G* geni. Sia $y_{jg}$ il logaritmo del valore di espressione osservato per il *g*-esimo gene nel *j*-esimo array. Sia $Y$ la matrice *n*x*G* di questi valori $y_{jg}$. Allora il modello ipotizzato per $Y$ è il seguente:

$$Y_{nxG} = X_{nxp}\beta_{pxG} + Z_{nxq}\gamma_{qxG} + W_{nxk}\alpha_{kxG} + \epsilon_{nxG} \quad (3.17)$$

dove $X$ è la matrice osservata del disegno, nelle cui colonne sono presenti i fattori di interesse (es: sani/malati), $Z$ è una matrice che ha nelle colonne le covariate osservate (es: etnia) e $W$ è la matrice delle covariate non osservate (es: batch effect). Si noti che *k*, ossia il numero di covariate non osservate, è ignoto. $\beta$, $\gamma$ e $\alpha$ sono gli ignoti coefficienti che determinano l'influenza di uno specifico fattore su un determinato gene. Infine, $\epsilon$ rappresenta il termine d'errore, per il quale si assume la distribuzione $\epsilon_{ij} \sim N(0, \sigma_j^2)$.

Il termine $Z_{nxq}\gamma_{qxG}$ è opzionale; un ricercatore potrebbe non avere nessuna covariata, o potrebbe voler trattare le covariate come se fossero inosservate. Inoltre, per facilitare l'esposizione, si supponga di disporre di un solo fattore di interesse. Il modello può essere riscritto come:

$$Y_{nxG} = X_{nx1}\beta_{1xG} + W_{nxk}\alpha_{kxG} + \epsilon_{nxG} \quad (3.18)$$

## 3.2.1. RUV-2

Come detto in precedenza, il punto chiave del RUV è l'utilizzo dei geni di controllo. Nel RUV-2 questi vengono utilizzati per la stima della matrice di covariate non osservate. In particolare, ora, si utilizzeranno i controlli negativi, per i quali è nota l'assenza di associazione con il fattore in analisi. Questa assunzione può essere tradotta, nel modello precedente, come $\beta_g = 0$.

I controlli negativi possono essere, ad esempio, dei geni *housekeeping* oppure, negli array di Affymetrix, dei controlli aventi prefisso AFFX. Per geni *housekeeping* si intendono dei geni che, solitamente, hanno funzioni generiche e il

cui livello di espressione non varia tra le condizioni che si vogliono studiare, proprio a causa della loro generalità. Si tratta tipicamente di geni fondamentali per la vita stessa della cellula e che, pertanto, devono essere sempre presenti ed egualmente espressi in ogni condizione.

I controlli negativi vengono utilizzati nel metodo RUV per la stima della matrice $W$. Potendo infatti assumere che, per questi geni, i parametri $\beta$ siano pari a zero, si ottiene la formula:

$$Y_c = W\alpha_c + \varepsilon_c \tag{3.19}$$

dove *c* rappresenta l'insieme dei geni di controllo negativi.

Fatta questa premessa, il RUV-2 prevede due passi per ricavare i valori dei coefficienti $\beta$ del modello:

1) *stima di $W$ tramite un'analisi fattoriale su $Y_c$;*
2) *stima di $\beta$ tramite una regressione di $Y$ su $X$ e la stima di $W$.*

Più esplicitamente, indicando con $\widehat{\mathrm{W}}^{(\mathrm{RUV}-2)}$ la stima di $W$, la stima di $\beta$ è:

$$\hat{\beta}^{(RUV-2)} = (X^T R_{\widehat{W}^{(RUV-2)}} X)^{-1} X^T R_{\widehat{W}^{(RUV-2)}} Y \tag{3.20}$$

dove con $R_{\widehat{W}^{(RUV-2)}}$ si indica l'operatore residuale della matrice $\widehat{W}^{(RUV-2)}$, che rimuove l'effetto della componente stimata di $W$. Data una matrice $A$ si ha $R_A = I - A(A^T A)^{-1} A^T$.

Per la stima di $W$ si veda l'appendice (§A.1) nella quale si introduce il concetto di analisi fattoriale e si mostra come, in realtà, per ottenere tale stima sia sufficiente utilizzare una semplice decomposizione a valori singolari (SVD) sulla matrice dei geni di controllo negativi, $Y_c$.

In generale, il RUV-2 può portare alla stima della matrice di espressione corretta:

$$\hat{Y} = Y - \widehat{W}\hat{\alpha} \tag{3.21}$$

Un punto decisamente non banale e di fondamentale importanza è la decisione del rango della matrice $W$, definito in precedenza come *k*. Non esiste un algoritmo ottimale per il suo calcolo per cui ciò che può essere fatto è provarne diversi e controllare alcuni indicatori di bontà. Una discussione su questi indicatori è riportata dopo la spiegazione del RUV-4.

Si osservi anche che $W$ può essere stimata solo assumendo che i geni di controllo negativi non siano associati con il fattore di interesse, ma piuttosto con i

fattori ignoti che si vogliono stimare. La scelta del gruppo di geni di controllo è fondamentale per la buona riuscita del metodo. Una cattiva scelta dei geni di controllo porta ad una cattiva stima di $W$.

### 3.2.2. RUV-4

Il RUV-4 è più complesso del precedente e prevede quattro passi per ricavare i valori dei coefficienti β del modello:

***1) Stima e rimozione di X***

Si consideri il modello precedente (3.18); si ha $Y = X\beta + W\alpha + \epsilon$. Moltiplicando a destra e a sinistra per l'operatore residuale $R_X$ si ottiene:

$$R_X Y = R_X X\beta + R_X W\alpha + R_X\epsilon = W_0\alpha + R_X\epsilon \quad (3.22)$$

dove $W_0 = R_X W$

Questa operazione di moltiplicazione ha lo scopo di rimuovere l'effetto del fattore biologico di interesse, $X$, in modo da non rischiare di considerarlo nel seguito come fattore da rimuovere.

***2) Analisi fattoriale***

Si utilizzi qualche variante dell'analisi fattoriale per ottenere la stima $\widehat{W_0\alpha}$ di $W_0\alpha$. Inoltre si definiscano le stime individuali $\widehat{W_0}$ e $\hat{\alpha}$ in modo che $\widehat{W_0}\hat{\alpha} = \widehat{W_0\alpha}$. Malgrado non sia obbligatorio utilizzare uno specifico metodo per l'analisi fattoriale, ciò che si richiede è che la stima di $\alpha$ sia $\hat{\alpha} = \left(\widehat{W_0}^T\widehat{W_0}\right)^{-1}\widehat{W_0}^T Y$.

***3) Stima di W***

Per la stima di $W$, la si riscrive come:

$$\begin{aligned} W &= W - X(X^TX)^{-1}X^TW + X(X^TX)^{-1}X^TW = \\ &= (I - X(X^TX)^{-1}X^T)W + X[(X^TX)^{-1}X^TW] = \\ &= R_X W + Xb_{WX} = W_0 + Xb_{WX} \end{aligned} \quad (3.23)$$

dove $b_{WX}$ indica il coefficiente di regressione parziale di $W$ su $X$; $b_{WX} = [(X^TX)^{-1}X^TW]$.

$X$ è noto e si dispone di una stima di $W_0$, ottenuta al punto precedente tramite analisi fattoriale. Ciò che si vorrebbe stimare è $b_{WX}$. Per farlo bisogna utilizzare l'identità:

$$b_{Y_cX} = b_{Y_cX.W} + b_{WX}b_{Y_cW.X} \tag{3.24}$$

Il coefficiente di regressione parziale di $Y_c$ su $X$ è pari alla somma tra il coefficiente di regressione parziale di $Y_c$ su $X$ al netto di $W$ e il prodotto tra il coefficiente di regressione parziale tra $W$ e $X$ per il coefficiente di regressione parziale di $Y_c$ su $W$ al netto di $X$.

Assumendo che esista $(b_{Y_cW.X}b_{Y_cW.X}^T)^{-1}$ si può risolvere per $b_{WX}$:

$$b_{WX} = (b_{Y_cX} - b_{Y_cX.W})b_{Y_cW.X}^T(b_{Y_cW.X}b_{Y_cW.X}^T)^{-1} \tag{3.25}$$

Notando che:

$$b_{Y_cX.W} \approx \beta_{Y_cX.W} = \beta_c = 0 \tag{3.26}$$

e che:

$$b_{Y_cW.X} \approx \beta_{Y_cW.X} = \alpha_c \approx \hat{\alpha}_c \tag{3.27}$$

allora:

$$b_{WX} \approx b_{Y_cX}\hat{\alpha}_c^T(\hat{\alpha}_c\hat{\alpha}_c^T)^{-1} \tag{3.28}$$

Dove con β si è indicato il vero parametro relativo a *b*.

Per cui la stima di W cercata è:

$$\widehat{W} = \widehat{W}_0 + Xb_{Y_cX}\hat{\alpha}_c^T(\hat{\alpha}_c\hat{\alpha}_c^T)^{-1} \tag{3.29}$$

***4) Stima di β tramite una regressione di Y su X e la stima di W.***

Come per il metodo RUV-2 si può inserire $\widehat{W}$nel modello e definire:

$$\hat{\beta}^{(RUV-4)} = (X^T R_{\widehat{W}^{(RUV-4)}}X)^{-1}X^T R_{\widehat{W}^{(RUV-4)}}Y \tag{3.30}$$

La diversità tra RUV-2 e RUV-4 sta sostanzialmente nel metodo utilizzato per la stima della matrice $W$. Mentre in RUV-2 $W$ può essere stimata semplicemente utilizzando il sottoinsieme dei geni di controllo negativi, nel RUV-4 la stima avviene utilizzando l'intero set di geni osservati. E' naturale, però, che questa operazione comporta il rischio che l'analisi fattoriale stimi come varianza non voluta anche una parte o tutta la varianza biologica derivante dal fattore di interesse. Mentre nel RUV-2 ciò non può accadere proprio perché i controlli negativi non hanno alcuna associazione con il fattore di interesse, nel RUV-4 è

necessario tutelarsi da questa evenienza. Ciò viene fatto rimuovendo prima la varianza connessa al fattore di interesse proiettando i dati nel complemento ortogonale dello spazio delle colonne di $X$. Solo dopo questa operazione si può procedere con l'analisi fattoriale. D'altro canto, però, svolgere un'analisi fattoriale su $R_X Y$ non produce una stima accurata di W, ma produce la stima di $W_0 = R_X W$. E' dunque necessario recuperare la parte di W che era stata eliminata dalla proiezione fatta al primo passo. Qui entrano in gioco i geni di controllo negativi, i quali soddisfano le assunzioni utilizzate nel terzo step, ossia $b_{Y_c X.W} \approx 0$, $b_{Y_c W.X} \approx \alpha_c$ e $\hat{\alpha}_c \approx \alpha_c$. Riguardo la prima assunzione, l'interpretazione esatta è sottile. $b_{Y_c X.W}$ è il coefficiente di regressione parziale di $Y_c$ su $X$ al netto di $W$, cioè è la stima di $\beta_c$ che si avrebbe in una regressione di $Y_c$ su $X$ e $W$ se il vero $W$ fosse noto. Dato che, per i controlli negativi, vale l'assunzione $\beta_c = 0$, essendo $b_{Y_c X.W}$ la stima di $\beta_c$, allora si può concludere che $b_{Y_c X.W} \approx 0$.

La seconda assunzione è del tutto simile alla prima, mentre la terza assunzione afferma sostanzialmente che la stima di $\alpha_c$ ottenuta al passo 2 tramite l'analisi fattoriale è una buona stima di $\alpha_c$.

Una interessante riformulazione per il calcolo di $\hat{\beta}^{(RUV-4)}$ è la seguente:

$$\hat{\beta}^{(RUV-4)} = (X^T X)^{-1} X^T (Y - \widehat{W}\hat{\alpha}) \qquad (3.31)$$

Anche in questo metodo è necessaia la stima del corretto *k* che indica la dimensione della matrice $W$. E' stato dimostrato che il RUV-4 può funzionare meglio del RUV-2 almeno in due casi: quando i geni di controllo negativi sono mal specificati, ossia alcuni di questi non sono veramente non associati con il fattore di interesse, e quando il parametro *k* è sovrastimato.

## 3.3. Valutazione della bontà di un metodo di aggiustamento

E' necessario avere a disposizione delle tecniche per valutare la qualità di un metodo di aggiustamento per la rimozione della variabilità non voluta. Esistono due buoni criteri per la verifica della bontà del metodo di aggiustamento utilizzato, ad ognuno dei quali è dedicato un breve paragrafo nel prosieguo del

trattato; altri criteri sono invece empirici e non sempre sono tutti verificati ed è, dunque, il ricercatore a doverli interpretare. Questi ultimi sono illustrati nel §3.3.3.

### 3.3.1. Geni di controllo positivi

Dato che i controlli positivi hanno certamente un'associazione con il fenomeno in esame, possono essere utilizzati per il controllo della bontà del metodo di aggiustamento. In particolare, per ogni gene, può essere calcolato il corrispettivo p-value e questi possono essere ordinati in ordine crescente. I controlli positivi dovrebbero risultare all'inizio di questa lista. Una possibile misura per la valutazione della bontà potrebbe essere la percentuale di controlli positivi che compaiono alle prime posizioni, ad esempio tra i primi 50 geni della lista ordinata. Se un metodo di aggiustamento fa aumentare questa percentuale si ha ragione di credere che il metodo sia efficace. E' interessante notare che si preferisce utilizzare il rank dei p-value anziché i p-value stessi. Questo perché, a seguito dell'aggiustamento, i p-value dei geni di controllo possono diminuire o anche aumentare in base alla natura della variazione non voluta.

### 3.3.2. Distribuzione dei p-value

In uno studio sulla differenziale espressione si assume che ci sia associazione tra il fenomeno di interesse e il livello di espressione solo per una piccola frazione di geni. La distribuzione dei p-value per i geni che non sono associati con il fattore di interesse sarà, dunque, idealmente uniformemente distribuita nell'intervallo unitario. Per questo un istogramma di tutti i p-value dovrebbe essere uniforme, con un picco di valori bassi per i p-value dei geni associati con il fattore di interesse. La varianza non voluta ha l'effetto di introdurre dipendenza spuria tra i livelli di espressione misurati; dato che un metodo di aggiustamento dovrebbe rimuovere questa dipendenza, ci si può aspettare che un buon metodo di produca un istogramma dei p-value più vicino a quello ideale.

### 3.3.3. Altri metodi empirici

Altri metodi per la valutazione della bontà dell'aggiustamento sono più empirici e possono non essere buoni indicatori se considerati singolarmente; per questo devono essere valutati complessivamente dal ricercatore. Questi sono:

- miglioramento nei grafici RLE (cfr. §4.3.1);
- identificazione di gruppi omogenei attraverso una cluster analysis;
- diminuzione delle varianze dei geni di controllo negativi;
- in caso di indipendenza tra fattore biologico di interesse e fattore che provoca il batch effect, ad ogni livello di FDR un numero maggiore di geni è chiamato come differenzialmente espresso.

# Capitolo 4

## 4. Applicazione a dati reali

In questo capitolo si analizzeranno dei dati di pazienti affette da tumore all'ovaio. Il primo paragrafo sarà dedicato ad un breve approfondimento su questa patologia, mettendone in evidenza i vari tipi e livelli di gravità. Si passerà poi ad una descrizione dei dati che si vogliono analizzare, mostrando quali caratteristiche sono state rilevate nelle pazienti e mettendo in evidenza l'obiettivo delle analisi. La fase di analisi vera e propria sarà divisa in due parti. Nella prima l'obiettivo sarà quello di ricercare il miglior metodo di rimozione del batch effect provocato dal fatto che si dispone di dati ottenuti con due piattaforme differenti, come verrà meglio spiegato nel §4.2; in questa fase verrà utilizzato solo un sottoinsieme dei dati di cui si dispone. Nella seconda parte, invece, verranno utilizzati i risultati ottenuti nella prima per la correzione del batch effect e, poi, si procederà ad analizzare tutti i campioni di cui si dispone per identificare quali geni si comportano in maniera significativamente diversa tra le pazienti caratterizzate da un'alta sopravvivenza globale e quelle che, invece, hanno una sopravvivenza bassa. Per sopravvivenza globale si intende il numero di mesi di sopravvivenza tra il momento della diagnosi del tumore e il momento del decesso o di fine del follow-up, avvenuto in data 1/05/2013.

## 4.1. Il tumore all'ovaio

Le ovaie sono due organi situati uno a destra e uno a sinistra dell'utero e sono connesse a quest'ultimo tramite le tube di Falloppio. Le loro funzioni sono sostanzialmente due: (i) produrre ormoni sessuali femminili e (ii) produrre gli ovociti, ossia cellule riproduttive femminili.

Il tumore dell'ovaio si sviluppa quando le cellule di quest'organo iniziano a moltiplicarsi in modo anomalo, perdendo qualsiasi meccanismo di controllo della loro crescita.

Il termine tumore all'ovaio può indurre a credere che si tratti di una sola malattia che si manifesta sempre nello stesso modo ma in realtà non è così. Già da tempo è noto che esiste una grande divisione tra tumori in base alla zona dalla quale essi si originano; dando origine ai tre tipi tumorali: *epiteliale*, *germinale* e *stromale*. I tumori epiteliali costituiscono circa il 90% di tutti i tumori ovarici e sono così definiti perché hanno origine nelle cellule dell'epitelio ovarico, cioè il tessuto che riveste superficialmente l'ovaio stesso. I tumori germinali originano, invece, dalle cellule germinali, che danno origine agli ovociti. I tumori stromali, infine, derivano dallo stroma gonadico, ossia il tessuto di sostegno dell'ovaio. Queste ultime due categorie sono le più rare e comprendono circa il 10% di tutti i tumori ovarici.

Come riportato in studi recenti (Oberaigner, et al., 2012) la sopravvivenza per il tumore all'ovaio è la più bassa tra tutti i tipi di tumore ginecologici. Il progetto EUROCARE-4, che ha prodotto un database di pazienti di 23 paesi europei alle quali è stato diagnosticato un tumore all'ovaio tra il 1978 e il 2002, riporta una sopravvivenza relativa a cinque anni pari al 36%. La ragione di ciò è la mancanza di un metodo per l'individuazione precoce di questo tipo di tumore, infatti più dei due terzi dei tumori vengono diagnosticati in stadi avanzati.

L'Organizzazione Mondiale della Sanità (OMS) classifica i tumori ovarici secondo sei isotipi principali: *sieroso*, *mucinoso*, *endometrioide*, *a cellule chiare*, *a cellule transizionali* e *squamoso*.

In base alla gravità e alla proliferazione, i tumori sono classificati in quattro stadi secondo il sistema FIGO (Federazione Internazionale di Ginecologia e Ostetricia). Capire lo stadio del tumore è essenziale per programmare il trattamento più appropriato. Gli stadi FIGO sono riportati nella Tabella 4.1.

Circa l'80-85% di tutti i carcinomi ovarici nei Paesi occidentali sono sierosi (Przybycin & Soslow, 2011). Questi sono generalmente rappresentati da uno stadio FIGO alto, III o IV; poche volte si presenta lo stadio II, mentre è molto raro il riscontro di un carcinoma sieroso allo stadio FIGO I.

| | | |
|---|---|---|
| **Stadio I** | | **Tumore limitato alle ovaie** |
| | **IA** | Tumore limitato ad un ovaio; assenza di ascite; capsula integra, assenza di tumore sulla superficie |
| | **IB** | Tumore limitato alle due ovaie, assenza di ascite, capsula integra, assenza di tumore sulla superficie |
| | **IC** | Tumore allo stadio IA o IB, ma con tumore sulla superficie di una o entrambe le ovaie o con capsula rotta o con citologia del liquido ascitico positiva o con washing peritoneale positivo |
| **Stadio II** | | **Tumore che interessa una o entrambe le ovaie con estensione pelvica** |
| | **IIA** | Estensione e/o metastasi all'utero e/o alle tube |
| | **IIB** | Estensione ad altri tessuti pelvici |
| | **IIC** | Tumore allo stadio IIA o IIB, ma con tumore sulla superficie di una o entrambe le ovaie o con capsula rotta o con citologia del liquido ascitico positiva o con washing peritoneale positivo |
| **Stadio III** | | **Tumore interessante una o entrambe le ovaie con metastasi peritoneali extrapelviche e/o linfonodi retroperitoneali o inguinali positivi, metastasi sulla superficie epatica, tumore limitato alla pelvi ma con dimostrazione istologica di metastasi all'omento e/o al piccolo intestino** |
| | **IIIA** | Tumore limitato alla piccola pelvi, linfonodi negativi, diffusione microscopica istologicamente confermata alla superficie peritoneale addominale |
| | **IIIB** | Tumore allo stadio IIIA con disseminazione alla superficie peritoneale addominale non eccedente i 2 cm |
| | **IIIC** | Metastasi addominali di diametro > 2 cm e/o linfonodi retroperitoneali o inguinali positivi |
| **Stadio IV** | | **Metastasi a distanza. Versamento pleurico con citologia positiva. Metastasi epatiche parenchimali** |

**Tabella 4.1.Stadiazione del tumore all'ovaio secondo il metodo FIGO.**

Oltre allo stadio, un'altra importante caratteristica del tumore è il grado. Questo parametro indica quanto le cellule tumorali si differenziano da quelle del tessuto sano e dà un'idea della velocità con cui il tumore si sviluppa. Si possono distinguere tre gradi:

- G1 o grado basso: le cellule del tumore sono molto simili a quelle del tessuto ovarico sano, crescono lentamente e di solito non si diffondono nel tessuto circostante;

- G2 o grado medio: le cellule sono simili a quelle dei tumori di basso grado. Esse sono cancerose ma non hanno ancora invaso e danneggiato i tessuti sani circostanti;
- G3 o grado alto: le cellule hanno un aspetto anomalo, crescono con grande rapidità e hanno un'elevata probabilità di diffondersi nei tessuti circostanti e nel sangue.

Riguardo il tumore sieroso, si sta diffondendo l'idea che quello di grado alto e quello di grado basso, più che rappresentare gradi di gravità opposti, rappresentino due tipi diversi di tumore.

## 4.2. I dati e gli obiettivi

I dati di cui si dispone sono 44 esperimenti di microarray riguardanti cellule di tessuti cancerosi di donne affette da tumore all'ovaio di tipo sieroso e di grado medio o alto. Questi sono stati ottenuti con due piattaforme differenti: *GeneChip Human Genome U133A* e *GeneChip Human Genome U133 Plus 2.0*. Nel seguito le due piattaforme verranno abbreviate con *U133A* e *U133 plus 2*. Tra i 44 esperimenti, 5 sono duplicati e, dunque, sono disponibili per entrambe le piattaforme. Ciò significa che il numero totale di campioni per i quali si dispone di valori di espressione è 39. La piattaforma U133A è in grado di misurare l'espressione di un numero totale di 12.098 geni; la piattaforma U133 plus 2, invece, misura 18.960 geni.

Per quanto discusso nel capitolo 3, l'utilizzo di dati provenienti da piattaforme diverse senza nessun genere di aggiustamento non è corretto, dato che i valori di espressione sono distorti dal batch effect e da altri fattori non biologici. Per questo motivo si vogliono utilizzare le 5 osservazioni duplicate e disponibili in entrambe le piattaforme per stimare l'effetto batch con i tre metodi esposti in quel capitolo (COMBAT, RUV-2 e RUV-4) e verificare la bontà degli stessi per decretare quale sia il metodo migliore. Come spiegato nel §3.3, esistono alcuni criteri per la valutazione della bontà del metodo di aggiustamento utilizzato. In questo caso, però, si dispone di un metodo diverso per fare tale valutazione: controllare, tra i soli 5 esperimenti ripetuti, il comportamento di una cluster

analysis prima e dopo la correzione dei valori di espressione. Il risultato che ci si attende è che, dopo la correzione, la cluster analysis identifichi cinque gruppi, ognuno dei quali formato dai due esperimenti ripetuti nelle due diverse piattaforme. Nei dati non aggiustati, invece, ci si attende che la cluster analysis identifichi principalmente due gruppi, ognuno dei quali formato dalle osservazioni della stessa piattaforma. Questo perché nei dati non aggiustati la maggiore fonte di variabilità presente è quella dovuta al fattore piattaforma (non biologico) che determina l'effetto batch; nei dati aggiustati, invece, se tale fattore è stato rimosso adeguatamente, la maggior fonte di variabilità presente è quella biologica. Nel caso in esame, dato che ci sono esperimenti ripetuti, quelli più vicini tra loro devono essere le coppie formate dallo stesso esperimento svolto su due piattaforme differenti.

Capito qual è il metodo migliore per la rimozione del batch effect, si vuole utilizzarlo per correggere i valori di espressione dei 39 esperimenti e poterli, dunque, trattare come se fossero tutti provenienti da uno stesso batch e privi di effetti provocati da fattori non biologici. Resta, per ora, in sospeso quali delle 5 osservazioni ripetute dovranno essere utilizzate nelle analisi dell'intero set di dati a disposizione. Tale scelta sarà fatta in seguito, in base anche ad una verifica della qualità dei campioni ottenuti.

Oltre a quanto detto, si dispone di alcune informazioni cliniche relative ad ognuno degli esperimenti. Queste sono riassunte in Tabella 4.2.

L'obiettivo di questa analisi, come accennato all'inizio di questo capitolo, è quello di individuare quali geni sono differenzialmente espressi tra le donne che sono caratterizzate da un'alta sopravvivenza globale e quelle che, invece, hanno una sopravvivenza bassa. La sopravvivenza è stata misurata in mesi di vita prima dell'evento di decesso o del termine del follow-up delle pazienti.

E' doveroso sottolineare che la selezione dei campioni non è avvenuta in maniera casuale ma sono state appositamente scelte pazienti con sopravvivenza alta e pazienti con sopravvivenza bassa cercando di separare il più possibile i due gruppi, in modo da non dover stabilire una soglia specifica per la decisione di appartenenza alle classi e sperando così di ottenere risultati migliori.

| *Variabile* | *Descrizione* |
|---|---|
| **Sopravvivenza** | Variabile quantitativa indicante la sopravvivenza, in mesi, della paziente. Per gli scopi dell'analisi verrà trattata come qualitativa con due livelli: *alta*, *bassa*. |
| **Stadio FIGO** | Stadiazione secondo il metodo FIGO. I livelli osservati sono i seguenti:<br>*IB, IIB, IIC, IIIC, IV*. |
| **Grado** | Grado del tumore, con due modalità osservate: *G2* o *G3*. |
| **Data nascita** | Data di nascita della paziente. |
| **Età diagnosi** | Età della paziente al momento della diagnosi. |
| **Menopausa** | Variabile indicante lo stato di menopausa o meno della donna al momento della diagnosi. Le modalità sono:<br>- *pre:* la donna non è in menopausa;<br>- *post:* la donna è in menopausa. |
| **Dimensione del tumore residuo** | Variabile qualitativa relativa alla dimensione, in cm, del tumore residuo dopo l'operazione, definito come massima dimensione del nodulo di maggior volume, evidenziabile alla fine della procedura chirurgica. I livelli sono i seguenti:<br>- $TR = 0$;<br>- $0 < TR \leq 1$;<br>- $TR > 1$. |
| **CA125 preoperatorio** | Livello rilevato nel sangue di CA125 prima dell'operazione. Il CA125 è un marker tumorale: per molti tipi di tumori ginecologici si rileva un rialzo del livello di CA125. |
| **Asciti** | Presenza di ascite, ossia accumulo di liquido nella cavità addominale. Livelli: *si* / *no*. |
| **Linfonodi** | Variabile qualitativa indicante se nei linfonodi sono presenti cellule cancerose o meno. La diffusione di cellule tumorali ai linfonodi, infatti, è simbolo di avanzamento del tumore. I livelli della variabile sono: *positivi*, *negativi*. |

| | |
|---|---|
| **Risposta prima linea di chemioterapia** | Tipo di risposta alla prima linea di chemioterapia. Le possibili risposte sono:<br>- *clinica completa:* se il CA125 dopo la prima linea di chemioterapia è minore della metà del CA125 iniziale;<br>- *clinica parziale:* se il CA125 dopo la prima linea di chemioterapia è minore di quello iniziale, ma non minore della metà;<br>- *non risposta:* se il CA125 dopo la prima linea di chemioterapia non è diminuito. |
| **Risposta chemioterapia al platino** | Tipologia di reazione alla chemioterapia al platino. Le possibili reazioni sono: *platino sensibile*, ossia sensibile alla chemioterapia al platino, *platino parzialmente sensibile*, ossia parzialmente sensibile alla chemioterapia al platino, *platino resistente*, ossia che la paziente è diventata resistente alla chemioterapia al platino e *platino refrattaria*, ossia che la paziente non è mai stata sensibile alla chemioterapia al platino. |
| **Recidiva** | Variabile qualitativa indicante se la paziente è recidiva nel tumore, ossia se dopo averlo asportato si è ripresentato. I livelli sono: *si*, *no* e *prog*. Quest'ultimo sta ad indicare la progressione della malattia, ossia che non c'è stata alcuna risposta al trattamento chemioterapico e quindi la paziente non è mai stata libera da malattia. |
| **Data prima recidiva** | Data della prima recidiva, se questa c'è stata. |

**Tabella 4.2: Informazioni cliniche disponibili per ciascun esperimento effettuato.**

## 4.3. Ricerca del miglior metodo di rimozione del batch effect

In questo paragrafo si cercherà di individuare quale tra i tre metodi esposti in questa tesi per la correzione del batch effect sia il migliore per i dati osservati. Per fare ciò sarà necessario, innanzitutto, svolgere alcune analisi esplorative per verificare la qualità degli esperimenti effettuati. Fatto questo si procederà ad ottenere le stime del batch effect con i tre metodi COMBAT, RUV-2 e RUV-4 e, infine, si spiegheranno i risultati ottenuti, individuando anche il miglior metodo di

correzione. Tale metodo sarà poi utilizzato nell'analisi finale di tutti i 39 campioni a disposizione.

### 4.3.1. Analisi esplorative

Negli esperimenti di microarray è prassi comune effettuare delle analisi esplorative sui dati osservati, le quali sono utili per verificare la buona riuscita degli esperimenti. Per ora ci si limiterà a svolgere queste analisi solo sui 5 esperimenti ripetuti, ossia su 10 array in totale.

Le analisi saranno condotte in parallelo sui dati derivanti dalla piattaforma U133A e su quelli della piattaforma U133 plus 2.

Una delle prime analisi che si è soliti effettuare è per verificare la bontà del processo di ibridazione, ossia per verificare se ci sono aree del vetrino nelle quali l'ibridazione non è avvenuta correttamente. Potrebbe essere, infatti, che in alcune zone l'ibridazione sia sistematicamente più alta o più bassa che nel resto dell'array. Queste valutazioni vengono fatte ricavando i valori di espressione senza effettuare alcuna normalizzazione né eliminare il background e creando delle immagini che mostrino i valori di espressione con gradazioni di colore. Si noti che la procedura per ricavare i valori di espressione è quella del metodo RMA esposto al §2.4.1.

Le immagini per l'analisi della bontà del processo di ibridazione per la piattaforma U133A sono riportate in Figura 4.1. Le immagini a sinistra mostrano tutti i valori ricavati per ogni probe come gradazioni tra grigio e marrone. Le immagini centrali rappresentano i residui positivi del modello previsto dal metodo RMA e quelle di destra rappresentano i residui negativi.

Si noti come nel secondo campione (OSPC29) vi sia un'area, nel basso del vetrino, nella quale l'ibridazione è avvenuta in maniera leggermente diversa dal resto dell'array. Malgrado ciò, non si può certo affermare che gli esperimenti non siano riusciti bene dato che, per il secondo campione, l'area in questione è abbastanza ridotta e con intensità non eccessivamente alte.

In Figura 4.2 sono riportate le immagini relative alla piattaforma U133 plus 2. E' evidente il problema di ibridazione nel terzo campione (OSPC29), dove

un'area abbastanza grande del vetrino ha valori di espressione sistematicamente bassi e l'altra area ha valori alti. Ciò è sintomo di una qualità non molto buona dell'esperimento e, malgrado le tecniche di rimozione del background e quelle di normalizzazione siano utili per risolvere questo tipo di problemi, bisognerà prestare una particolare attenzione al comportamento dello stesso nelle analisi successive.

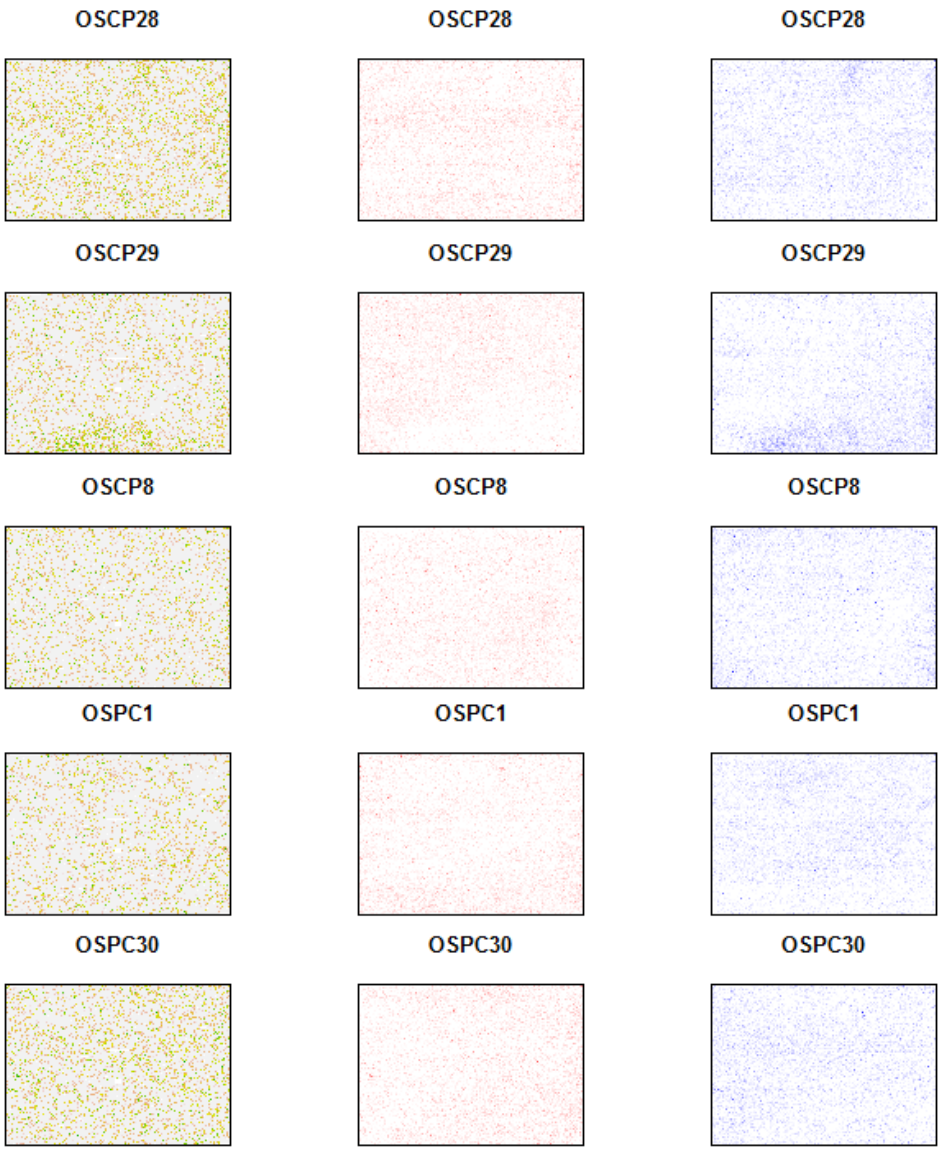



**Figura 4.1:Analisi della bontà del processo di ibridazione nella piattaforma U133A. Le immagini a sinistra mostrano tutti i valori ricavati per ogni probe come gradazioni tra grigio e marrone. Le immagini centrali rappresentano i residui positivi del modello previsto dal metodo RMA e quelle di destra rappresentano i residui negativi.**

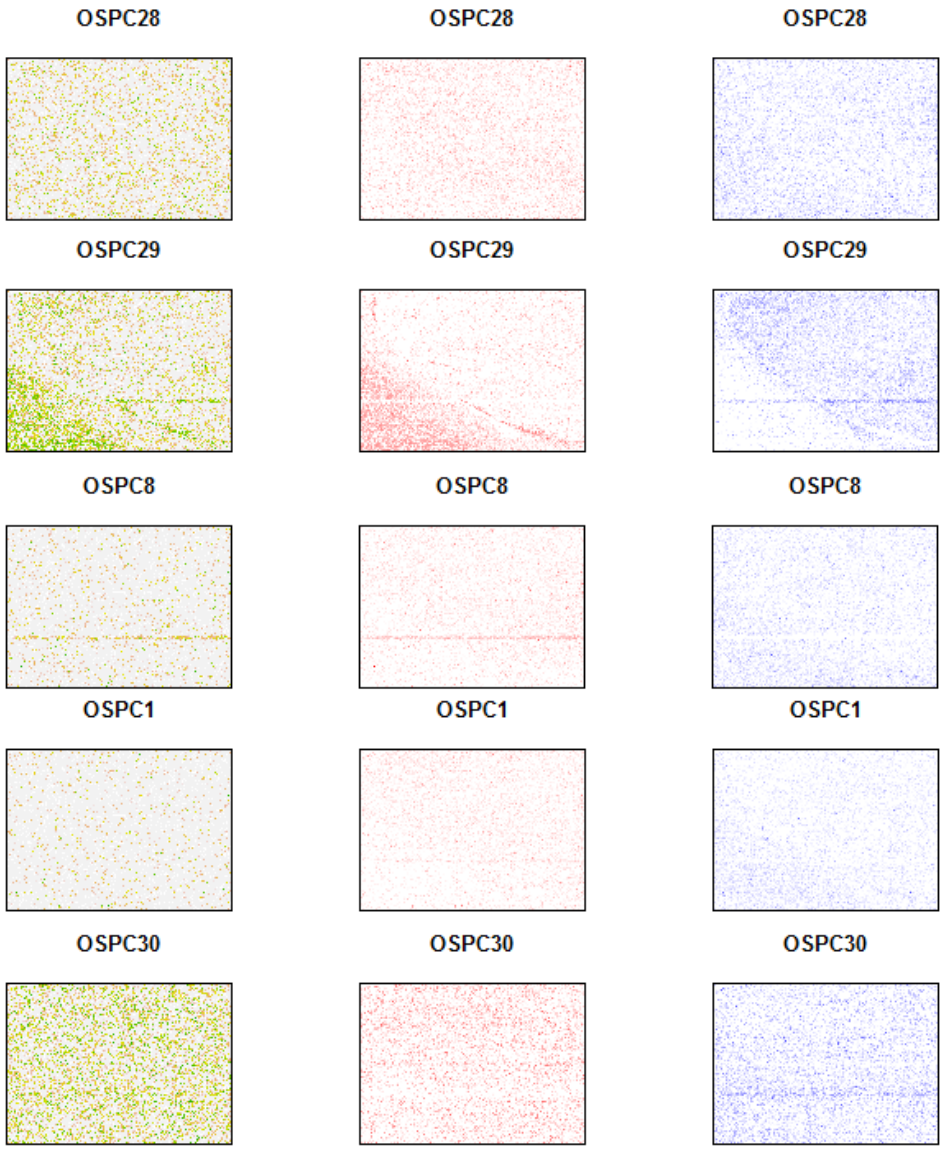



**Figura 4.2: Analisi della bontà del processo di ibridazione nella piattaforma U133 plus 2. Le immagini a sinistra mostrano tutti i valori ricavati per ogni probe come gradazioni tra grigio e marrone. Le immagini centrali rappresentano i residui positivi del modello previsto dal metodo RMA e quelle di destra rappresentano i residui negativi.**

Altro grafico importantissimo per valutare la bontà degli esperimenti è il grafico di degradazione dell'RNA. Quando si estrae l'RNA dalle cellule di un tessuto, questo si degrada molto velocemente. Per questo motivo deve passare pochissimo tempo da quando l'RNA viene estratto a quando lo si mette in un freezer a -80°C per mantenerne intatte le proprietà. La degradazione avviene secondo la direzione 5'-3'.

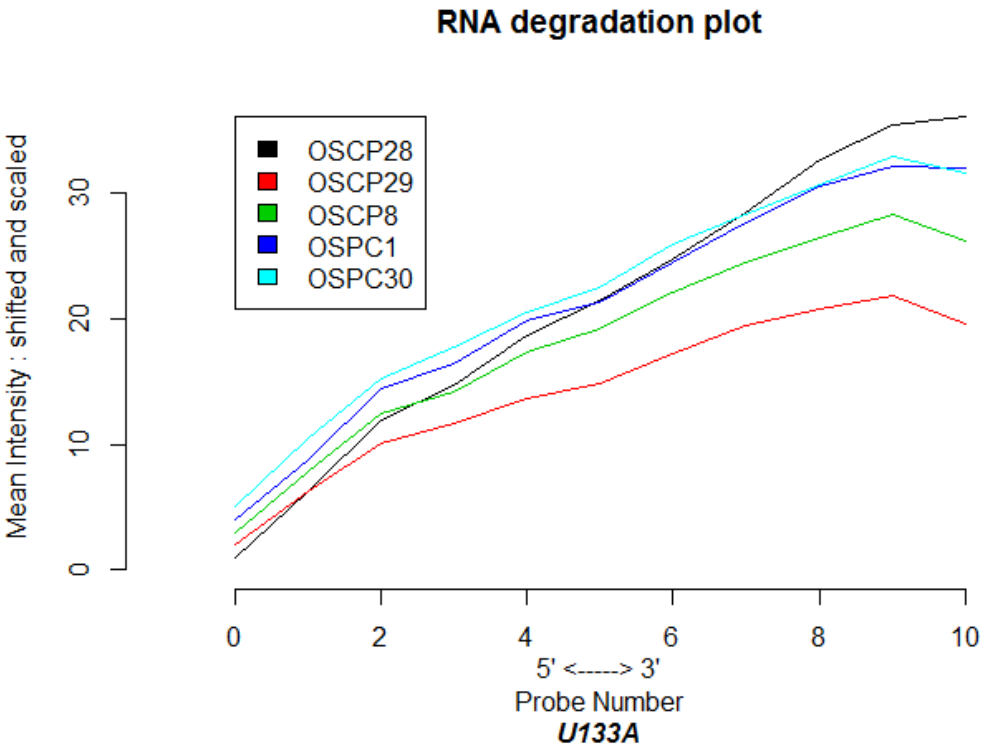



**Figura 4.3: Grafico di degradazione dell'RNA per la piattaforma U133A. Ogni spezzata è relativa ad uno dei 5 campioni. Il primo valore di una spezzata equivale al valore medio delle log-intensità di tutti i primi probe (quelli al 5') di quel campione, adeguatamente ricentrato e riscalato.**

Per valutare il livello di degradazione si calcolano i valori medi di log-intensità di ogni probe. Dato che i probe di un probeset sono costruiti su porzioni diverse del gene (dal 5'-3'), si costruisce un grafico nel quale si mostrano le medie dei valori delle log-intensità di tutti i probe dal 5' al 3'. Queste, prima di essere rappresentate, vengono riposizionate e riscalate per ottenere un grafico in cui il primo valore della prima spezzata sia centrato nell'origine. Ciò che ci si aspetta è che le intensità medie dei probe al 5' siano più basse perché l'RNA di questi probe è degradato, mentre al 3' siano più alte. I grafici di degradazione dell'RNA sono mostrati in Figura 4.3 (piattaforma U133A) e Figura 4.4 (piattaforma U133 plus 2).

Si nota che nei campioni ottenuti con la piattaforma U133A la degradazione dell'RNA è avvenuta in modo simile in tutti gli array; nella piattaforma U133 plus 2, invece, l'RNA si è degradato in modo molto disomogeneo tra gli array: il primo e l'ultimo campione hanno una degradazione simile tra loro e proprio analoga alla

situazione che ci si attende normalmente, mentre gli altri tre si comportano in maniera diversa. In particolare, nel campione OSPC30, l'RNA si è degradato notevolmente anche al 3' e ciò significa che la qualità dell'esperimento è bassa. Anche di questo, dunque, dovrà essere controllato il comportamento nelle successive analisi.

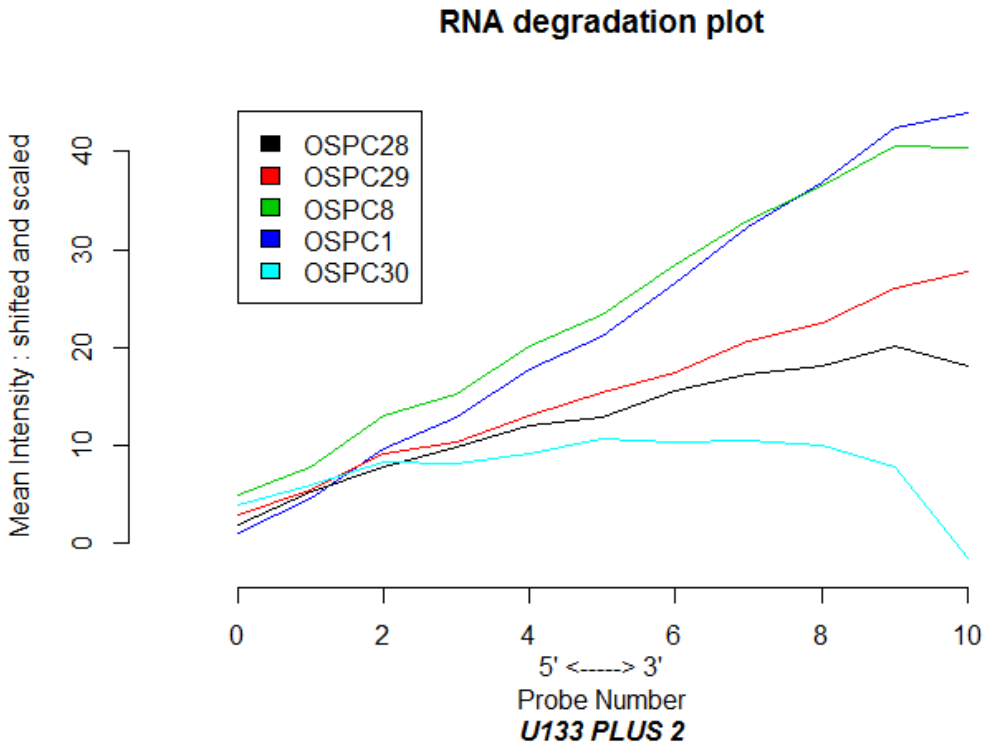



**Figura 4.4: Grafico di degradazione dell'RNA per la piattaforma U133A. Ogni spezzata è relativa ad uno dei 5 campioni. Il primo valore di una spezzata equivale al valore medio delle log-intensità di tutti i primi probe (quelli al 5') di quel campione, adeguatamente ricentrato e riscalato.**

Altri due grafici che si è soliti analizzare in questo tipo di studi sono riguardanti la distribuzione dei valori di espressione. Questa viene visualizzata tramite dei boxplot e dei density plot. In Figura 4.5 sono riportati quelli relativi alla piattaforma U133A e in Figura 4.6 quelli della piattaforma U133 plus 2. I grafici confermano la non perfetta riuscita di alcuni esperimenti nella piattaforma U133 plus 2 e, invece, la buona riuscita degli esperimenti fatti con la piattaforma U133A.

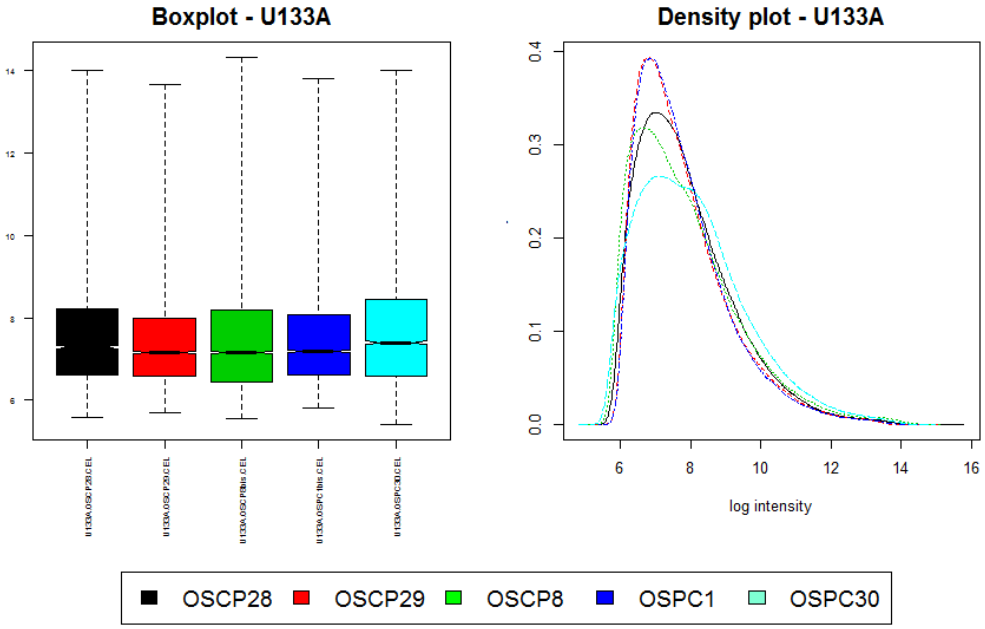



**Figura 4.5: Boxplot e Density plot dei valori di espressione nei 5 esperimenti con piattaforma U133A. Dati non normalizzati.**

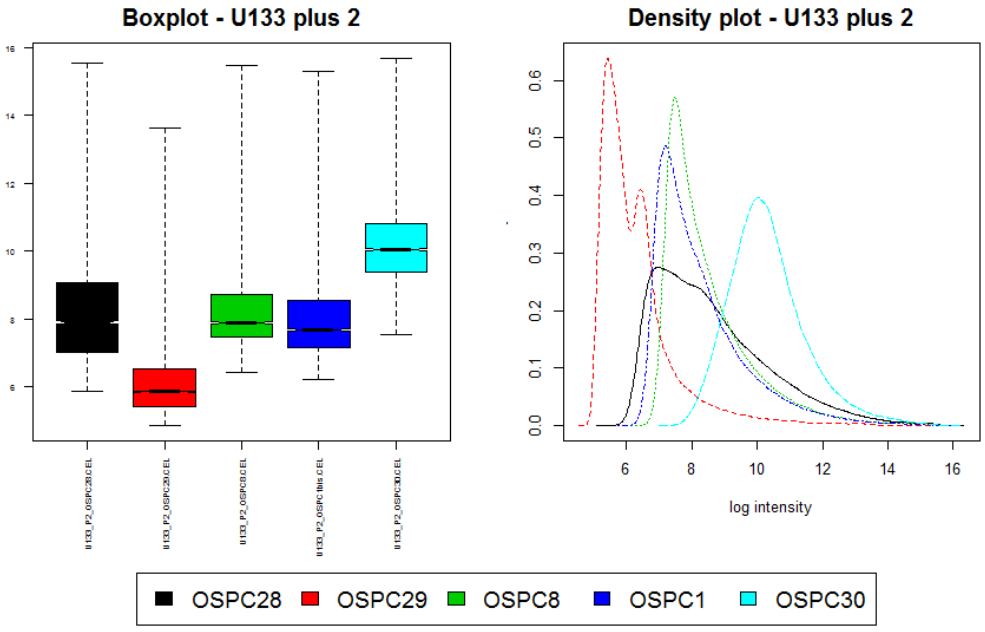



**Figura 4.6: Boxplot e Density plot dei valori di espressione nei 5 esperimenti con piattaforma U133 plus 2. Dati non normalizzati.**

Altri due grafici utili non solo per valutare la bontà degli esperimenti, ma anche per valutare la buona riuscita di una normalizzazione, sono il grafico RLE (Relative Log Expression) e il grafico NUSE (Normalized Unscaled Standard Errors). I valori del grafico RLE vengono ricavati per ogni probeset confrontando il valore di espressione di ogni array con il valore di espressione mediano per quel probeset tra gli esperimenti. Dato che si assume che la maggior parte dei geni non abbia variazione di espressione tra gli array, allora la maggior parte dei valori RLE dovrebbe essere attorno allo zero. Il grafico RLE prevede di effettuare un boxplot dei valori e analizzarne, dunque, la distribuzione.

Per la costruzione del grafico NUSE, invece, le stime degli standard error ottenute per ogni gene di ogni array vengono standardizzate tra gli array in modo che lo standard error mediano per quei geni sia pari a uno. Array nei quali, rispetto agli altri, ci sono molti standard error elevati, sono da considerarsi di qualità bassa.

I grafici RLE e NUSE dei dati non normalizzati sono riportati in Figura 4.7 e 4.8, rispettivamente. Si noti che nel grafico RLE di destra della Figura 4.7, relativo alla piattaforma U133 plus 2, la scala dell'asse *y* è diversa da quella del grafico di sinistra; ciò a conferma del fatto che alcuni esperimenti in quella piattaforma non hanno una buona qualità.

L'aver rilevato una scarsa qualità di alcuni campioni potrebbe rappresentare un problema nelle prossime analisi. Malgrado ciò, la valutazione finale della bontà del metodo di correzione del batch effect sarà incentrata prevalentemente sulla verifica, tramite cluster analysis, della creazione di gruppi contenenti le esatte coppie relative allo stesso campione di tessuto nelle due piattaforme diverse e, dunque, permetterà di controllare quali campioni risultano correttamente appaiati e quali no. In pratica, a differenza della maggior parte degli altri casi in cui si valuta la bontà di un metodo di aggiustamento, si conosce già qual è la vera suddivisione che si dovrebbe ottenere se il batch effect fosse eliminato e rimanesse solo il fattore biologico di interesse. Questa suddivisione consisterebbe nella creazione di 5 gruppi, uno con i due campioni OSPC28 delle due diverse piattaforme, un altro con i due campioni OSPC29, e via così. La presenza di campioni di scarsa qualità, quindi, non rappresenta un problema in quanto si potrà verificare se i metodi di aggiustamento riescono comunque a correggere i difetti

dovuti alle variazioni non biologiche in modo che una cluster analysis identifichi i gruppi di campioni correttamente. Se ciò non dovesse accadere, si procederà a verificare se i campioni non correttamente appaiati sono quelli caratterizzati da una scarsa qualità o anche gli altri.

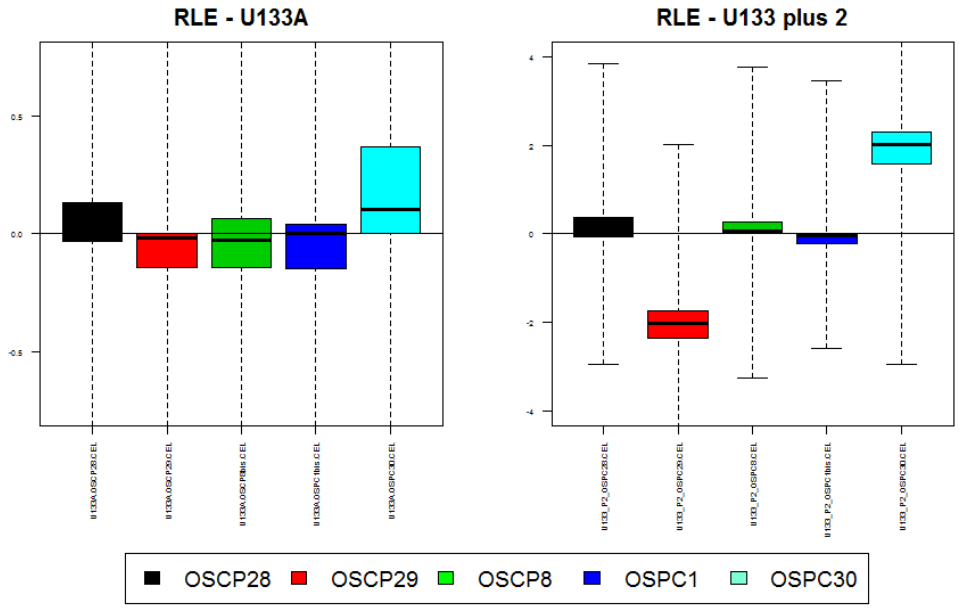



**Figura 4.7: Grafici RLE dei valori di espressione non normalizzati di entrambe le piattaforme. Notare la scala diversa in ordinata per il grafico di destra, relativo alla piattaforma U133 plus 2. In questo, alcuni esperimenti si discostano di molto dalla situazione ideale.**

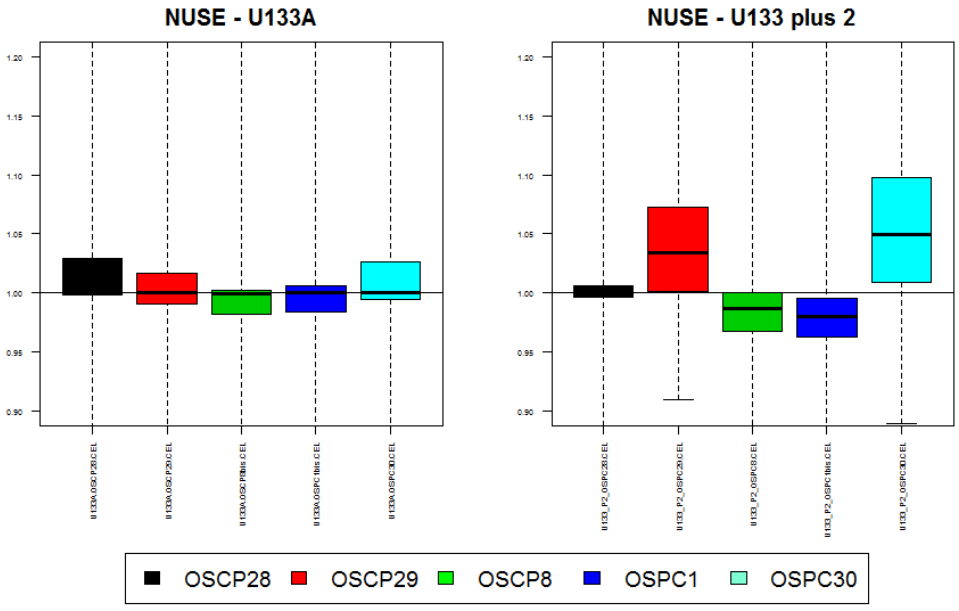



**Figura 4.8: Grafici NUSE dei valori di espressione non normalizzati di entrambe le piattaforme.**

### 4.3.2. Stime del batch effect

Per poter stimare il batch effect è necessario, innanzitutto, ottenere i valori di espressione tramite il metodo RMA, dopo la normalizzazione dei dati e l'eliminazione del background stimato, costruendo un modello lineare per ogni gene. Fatto questo, siccome i dati provengono da due differenti piattaforme e queste sono state progettate per testare un numero diverso di geni, è necessario trovare i geni comuni alle due piattaforme e costruire una matrice di valori di espressione per solo questo sottogruppo di geni. Come detto nel §4.2, la piattaforma U133A è in grado di misurare l'espressione di un numero totale di 12.098 geni; la piattaforma U133 plus 2, invece, misura 18.960 geni. I geni comuni alle due piattaforme sono 12.092, quindi solo su questi verrà condotta l'analisi per la correzione del batch effect.

Come spiegato all'inizio del capitolo 3, i metodi di rimozione del batch effect e degli altri fattori non biologici sono "application specific", ossia sono legati all'obiettivo dell'analisi che si vuole condurre. In questo elaborato, come più volte detto, l'obiettivo è quello di effettuare uno studio di differenziale espressione tra le donne con tumore all'ovaio con alta sopravvivenza globale e quelle con sopravvivenza bassa. Tutti e tre i metodi di rimozione del batch effect (COMBAT, RUV-2 e RUV-4) non potranno prescindere da ciò e, dunque, dovrà essere definita una matrice del disegno, denominata sia da COMBAT (cfr. §3.1) che da RUV-2 e RUV-4 (cfr. §3.2) come $X$. Ciò che si vuole è verificare quale sia il miglior metodo di rimozione del batch effect per questi dati, in modo poi da utilizzarlo come metodo di correzione nell'intero set dei 39 esperimenti di cui di dispone.

In Figura 4.9 è riportato il dendrogramma relativo ad una cluster analysis svolta nei 5 esperimenti ripetuti in doppia piattaforma senza alcun aggiustamento per il batch effect. L'analisi è stata svolta con un metodo di aggregazione agglomerativo gerarchico basato sul legame completo, ossia calcolando le nuove distanze tra gruppi come massimo tra le distanze dei singoli elementi che li costituiscono. La matrice delle distanze, invece, è basata su distanze euclidee.

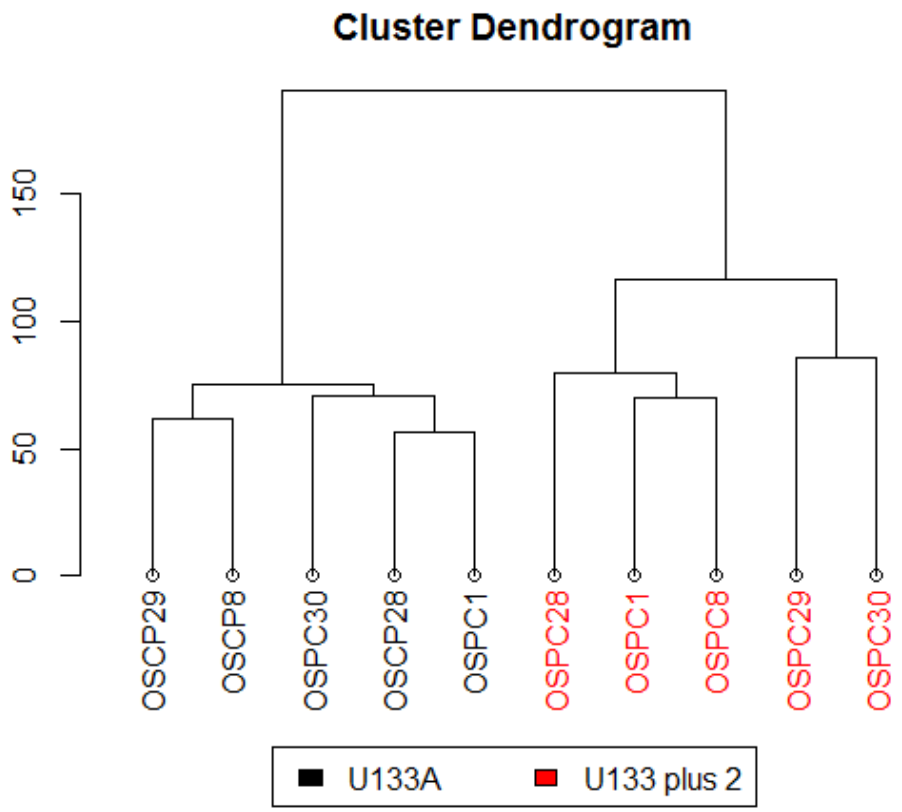



**Figura 4.9: Cluster Analysis eseguita sulla matrice di distanze ricavata dai valori di espressione normalizzati con il metodo RMA, senza alcuna correzione del batch effect, nei 5 campioni ripetuti.**

In nero sono evidenziati gli esperimenti svolti con la piattaforma U133A e in rosso quelli della piattaforma U133 plus 2. Si noti che i valori di espressione sono stati ricavati con il metodo RMA e, dunque, sono già stati normalizzati con la normalizzazione Quantile. Come anticipato, il metodo di aggregazione identifica come maggior fonte di variabilità la piattaforma utilizzata e, dunque, riconosce come più vicini gli esperimenti della stessa piattaforma. La normalizzazione non è stata in grado di rimuovere questo effetto ed è appunto per questo motivo che si sono sviluppati specifici metodi di rimozione del batch effect.

L'efficacia dei metodi di rimozione dell'effetto batch che si andranno ora ad utilizzare sui dati sarà verificata inizialmente controllando l'aspetto del dendrogramma sui dati aggiustati, salvo poi analizzare anche qualche altro indicatore di bontà.

### *1) COMBAT*

Il primo metodo di aggiustamento per il batch effect che si vuole analizzare è COMBAT, così come descritto al §3.1. Il metodo richiede la definizione della matrice $X$ del disegno, la quale contiene il fattore di interesse, ossia la sopravvivenza generale; inoltre si deve indicare esplicitamente a quale batch appartiene ogni esperimento. Ciò che viene restituito è una matrice di valori di espressione corretti utilizzando l'approccio Bayesiano empirico. Come si ricorderà, questo metodo prevede la specificazione di distribuzioni a priori per la media e la varianza delle variabili standardizzate. Il rispetto di tali assunzioni va, giustamente, controllato, rammentando però che il metodo è particolarmente robusto alla violazione degli assunti. In Figura 4.10 sono riportati i grafici di verifica del rispetto delle distribuzioni a priori. I due grafici in alto si riferiscono all'assunzione di normalità per la media dei valori standardizzati, mentre i due grafici in basso verificano l'assunto di distribuzione Gamma Inversa per la varianza. Nei grafici di sinistra le linee nere indicano il density plot stimato dai dati, mentre quelle rosse le distribuzioni teoriche Normale e Gamma Inversa. I grafici di destra, invece, sono i Q-Q plot che mostrano il discostamento tra quantili empirici e teorici delle distribuzioni. E' evidente che le code sono piuttosto pesanti per poter affermare che le ipotesi siano davvero rispettate. D'altra parte il metodo è robusto ad un'errata specificazione delle distribuzioni a priori, quindi si è deciso di provare a correggere i valori anche senza effettuare assunzioni parametriche (COMBAT non parametrico). Per verificare i risultati ottenuti si confronti l'appendice (§A.2), nella quale è introdotto il principio alla base del metodo COMBAT non parametrico e sono presenti (§A.2.1) alcuni risultati utili per verificare che effettivamente, per questi dati, l'applicazione del metodo non parametrico non comporta una grande differenza rispetto a quello parametrico. D'altro canto, però, il carico computazionale richiesto per l'utilizzo del metodo non parametrico è particolarmente alto, anche con soli 10 esperimenti. Per questi motivi si è scelto di applicare il metodo non parametrico, consci anche del fatto che è nota la situazione che si dovrebbe presentare nel caso in cui il metodo di aggiustamento avesse l'effetto auspicato.

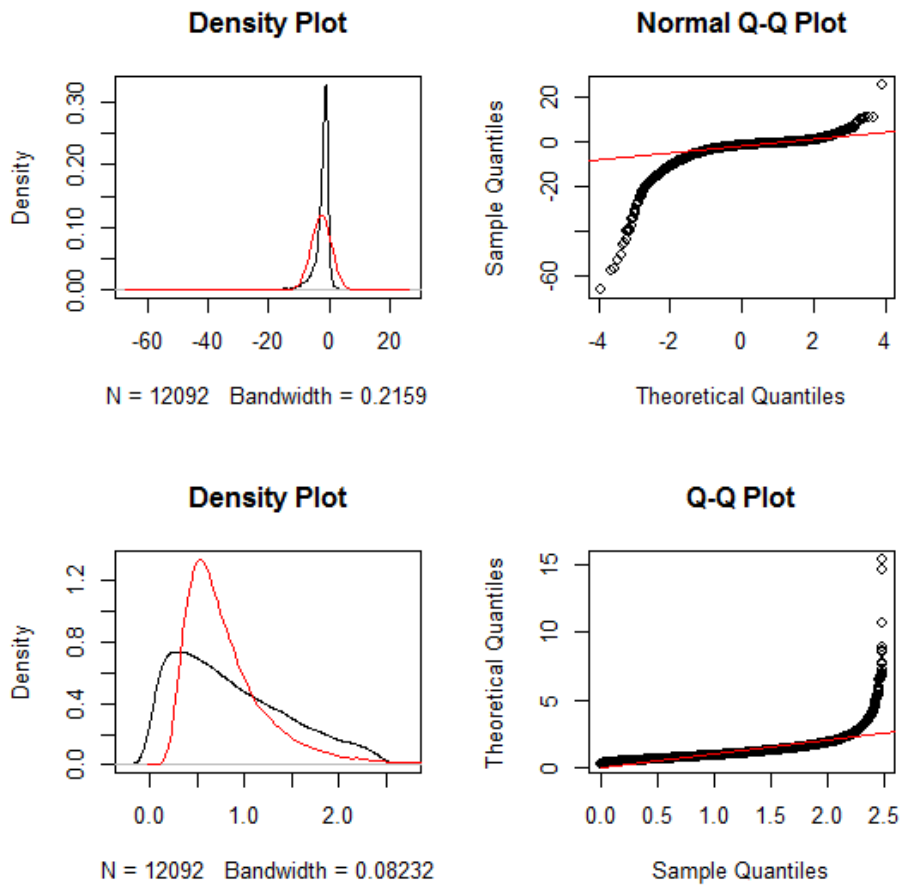



**Figura 4.10: Grafici di verifica del rispetto delle distribuzioni a priori. Nei grafici di sinistra sono mostrati in nero i density plot stimati dai dati e in rosso le distribuzioni teoriche. Nei grafici di destra sono mostrati i Q-Q plot che mostrano quantili teorici e campionari.**

In Figura 4.11 è riportato il dendrogramma relativo ad una cluster analysis effettuata sulla matrice corretta. Si nota che il metodo di correzione ha funzionato abbastanza bene, eliminando la forte fonte di variazione dovuta alla piattaforma e presente nei dati non corretti. E' però presente un errore nella creazione delle coppie, ossia quello di considerare più vicini gli esperimenti OSPC29 e OSPC28 delle due diverse piattaforme. Inoltre, si nota che i due campioni OSPC30 si uniscono ad una distanza molto più alta rispetto a tutti gli altri.

Un'analisi più approfondita sarà fatta nel prossimo paragrafo, dopo aver mostrato i risultati ottenuti con tutti e tre i metodi di aggiustamento che si vogliono studiare.

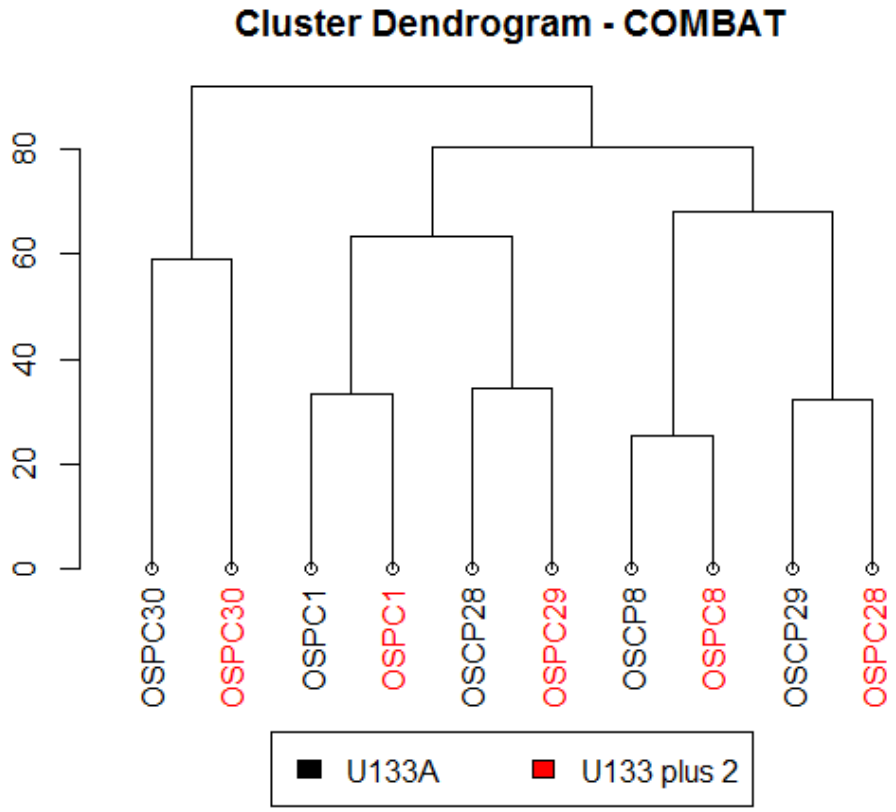



**Figura 4.11: Dendrogramma di una cluster analysis svolta sulla matrice dei valori di espressione corretti con il metodo COMBAT parametrico.**

### *2) RUV-2*

Il metodo RUV-2, così come spiegato nel §3.2.1, permette di correggere dati di espressione affetti da batch effect anche se questo effetto non è stato osservato ed è, dunque, ignoto.

RUV, indipendentemente dalla versione che se ne utilizza, è basato sull'utilizzo dei geni di controllo. In particolare i geni di controllo negativi vengono utilizzati per la stima dell'effetto batch e quelli positivi per la valutazione del modello e, talvolta, per la scelta del numero *k* più adeguato di fattori da stimare. In questo caso, dato che la valutazione del modello può essere fatta con la cluster analysis, i controlli più importanti sono quelli negativi, essenziali per una buona stima delle variazioni non biologiche da rimuovere.

I geni che qui verranno considerati come controlli negativi sono i controlli inseriti da Affymetrix nei suoi array. Questi sono facilmente individuabili nella matrice dei dati grazie al loro prefisso "AFFX". I controlli comuni ad entrambe le piattaforme sono 62, un numero non particolarmente elevato.

In Figura 4.12 è riportato il dendrogramma della cluster analysis eseguita sui valori di espressione corretti con RUV-2. Come per i valori aggiustati con COMBAT, la cluster analysis sbaglia l'aggregazione degli esperimenti OSPC28 e OSPC29. In questo caso, però, anche l'aggregazione degli esperimenti OSPC30 è sbagliata e per questi, a differenza di COMBAT, non viene creato un gruppo a sé stante.

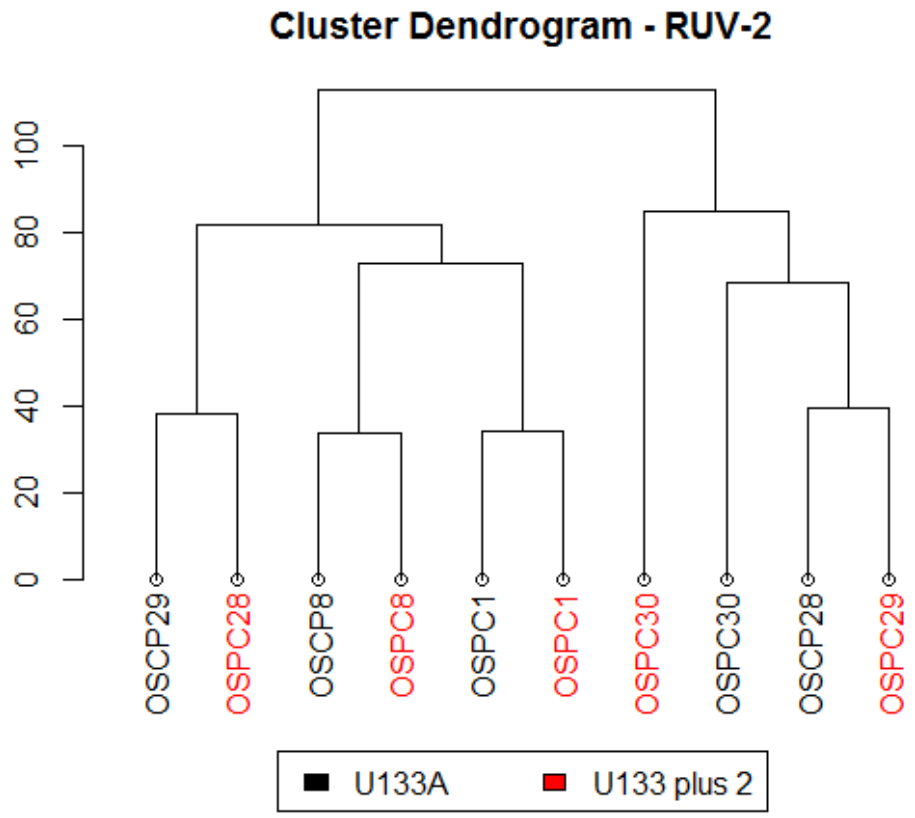



**Figura 4.12: Dendrogramma di una cluster analysis svolta sulla matrice dei valori di espressione corretti con il metodo RUV-2.**

Riguardo la scelta del numero $k$ di fattori identificati e rimossi dall'analisi fattoriale, si è scelto di utilizzare $k = 1$. La scelta è stata dettata dal fatto che si è a conoscenza dell'esistenza di un solo fattore, ossia quello della piattaforma; inoltre si è provato a scegliere valori di $k$ più alti ma la cluster analysis evidenziava un evidente peggioramento rispetto a $k = 1$.

### *3) RUV-4*

La differenza sostanziale tra il metodo RUV-2 e il RUV-4 è sulla procedura per la stima dell'ignota matrice $W$ delle covariate non osservate (cfr. §3.2.2). Il RUV-4 prevede anzitutto la stima e la rimozione della matrice $X$ del fattore di interesse, poi la stima, tramite analisi fattoriale, della matrice $W_0$ (ossia $W$ da cui è stato tolto l'effetto di $X$) e solo poi la stima di $W$ recuperando la parte rimossa con la rimozione di $X$. Come per il RUV-2, i geni di controllo negativi sono i 62 controlli con prefisso "AFFX". Il risultato della cluster analysis svolta sui valori corretti con il RUV-4 è in Figura 4.13.

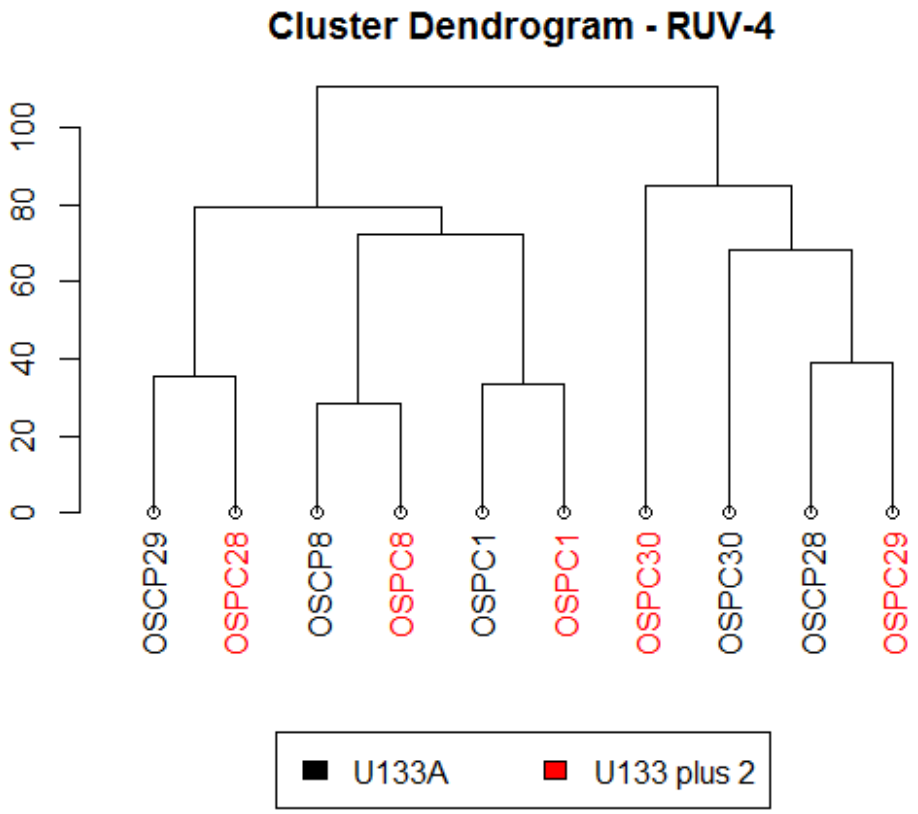



**Figura 4.13: Dendrogramma di una cluster analysis svolta sulla matrice dei valori di espressione corretti con il metodo RUV-4.**

Anche in questo caso il valore scelto per $k$ è stato $k = 1$. Come per il RUV-2, la scelta è stata fatta provando anche valori maggiori, i quali conducevano però a risultati peggiori.

### 4.3.3. Risultati

Dalle analisi del paragrafo precedente si nota che la cluster analysis svolta sui dati aggiustati non mostra evidenti effetti residui della piattaforma, ciò a conferma del fatto che i metodi utilizzati funzionano nella pratica. Malgrado ciò, la situazione che si presenta non è la migliore che ci si potesse aspettare, dato che nessuno dei tre metodi utilizzati ha saputo aggiustare i valori in modo che le coppie di esperimenti uguali fatti con piattaforma diversa formassero gruppi a sé stanti. Il problema, però, è capire se la colpa di ciò sia imputabile ai metodi di aggiustamento oppure se la qualità degli esperimenti rendesse impossibile il verificarsi della situazione ottimale. In particolare, come evidenziato dalle analisi esplorative del §4.3.1, il campione OSPC30 della piattaforma U133 plus 2 potrebbe avere problemi di qualità, a causa della forte degradazione dell'RNA che si è verificata. Sempre nella seconda piattaforma, poi, il campione OSPC29 è affetto da un'ibridazione non particolarmente ben riuscita e che, probabilmente, la normalizzazione non è stata in grado di correggere. Si consideri che, per maggior completezza, si sono ripetute tutte le precedenti analisi anche con una normalizzazione diversa, ossia la VSN introdotta nel §2.3.2. Queste hanno condotto a risultati del tutto analoghi a quelli ottenuti con la normalizzazione Quantile adottata dal metodo RMA che si è utilizzato per ricavare i valori di espressione (cfr. §A.3).

E' possibile che le cause degli errori siano imputabili ai problemi appena evidenziati. Tra i metodi di correzione utilizzati è evidente che COMBAT parte da una posizione di vantaggio datagli dal conoscere il vero batch effect, ossia la piattaforma. Questo si trasformerebbe in uno svantaggio se ci fossero anche altri effetti batch non noti, che non potrebbero essere stimati e rimossi. In questo caso sembrano non essere presenti altri effetti batch, dato che la correzione di COMBAT permette di ottenere gruppi migliori rispetto ai due concorrenti e, tutto sommato, i problemi rilevati nelle analisi esplorative spiegano le carenze dei risultati. Riguardo le più scarse performance del metodo RUV, c'è da ricordare che il metodo è sensibile alla scelta dei geni di controllo utilizzati. In questo caso i geni di controllo negativi rispettano le condizioni richieste dal procedimento, ossia (i) sono colpiti dal fattore che si vuole rimuovere e (ii) non sono associati

con il fattore di interesse; malgrado ciò il loro numero è abbastanza ridotto, dato che sono solo 62 controlli.

Per decidere quale metodo di aggiustamento sia migliore si può provare ad utilizzare anche qualcuno dei metodi introdotti nel §3.3. Ad esempio, si può controllare il comportamento dei grafici RLE, per vedere se gli aggiustamenti comportano un suo miglioramento. In Figura 4.14 sono riportati i grafici RLE dei dati normalizzati con la Quantile, senza nessun aggiustamento per il batch effect; mentre i grafici RLE per i valori di espressione aggiustati sono riportati in Figura 4.15.

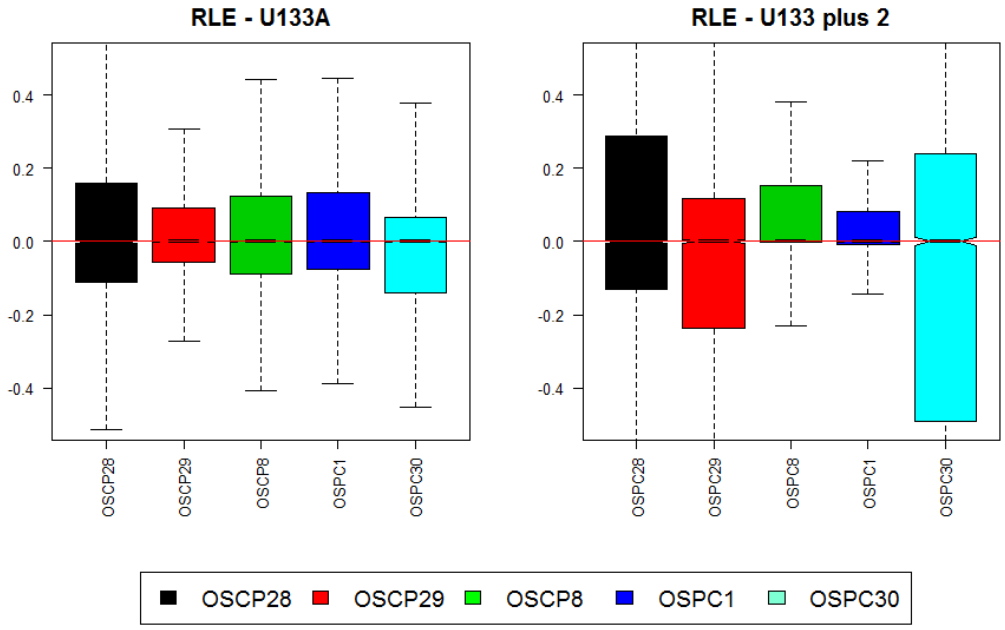



**Figura 4.14: Grafici RLE dei dati normalizzati con la Quantile, senza nessun aggiustamento per il batch effect.**

Si nota che la correzione con il metodo COMBAT lascia i grafici RLE pressoché invariati, mantenendoli centrati nello zero e riducendo in maniera quasi impercettibile, per l'esperimento OSPC30, l'ampiezza della scatola del boxplot. Gli altri due metodi di aggiustamento, invece, peggiorano i grafici RLE spostando, per alcuni esperimenti, la mediana delle distribuzioni.

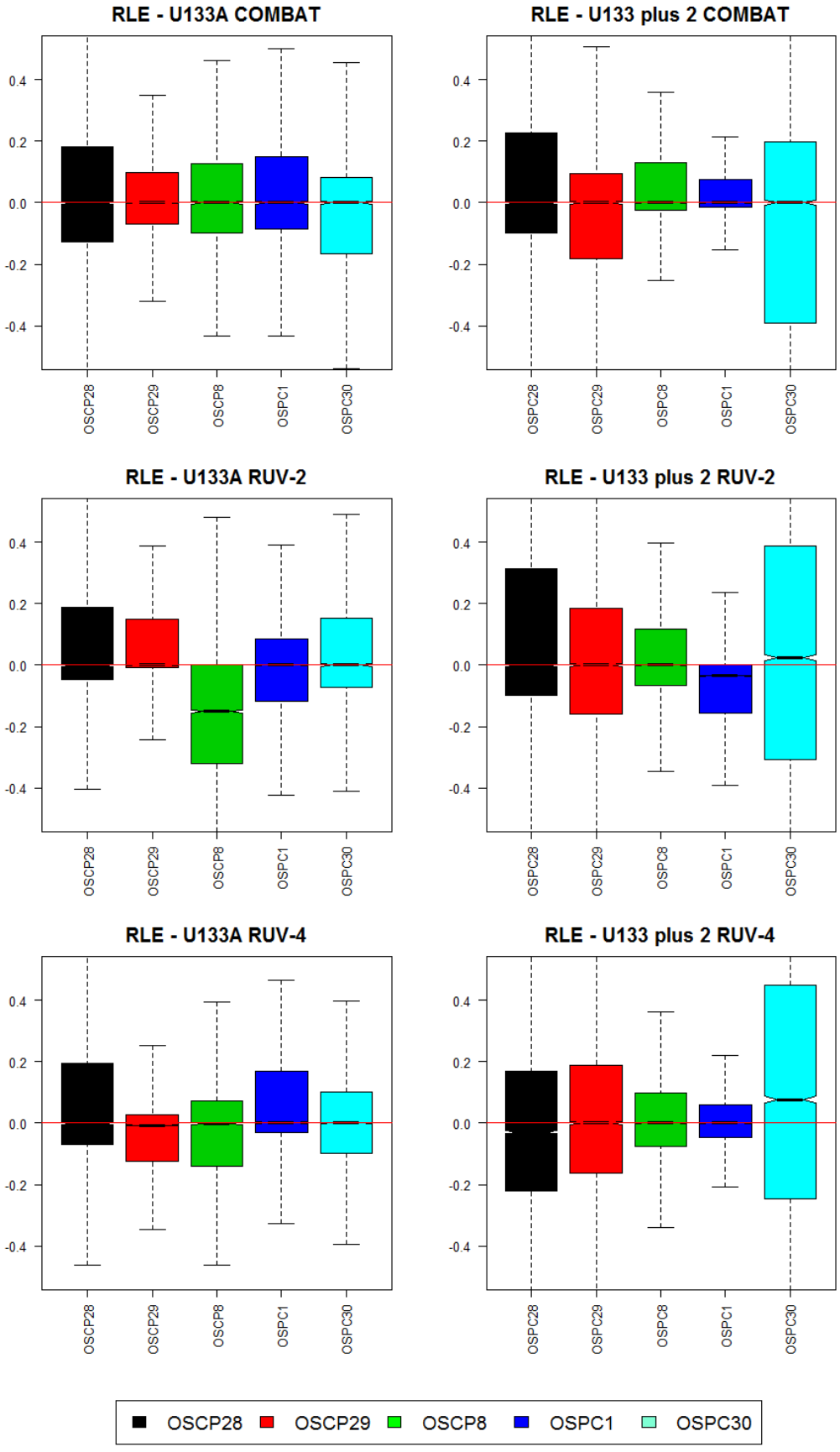



**Figura 4.15: Grafici RLE dei dati corretti con il metodo COMBAT (sopra), il metodo RUV-2 (al centro) e con il metodo RUV-4 (sotto).**

Un altro metodo di valutazione prevede di controllare che, mediamente, la varianza dei geni di controllo negativi non aumenti; idealmente dovrebbe diminuire.

In Figura 4.16 sono riportati i boxplot delle varianze dei geni di controllo negativi. Si nota che queste diminuiscono dopo l'applicazione di tutti e tre i metodi di correzione; il valore mediano più basso è comunque quello relativo a COMBAT.

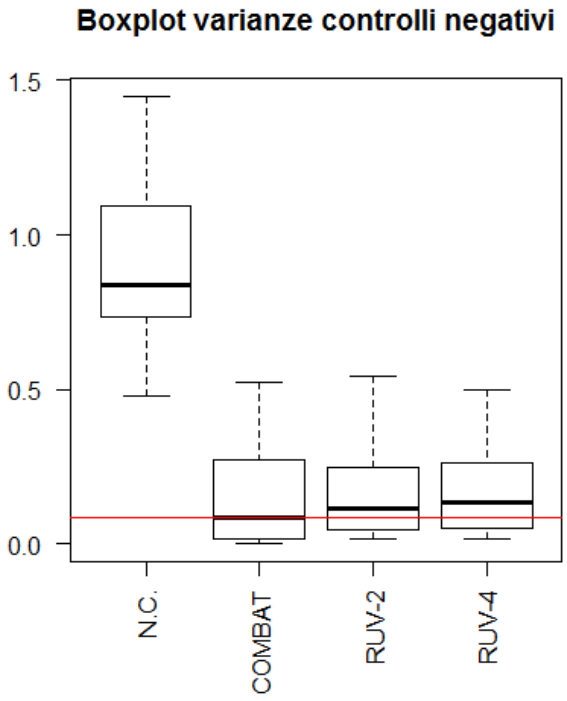



**Figura 4.16: Boxplot delle varianze dei valori di espressione dei geni di controllo negativi non aggiustati (primo), aggiustati con COMBAT (secondo), con RUV-2 (terzo) e con RUV-4 (quarto). La linea rossa è in corrispondenza della mediana minima tra le quattro.**

Come spiegato nel §3.3.2, un altro possibile controllo per la verifica della bontà degli aggiustamenti è l'analisi della distribuzione dei p-value ottenuti effettuando un test su ogni gene. Solitamente il test viene condotto sul fattore biologico di interesse, che in questo caso è la diversa sopravvivenza delle pazienti. Qui si dispone, però, di esperimenti ripetuti in doppia piattaforma e, dunque, si può pensare di eseguire un test che sfrutti la natura delle osservazioni. Ciò potrebbe essere fatto conducendo dei test per dati appaiati, per verificare l'ipotesi

che tra le coppie di osservazioni relative alla stessa paziente nessuno dei geni sia differenzialmente espresso. Il disegno sperimentale è quello descritto nella Tabella 4.3. I p-value devono essere calcolati sul contrasto relativo alla diversa piattaforma. Ripetendo i test sui dati non corretti e su quelli corretti con COMBAT, RUV-2 e RUV-4, ci si aspetta che nei primi sia rilevato un numero di geni differenzialmente espressi alto, mentre nei dati corretti i p-value tendano ad essere distribuiti su valori alti, dato che è nota l'assenza di differenziale espressione.

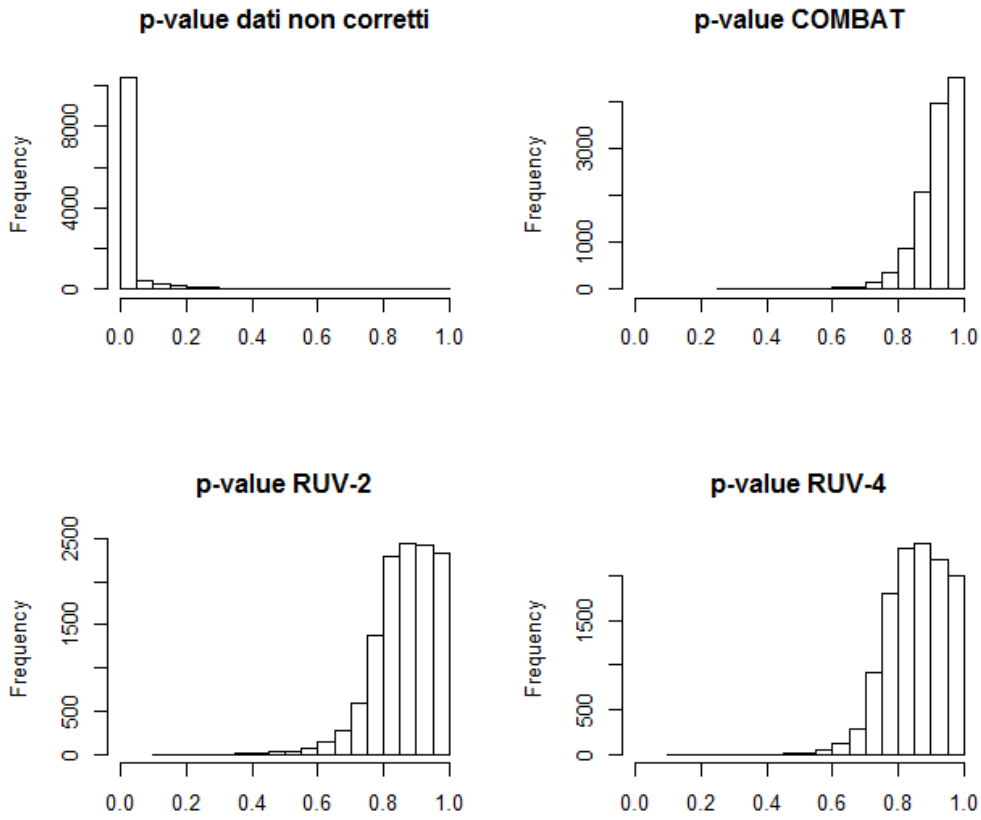



**Figura 4.17: Istogramma dei p-value ottenuti effettuando un test Ebayes per dati appaiati per ogni gene. In tutti i casi si è stimato un modello lineare per la verifica dell'ipotesi di differenziale espressione tra i dati nelle due diverse piattaforme, tenendo conto del fatto che le osservazioni sono ripetute, come descritto nella Tabella 4.3.**

Osservando i grafici della Figura 4.17 si nota che i p-value si comportano sostanzialmente come atteso. E' molto confortante notare che a seguito di tutte le correzioni non vengono identificati geni differenzialmente espressi, i quali

sarebbero da considerarsi falsi positivi. Tra tutti i metodi, quello che sembra essere migliore è ancora una volta COMBAT, il quale comporta uno schiacciamento della distribuzione dei p-value verso i valori più alti. Il valore minimo dei p-value per i dati corretti con COMBAT è 0.264; per i dati corretti con RUV-2 è 0.121 e per quelli corretti con RUV-4 è 0.1.

In definitiva, considerando tutte le analisi svolte, la correzione più adatta per questi dati sembra essere quella di COMBAT, per la quale risultano migliori tutti gli indicatori di bontà.

Per quanto detto, nel prossimo paragrafo, sarà utilizzato COMBAT per correggere i valori di espressione di tutti i 39 esperimenti di cui si dispone; in modo poi da eseguire le analisi per l'identificazione dei geni differenzialmente espressi.

| Campione | Appaiamento | Piattaforma |
|---|---|---|
| *U133A-OSPC28* | 1 | 0 |
| *U133A-OSPC29* | 2 | 0 |
| *U133A-OSPC8* | 3 | 0 |
| *U133A-OSPC1* | 4 | 0 |
| *U133A-OSPC30* | 5 | 0 |
| *U133P2-OSPC28* | 1 | 1 |
| *U133P2-OSPC29* | 2 | 1 |
| *U133P2-OSPC8* | 3 | 1 |
| *U133P2-OSPC1* | 4 | 1 |
| *U133P2-OSPC30* | 5 | 1 |

**Tabella 4.3: Descrizione del disegno sperimentale per il test Ebayes su dati appaiati.**

## 4.4. Analisi dei dati di tumore all'ovaio

Come introdotto all'inizio del §4.2 si è in possesso di un totale di 44 array ottenuti con le due piattaforme Affymetrix U133A e U133 plus 2. In realtà però, mentre per 34 pazienti si dispone di un solo array, per 5 pazienti si dispone di

osservazioni doppie, effettuate cioè con entrambe le piattaforme. In definitiva, dunque, si hanno 39 singoli esperimenti relativi a 39 pazienti affette da tumore all'ovaio. La prima scelta che si deve effettuare è la decisione su quali esperimenti utilizzare per le pazienti rilevate in doppia piattaforma. Questa decisione può essere presa abbastanza tranquillamente grazie alle analisi esplorative svolte su questi dati nel §4.3.1. Come emerso in quel paragrafo, infatti, alcuni esperimenti effettuati con la piattaforma U133 plus 2 erano affetti da problemi di ibridazione e di qualità degli esperimenti. Per questo motivo sembra preferibile utilizzare, per ognuna di queste cinque pazienti, l'array ottenuto con la piattaforma U133A. Presa questa decisione, si dispone di 25 osservazioni per la piattaforma U133A e di 14 per la piattaforma U133 plus 2.

### 4.4.1. Analisi esplorative

Come fatto nel §4.3.1, è necessario svolgere delle analisi esplorative sull'insieme di tutti i dati disponibili. Queste verranno svolte in parallelo per entrambe le piattaforme, proprio come fatto in quel paragrafo.

Nel §A.4 dell'appendice, nelle Figure A.5 e A.6, sono riportate le immagini in gradazione di colore utili per capire se l'ibridazione sia avvenuta in maniera corretta o meno, rispettivamente, nella piattaforma U133A e U133 plus 2. Si nota che in tutti gli esperimenti della piattaforma U133 plus 2 l'ibridazione è avvenuta correttamente: non sono presenti aree dei vetrini caratterizzate da colori intensi rispetto al resto del chip. Alcuni esperimenti della piattaforma U133A, invece, hanno aree del vetrino caratterizzate da un'ibridazione molto alta o molto bassa. Si osservino, ad esempio, gli angoli inferiori del vetrino dell'esperimento met-7310002; oppure anche i lati del vetrino degli esperimenti met-7310003 e met-7310015. Purtroppo eventi di questo tipo sono abbastanza frequenti in questo genere di esperimenti. Qui, in quasi tutti i casi, si tratta di aree abbastanza piccole e la normalizzazione dovrebbe essere in grado di rimuovere anche questo tipo di problemi.

In Figura 4.18 è riportato il grafico di degradazione dell'RNA per la piattaforma U133A. Le spezzate sono molto simili tra loro per cui la degradazione

dell'RNA è avvenuta allo stesso modo per tutti i campioni; inoltre, come ci si aspetta, i probe al 5' sono più degradati di quelli al 3'.

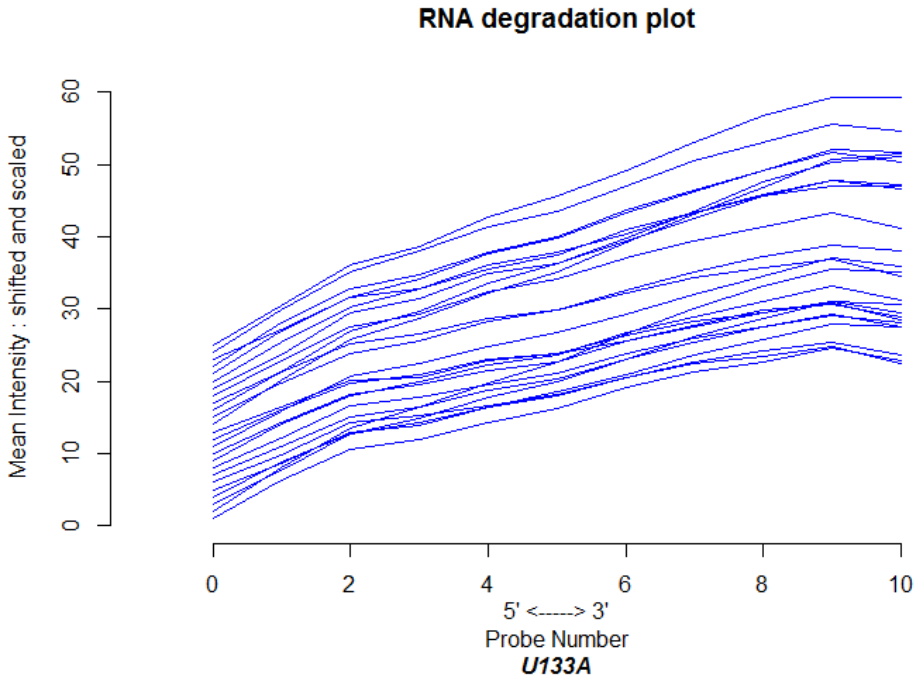



**Figura 4.18: Grafico di degradazione dell'RNA per la piattaforma U133A.**

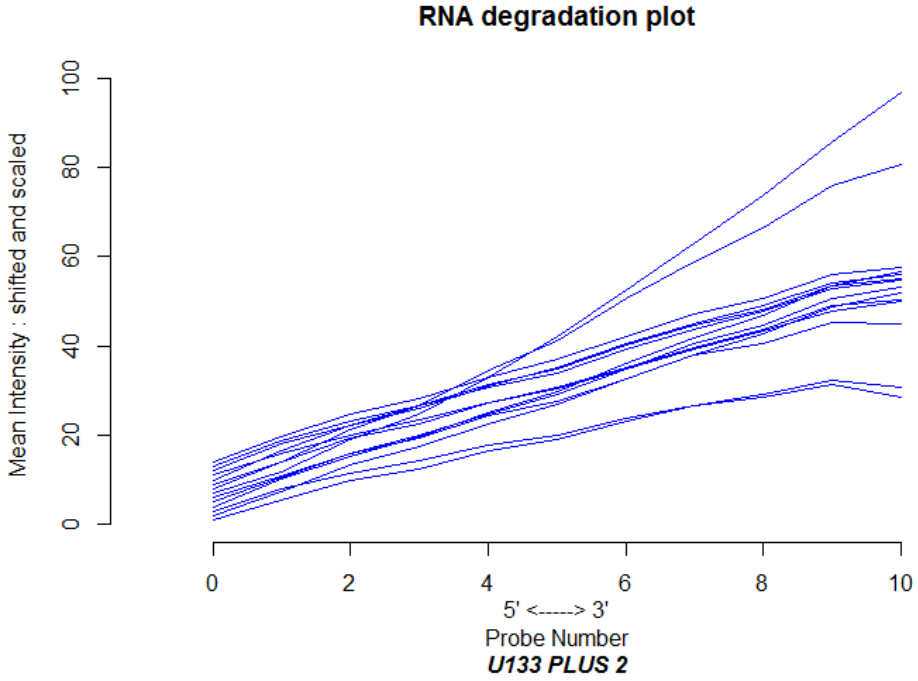



**Figura 4.19: Grafico di degradazione dell'RNA per la piattaforma U133 plus 2.**

Le stesse considerazioni possono essere fatte per la piattaforma U133 plus 2, nella quale si assiste, però, ad una differenza di comportamento per un paio di campioni. Ciò non è un problema dato che, per questi, l'RNA risulta meno degradato degli altri. Il grafico per questa piattaforma è riportato in Figura 4.19.

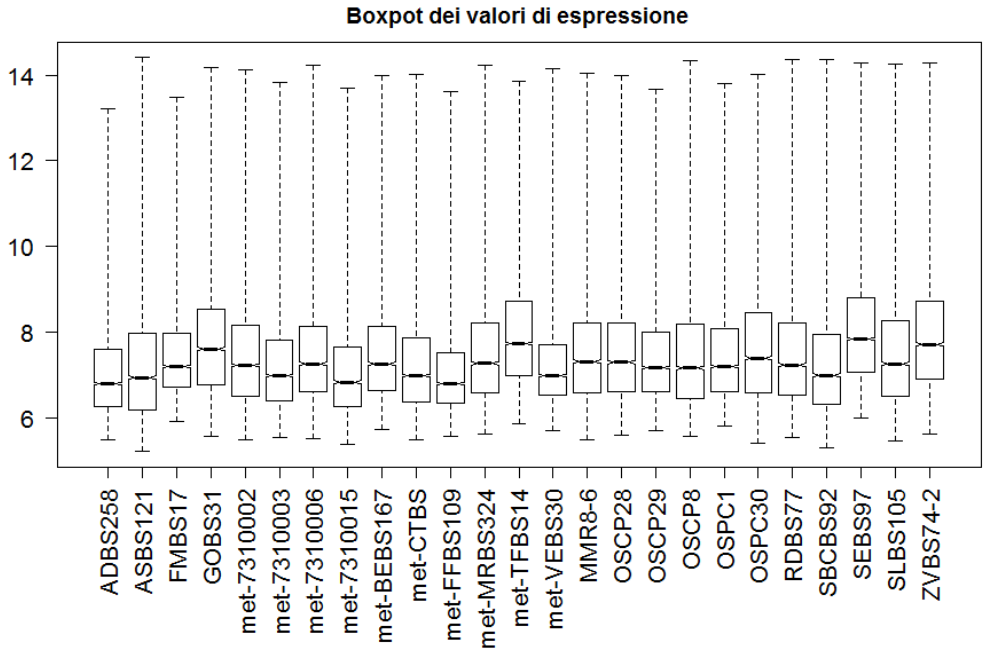



**Figura 4.20: Boxplot dei valori di espressione non corretti né normalizzati in ogni esperimento (piattaforma U133A).**

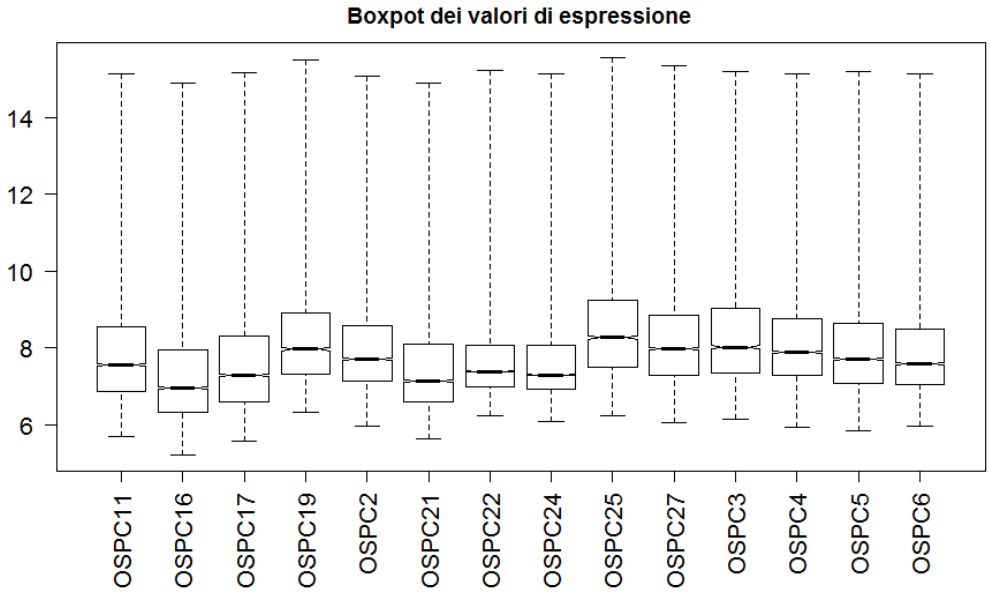



**Figura 4.21: Boxplot dei valori di espressione non corretti né normalizzati in ogni esperimento (piattaforma U133 plus 2).**

Nelle Figure 4.20 e 4.21 sono riportati i boxplot delle distribuzioni dei valori non corretti né normalizzati per la piattaforma U133A e U133 plus 2, rispettivamente. E' evidente che i dati risentono della non normalizzazione, ma le distribuzioni sono comunque abbastanza omogenee e questo è indice di buona qualità degli esperimenti.

### 4.4.2. Rimozione delle variazioni non biologiche

Nel §4.3.3 si sono spiegati i motivi per cui, per i dati a disposizione, il metodo migliore di eliminazione del batch effect causato dalla piattaforma sembra essere COMBAT. In questo paragrafo, dunque, si ricaveranno i valori di espressione normalizzati e senza background, per poi utilizzare tale metodo di correzione e verificarne l'adeguatezza analizzando alcuni degli indicatori descritti nel §3.3 ed utilizzati anche nel §4.3.3. Riguardo il metodo della cluster analysis, questo è stato utilizzato nei precedenti paragrafi come tecnica per la verifica dell'efficacia dei metodi di correzione. Ciò è stato reso possibile dalla conoscenza che si aveva dell'esatto appaiamento dei gruppi nel caso in cui il metodo di aggiustamento avesse funzionato perfettamente. E' necessario puntualizzare che, ora, la stessa tecnica non può funzionare come in precedenza, dato che non si ha più la conoscenza sui gruppi che dovrebbero formarsi. L'unica utilità che adesso si può trarre dalla cluster analysis è verificare se l'effetto piattaforma sembra sparire oppure no dai dati corretti. In merito a questo argomento si osservino le Figure 4.22 e 4.23, nelle quali sono riportati, rispettivamente, il dendrogramma ricavato da una cluster analysis eseguita sui valori di espressione non corretti e il dendrogramma per i valori di espressione corretti con il metodo COMBAT. E' evidente l'effetto piattaforma presente nei dati non corretti; questo sembra essere stato rimosso dopo la correzione anche se, come detto, la cluster analysis non è più il metodo migliore per accertarsene.

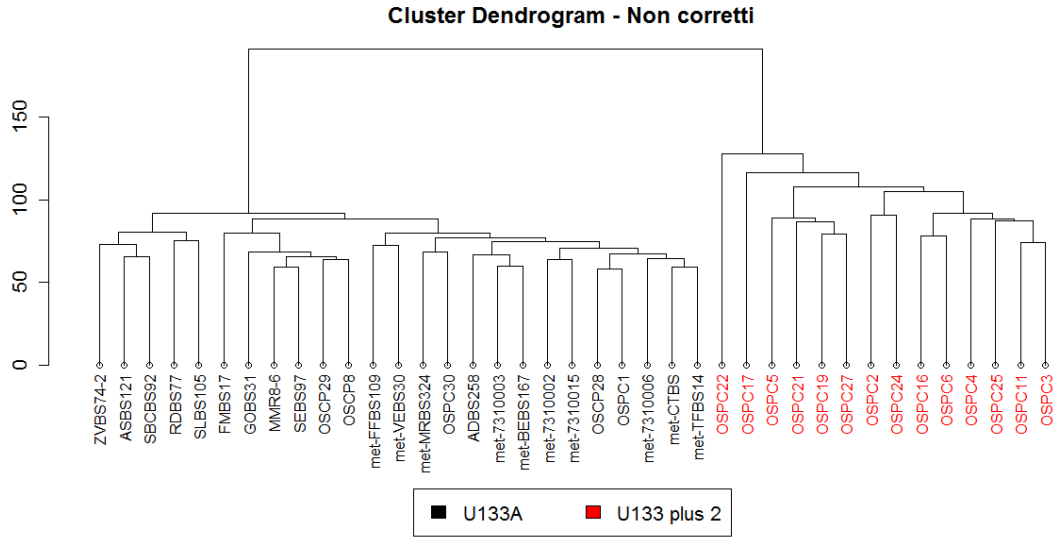



**Figura 4.22: Dendrogramma di una cluster analysis sui valori di espressione non corretti.**

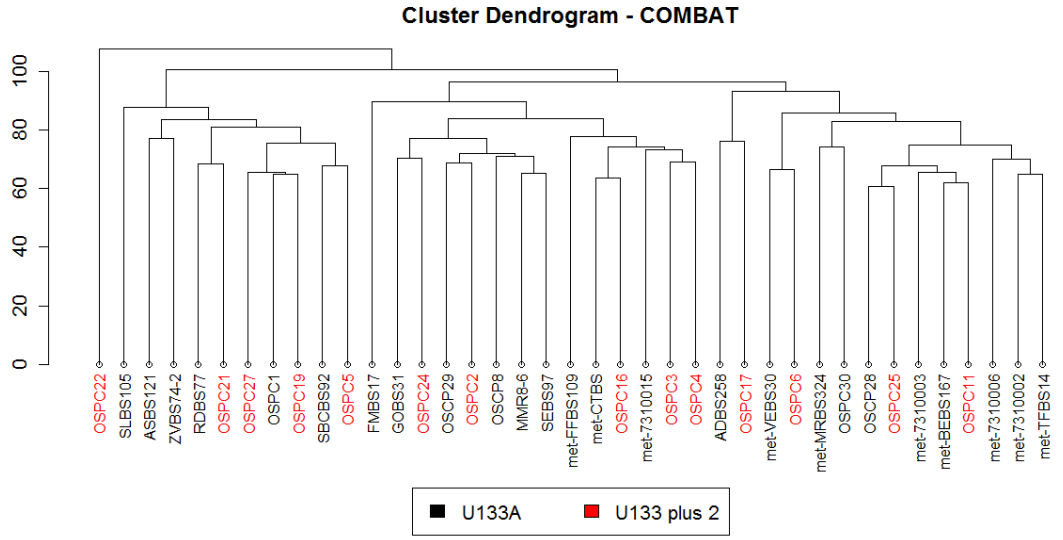



**Figura 4.23: Dendrogramma di una cluster analysis sui valori di espressione corretti con COMBAT.**

Per controllare l'efficacia del metodo di aggiustamento si osservino anche i grafici RLE dei valori di espressione grezzi (senza normalizzazione), normalizzati e corretti per il batch effect. Mentre è evidente il miglioramento apportato dalla normalizzazione, non è visibile in maniera chiara un miglioramento prodotto dal metodo di correzione, ma nemmeno si assiste ad un peggioramento. Si noti, ad esempio, la piccola riduzione dell'ampiezza della scatola del boxplot relativa al campione OSPC22.

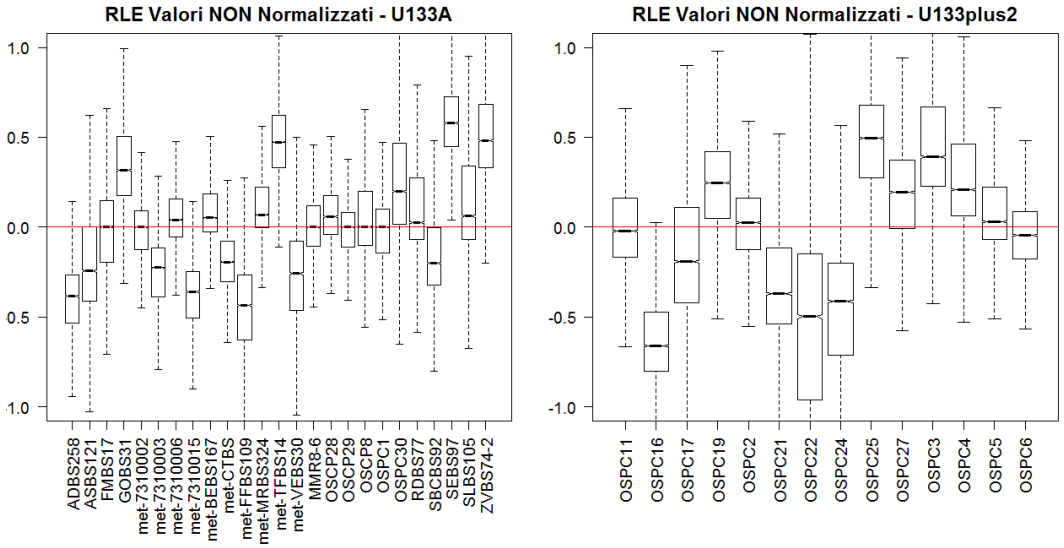



**Figura 4.24: Grafici RLE dei valori di espressione non normalizzati né corretti per il background.**

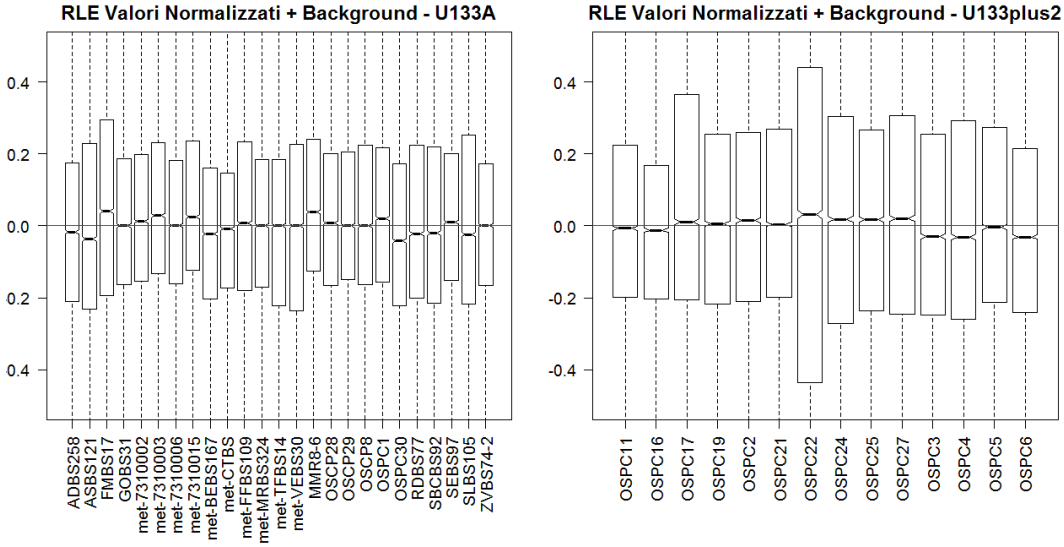



**Figura 4.25: Grafici RLE dei valori di espressione normalizzati (Quantile) e corretti per il background secondo il metodo RMA.**

In Figura 4.27 sono riportati i boxplot delle varianze campionarie dei valori di espressione dei geni di controllo negativi. Come ci si aspetta, i valori di espressione corretti con COMBAT, per questi geni, hanno variabilità minore rispetto a quelli non corretti. Ciò indica che l'aggiustamento ha funzionato correttamente.

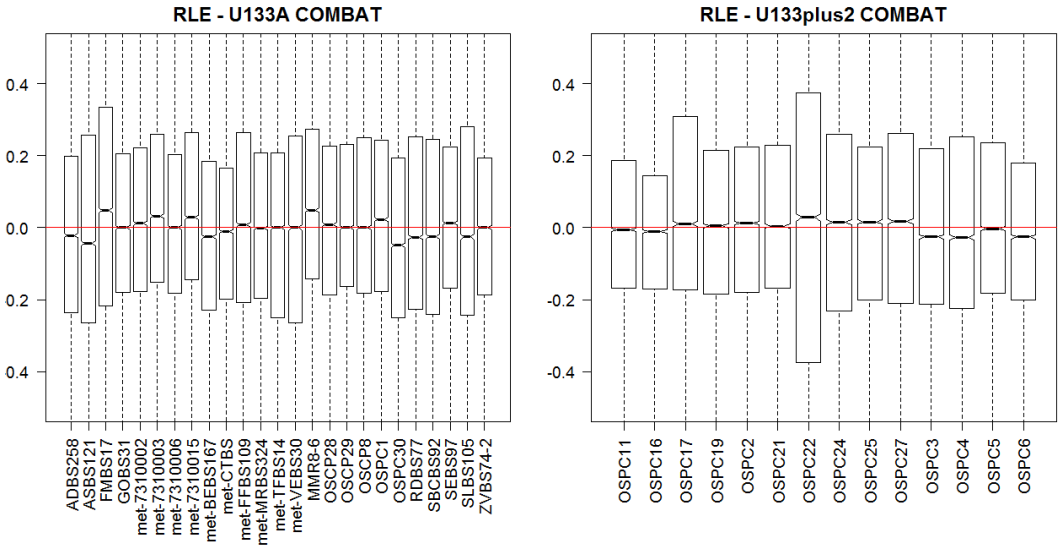



**Figura 4.26: Grafici RLE dei valori di espressione corretti con il metodo COMBAT.**

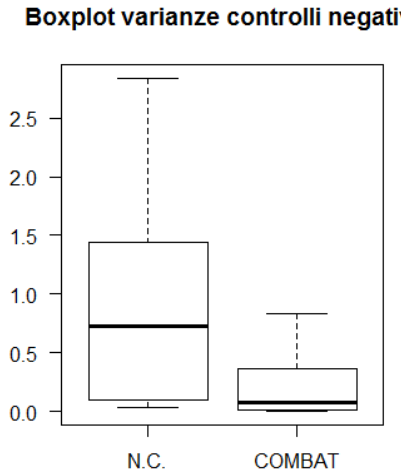



**Figura 4.27: Boxplot delle varianze campionarie dei valori di espressione dei geni di controllo negativi.**

Altro grafico che aiuta a capire l'efficacia del metodo di aggiustamento è quello della distribuzione dei p-value. Questi sono stati ricavati stimando un modello lineare per la verifica dell'ipotesi di differenziale espressione tra pazienti con sopravvivenza elevata e pazienti con sopravvivenza bassa nei dati non corretti e in quelli corretti con COMBAT. Questo modello sarà approfondito nel paragrafo successivo, relativo all'identificazione dei geni differenzialmente espressi. Si è poi utilizzato il test Empirical Bayes (cfr. §2.5.3) per calcolare il valore della

statistica test e il relativo p-value associato. Di questi valori si è poi fatto un istogramma e ne risulta che i p-value corretti con COMBAT sono più vicini alla situazione auspicabile di uniformità, sintomo del buon funzionamento dell'aggiustamento.

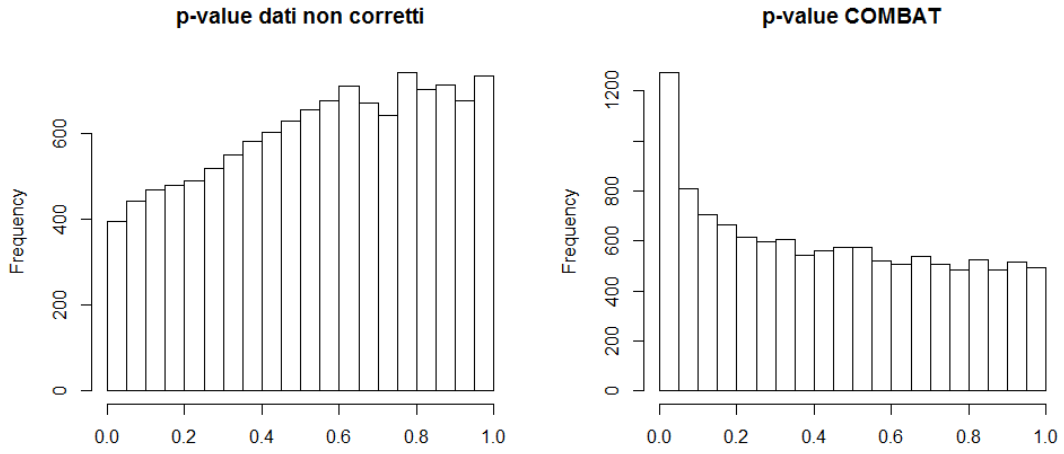



**Figura 4.28: Istogramma dei p-value ottenuti effettuando un test per ogni gene. Per i dati non corretti e per quelli corretti con COMBAT si è stimato un modello lineare per la verifica dell'ipotesi di differenziale espressione tra pazienti con sopravvivenza elevata e pazienti con sopravvivenza bassa.**

Malgrado nel §4.3.3 sia stata ampiamente argomentata la maggior efficacia del metodo COMBAT per la correzione dei dati disponibili in doppia piattaforma, cercando di analizzare criticamente anche le motivazioni di tali risultati, ci si è voluti accertare della correttezza dell'applicazione di questo metodo anche nell'insieme di tutti i campioni disponibili. A tal proposito, nel §A.5 sono riportati i segnali di bontà che mostrano l'applicazione dei metodi di aggiustamento RUV-2 e RUV-4 ai dati completi. Questi mostrano una inadeguatezza del metodo RUV-2 ma, d'altra parte, una buona correzione fatta dal metodo RUV-4. Purtroppo non è possibile sapere per certo quale delle due correzioni sia la migliore, ma certamente i risultati ottenuti per i dati in doppia piattaforma fanno propendere per la preferenza di COMBAT. Per questo motivo le analisi di differenziale espressione e di gene set saranno condotte principalmente sui dati corretti con questo metodo, lasciando però spazio a dei confronti con i risultati che si ottengono applicando la correzione di RUV-4.

### 4.4.3. Analisi descrittive sulle variabili cliniche

Come specificato nel §4.2 sulle pazienti sono state rilevate delle variabili cliniche, le quali possono essere utilizzate per svolgere alcune analisi descrittive. La descrizione delle variabili è riportata nella Tabella 4.2 del suddetto paragrafo.

Come già osservato più volte, le pazienti sono tutte affette da tumore all'ovaio sieroso di grado medio-alto. In particolare sono presenti 5 pazienti con tumore di grado medio (G2) e 34 con tumore di grado alto (G3). Per quanto riguarda la stadiazione dei tumori, sono solo 3 le pazienti con tumori di stadio basso (IB, IIB e IIC), 26 hanno tumore di stadio IIIC e 10 di stadio IV.

La sopravvivenza globale delle pazienti è stata divisa in due classi (alta e bassa) e, come detto, la scelta delle donne da includere nel campione non è avvenuta in maniera casuale, ma cercando di ottenere una netta separazione tra le classi. Risultano, dunque, 12 pazienti caratterizzate da una sopravvivenza alta, la quale è compresa tra 82 mesi e 120 mesi, e 27 pazienti con sopravvivenza bassa, compresa tra 1 mese e 33 mesi.

E' risaputo che il rischio di contrarre questo tipo di tumore aumenta con l'età. E' raro che si presenti in donne di età inferiore a 40 anni e, generalmente, colpisce donne già in menopausa. Nel campione osservato è presente solo una donna con età alla diagnosi inferiore a 40 anni, in particolare di 24 anni di età. Oltre questa, nella fascia 40 – 50 anni sono presenti 6 pazienti, nella fascia 51 – 60 ce ne sono 4, in quella 61 – 70 ce ne sono 14, in quella 71 – 80 ce ne sono 12 e, infine, sopra gli 81 anni ce ne sono solo 2, con l'età massima di 84 anni. L'età mediana è 66 anni. 8 pazienti non sono in menopausa, mentre le rimanenti 31 lo sono.

Per alcune delle variabili cliniche di cui si dispone può essere interessante verificare l'associazione con la variabile di interesse, ossia la sopravvivenza globale. Ciò può essere utile per capire se, tra le variabili a disposizione, alcune possano essere utili per la previsione della sopravvivenza delle pazienti.

Alcune variabili che potrebbero essere utili per questo scopo sono: lo stadio FIGO, la dimensione del tumore residuo dopo l'operazione, la presenza di asciti, la positività dei linfonodi, il CA125 preoperatorio, l'età alla diagnosi, lo stato di menopausa, il grado del tumore, la risposta alla chemioterapia e la recidività.

| Sopravvivenza | Stadio FIGO | | $Tot$ |
|---|---|---|---|
| | *I-II* | *III-IV* | |
| *Alta* | 3 | 9 | 12 |
| *Bassa* | 0 | 27 | 27 |
| *Tot* | 3 | 36 | 39 |

**Tabella 4.4: Frequenze delle osservazioni per le classi di sopravvivenza secondo la stadiazione secondo il metodo FIGO.**

| Sopravvivenza | Dimensione del tumore residuo | | $Tot$ |
|---|---|---|---|
| | $TR \leq 1$ | $TR > 1$ | |
| *Alta* | 7 | 5 | 12 |
| *Bassa* | 5 | 22 | 27 |
| *Tot* | 12 | 27 | 39 |

**Tabella 4.5: Frequenze delle osservazioni per le classi di sopravvivenza secondo la dimensione del tumore residuo.**

| Sopravvivenza | Asciti | | $Tot$ |
|---|---|---|---|
| | *Sì* | *No* | |
| *Alta* | 5 | 7 | 12 |
| *Bassa* | 25 | 2 | 27 |
| *Tot* | 30 | 9 | 39 |

**Tabella 4.6: Frequenze delle osservazioni per le classi di sopravvivenza secondo la presenza o meno di asciti.**

| Sopravvivenza | Linfonodi | | *Tot* |
|---|---|---|---|
| | *Positivi* | *Negativi* | |
| *Alta* | 2 | 8 | 10 |
| *Bassa* | 7 | 5 | 12 |
| *Tot* | 9 | 13 | 22 |

**Tabella 4.7: Frequenze delle osservazioni per le classi di sopravvivenza secondo la presenza o meno di asciti.**

| Sopravvivenza | Menopausa | | *Tot* |
|---|---|---|---|
| | *Sì* | *No* | |
| *Alta* | 8 | 4 | 12 |
| *Bassa* | 23 | 4 | 27 |
| *Tot* | 31 | 8 | 39 |

**Tabella 4.8: Frequenze delle osservazioni per le classi di sopravvivenza secondo lo stato o meno di menopausa.**

| Sopravvivenza | Grado | | *Tot* |
|---|---|---|---|
| | *G2* | *G3* | |
| *Alta* | 0 | 12 | 12 |
| *Bassa* | 5 | 22 | 27 |
| *Tot* | 5 | 34 | 39 |

**Tabella 4.9: Frequenze delle osservazioni per le classi di sopravvivenza secondo il grado del tumore.**

| Sopravvivenza | Risposta alla prima linea di chemioterapia | | *Tot* |
|---|---|---|---|
| | *Risposta* | *Non risposta* | |
| *Alta* | 12 | 0 | 12 |
| *Bassa* | 14 | 11 | 25 |
| *Tot* | 26 | 11 | 37 |

**Tabella 4.10: Frequenze delle osservazioni per le classi di sopravvivenza secondo la risposta alla prima linea di chemioterapia.**

| Sopravvivenza | Risposta alla chemioterapia al platino | | *Tot* |
|---|---|---|---|
| | *Risposta* | *Non risposta* | |
| *Alta* | 12 | 0 | 12 |
| *Bassa* | 5 | 20 | 25 |
| *Tot* | 17 | 20 | 37 |

**Tabella 4.11: Frequenze delle osservazioni per le classi di sopravvivenza secondo la risposta alla chemioterapia al platino.**

| | Recidiva | | | *Tot* |
|---|---|---|---|---|
| **Sopravvivenza** | *Si* | *No* | *Prog* | |
| *Alta* | 7 | 5 | 0 | 12 |
| *Bassa* | 10 | 0 | 15 | 25 |
| *Tot* | 17 | 5 | 15 | 37 |

**Tabella 4.12: Frequenze delle osservazioni per le classi di sopravvivenza secondo la presenza di recidività o meno. Progressione indica che non si è mai verificato un miglioramento, quindi non si può parlare di recidività.**

La Tabella 4.4 mostra la distribuzione congiunta di sopravvivenza e *stadiazione FIGO*, nella quale si sono aggregati gli stadi più bassi e quelli più alti. E' possibile verificare, con il test esatto di Fisher, la presenza di dipendenza tra le due variabili; tale test ha come ipotesi nulla l'assenza di dipendenza. Il p-value osservato è pari a 0.024 e ad un livello di significatività fissato al 5% l'ipotesi nulla è rifiutata.

La *dimensione del tumore residuo* dopo l'operazione è definita come massima dimensione del nodulo di maggior volume, evidenziabile alla fine della procedura chirurgica, in centimetri. La Tabella 4.5 ne mostra la distribuzione congiunta insieme alla sopravvivenza. Si sono aggregati i livelli $TR = 0$ e $0 < TR \leq 1$ a causa del basso numero di osservazioni per alcune classi. Il p-value osservato per un test esatto di Fisher è pari a 0.023 e ad un livello di significatività del 5%, anche in questo caso, l'ipotesi nulla è rifiutata.

Riguardo la *presenza di asciti* e la *positività dei linfonodi*, le tabelle di contingenza sono riportate nelle Tabelle 4.6 e 4.7. La presenza di asciti sembra avere un'influenza sulla sopravvivenza, infatti il p-value relativo alla verifica d'ipotesi svolta con test esatto di Fisher è pari a 0.0014. Riguardo la positività dei linfonodi, si hanno parecchie osservazioni mancanti. Decidendo di non considerarle, un test esatto di Fisher produce un p-value pari a 0.099, per cui ad un livello di significatività del 5% l'ipotesi nulla non viene rifiutata.

Per capire se il *CA125 preoperatorio* è legato alla sopravvivenza si può eseguire un test *t* per verificare se le distribuzioni nei due gruppi hanno la stessa media. Il p-value osservato per questo test è pari a 0.622, per cui si propende per l'accettazione dell'ipotesi nulla.

Anche per l'*età alla diagnosi* si può verificare se le distribuzioni nei due gruppi hanno la stessa media con un test *t*. Il p-value, in questo caso, è 0.353 per cui l'ipotesi nulla è accettata. Riguardo lo stato di *menopausa*, invece, un test esatto di Fisher ha p-value pari a 0.221, facendo propendere anche in questo caso per un'assenza di relazione con la sopravvivenza.

Anche la relazione tra *grado* del tumore e sopravvivenza è stata verificata tramite il test esatto di Fisher, conducendo all'accettazione dell'ipotesi nulla con un p-value osservato pari a 0.299.

Riguardo la risposta alla chemioterapia si dispone di due variabili, una che identifica il tipo di *risposta alla prima linea di chemioterapia* e l'altra che classifica la *risposta alla chemioterapia al platino*. A causa del basso numero di osservazioni per alcune classi, si è scelto di accorpare alcune delle modalità in entrambe. Nella risposta alla prima linea di chemioterapia sono state accorpate le modalità "clinica parziale" e "clinica completa", creando la nuova modalità "risposta". Una verifica d'ipotesi di assenza di dipendenza tra questa nuova variabile e la sopravvivenza delle pazienti, fatta con test esatto di Fisher, ha p-value pari a 0.0066, conducendo al rifiuto dell'ipotesi nulla. Nella variabile relativa alla risposta alla chemioterapia al platino, invece, si sono aggregate le due modalità "platino refrattaria" e "platino resistente" e anche le due modalità "platino sensibile" e "platino parzialmente sensibile", creando le due nuove modalità: "risposta" e "non risposta". Anche per questa variabile si è verificata l'ipotesi di assenza di relazione con la risposta tramite test esatto di Fisher, il quale ha p-value 3.34e-06 per la risposta alla chemioterapia al platino. Le relative tabelle di contingenza sono riportate nelle Tabelle 4.10 e 4.11, rispettivamente.

Ultima variabile a disposizione è la *recidività del tumore*, la quale indica se il tumore si è ripresentato dopo l'asportazione. Per le pazienti che non hanno mai avuto una totale asportazione la classe di appartenenza è "progressione". La distribuzione congiunta di recidività e sopravvivenza è riportata nella Tabella 4.12. In questo caso il test del Chi Quadrato ha un p-value basso, pari a 0.0001, che conduce al rifiuto dell'ipotesi nulla di assenza di relazione tra le due variabili.

Tutti i risultati ottenuti in queste analisi sono riportati nella Tabella 4.13.

Ciò che si è fatto finora è un'analisi marginale di tutte le variabili a disposizione. Questa, però, non è d'aiuto per la previsione della sopravvivenza, ma lo può essere per una selezione preventiva delle variabili da includere in un modello. A questo punto è necessario fare una riflessione su alcune delle variabili che si sono analizzate. Le due variabili relative alla risposta alla chemioterapia sono di fatto delle variabili proxy, ossia variabili sostitutive della sopravvivenza. Si ragioni, ad esempio, sul fatto che se una paziente viene curata con una chemioterapia, ma su di lei questa non ha effetto, è evidente che la sopravvivenza non potrà che essere bassa. Inoltre, la sensibilità o meno della paziente alla chemioterapia è nota solo in un momento avanzato, per cui non ha senso includere

questa variabile tra le esplicative della sopravvivenza. Anche nella variabile recidiva la categoria "progressione" è sostanzialmente legata ad una risposta o meno della paziente alle cure cui è stata sottoposta e, dunque, considerarla come una variabile esplicativa non avrebbe significato.

In definitiva, le unica variabile che potrebbe essere utili come predittori in un modello per la spiegazione della sopravvivenza sembrano essere la presenza di asciti, la dimensione del tumore residuo e lo stadio FIGO. Riguardo quest'ultima, data la presenza di frequenze nulle nella Tabella 4.4, i coefficienti stimati risulterebbero inattendibili. La dimensione del tumore residuo, invece, non risulta significativa e, dunque, è stata rimossa. Un modello logistico con la sola variabile esplicativa asciti ha le stime riportate nella Tabella 4.14.

| **Variabile** | **p-value del test** |
|---|---|
| *Stadio FIGO* | 0.024 |
| *Dimensione tumore residuo* | 0.023 |
| *Asciti* | 0.0014 |
| *Linfonodi* | 0.099 |
| *CA125 preoperatorio* | 0.622 |
| *Età alla diagnosi* | 0.353 |
| *Menopausa* | 0.221 |
| *Grado* | 0.299 |
| *Risposta alla prima linea di chemioterapia* | 0.0066 |
| *Risposta alla chemioterapia al platino* | 3.34e-06 |
| *Recidività* | 0.0001 |

**Tabella 4.13: Resoconto dei risultati ottenuti dalle analisi marginali sulle variabili cliniche. Tutti i p-value sono stati ottenuti con test esatto di Fisher.**

| **Coefficienti** | **Stima** | **Std. Error** | **p-value** |
|---|---|---|---|
| *Intercetta* | -1.2528 | 0.8018 | 0.118 |
| *Asciti: SI* | 2.8622 | 0.9396 | 0.002 |

**Tabella 4.14: Coefficienti stimati per un modello di regressione logistica con la sola variabile esplicativa "Asciti" e risposta la sopravvivenza.**

Tale modello è stato fatto considerando come successo l'evento *sopravvivenza bassa* per cui, come ci si può attendere, la presenza di asciti risulta essere un fattore di rischio significativo per la probabilità di sopravvivenza bassa. L'odds ratio, dato dal rapporto tra i due odds della sopravvivenza rispetto alla presenza di asciti e della sopravvivenza rispetto all'assenza di asciti, è pari a 17.5.

### 4.4.4. Analisi di differenziale espressione

Come spiegato nel §2.5, la fase successiva alla normalizzazione e all'eventuale aggiustamento dei dati è quella dell'identificazione dei geni differenzialmente espressi, definendo in questo modo i geni che hanno un valore di espressione significativamente diverso tra due o più condizioni. In questo caso le diverse condizioni che si vogliono analizzare sono relative alla durata della sopravvivenza, distinguendo tra pazienti con bassa e con elevata sopravvivenza. Obiettivo di questo paragrafo è, dunque, utilizzare un qualche test statistico per ogni gene nelle due condizioni ed ottenere una lista di geni differenzialmente espressi. Per fare ciò, la prima decisione da prendere riguarda la statistica test che si vuole utilizzare. Nel §2.5 e nei suoi sottoparagrafi sono state introdotte le due statistiche test più impiegate in quest'ambito, Empirical Bayes e SAM. Queste statistiche agiscono in modi diversi: SAM modera il test aggiungendo al denominatore il fudge factor $s_0$ e utilizza un approccio permutazionale per calcolare la distribuzione sotto $H_0$; Ebayes costruisce un modello lineare e sfrutta la natura parallela dell'analisi definendo delle distribuzioni a priori sui parametri. Entrambe le statistiche test sono state utilizzate nei dati aggiustati con COMBAT, in modo da produrre due diverse liste di geni differenzialmente espressi e analizzarne le differenze.

Per il test SAM si sono utilizzate mille permutazioni dei dati per calcolare la distribuzione nulla. Per quanto discusso nel §2.5.4 relativamente ai test multipli, si è deciso di adottare il FDR come metodo di controllo degli errori commessi in quanto risulta essere meno stringente rispetto al controllo del FWER. Come si è spiegato in quel paragrafo, il FDR è la frazione attesa di falsi positivi nella lista dei geni differenzialmente espressi e questa, in relazione al test SAM, è un

quantità che varia al variare della soglia $\Delta$ scelta. In particolare, all'aumentare della quantità $\Delta$ il numero di geni identificati come differenzialmente espressi diminuisce e lo stesso accade per il FDR. E' dunque possibile fissare il FDR che si è disposti a sopportare e scegliere la soglia $\Delta$ di conseguenza. Per i dati a disposizione si è scelto di fissare il FDR al 10%, identificando una soglia $\Delta = 0.66$. Così facendo si è individuata una lista di 304 geni differenzialmente espressi, dei quali 212 sono up-regolati e 92 sono down-regolati.

In Figura 4.29 è riportato il grafico relativo al test SAM. In ascissa sono presenti i valori medi $\bar{T}_{(g)}$ calcolati tramite mille permutazioni dei dati; in ordinata sono rappresentati i valori osservati della statistica. L'ampiezza delle due linee diagonali tratteggiate è $\Delta$, mentre le due linee orizzontali rappresentano i corrispondenti valori delle soglie di rifiuto del test ($t_1$ e $t_2$).

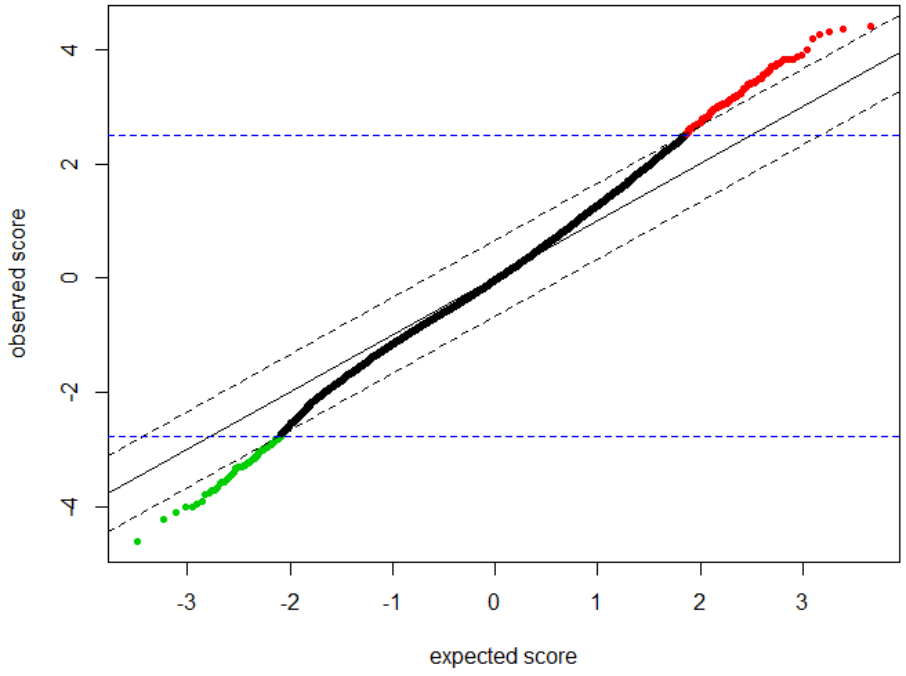



**Figura 4.29: Grafico test SAM. In ascissa sono riportati i valori medi $\bar{T}_{(g)}$ e in ordinata quelli delle statistiche originali $T_{(g)}$. Le due fasce parallele tratteggiate hanno ampiezza pari a $\Delta$, mentre le due linee orizzontali definiscono le due soglie $t_1$ e $t_2$.**

Riguardo il test Empirical Bayes, si è dapprima stimato un modello lineare per ciascun gene e si sono poi stimati i valori della statistica test sul contrasto di interesse, ossia la differenza tra pazienti con sopravvivenza alta e quelle con

sopravvivenza bassa. Anche in questo caso, trattandosi di test multipli, si è deciso di controllare il FDR. Per avere risultati confrontabili con il test precedente si è considerato un q-value pari a 0.1, il quale garantisce un FDR del 10% nella lista dei geni identificati come differenzialmente espressi. Così facendo si sono individuati 80 geni con differenziale espressione.

E' interessante notare che tutti gli 80 geni identificati dal test Ebayes sono contenuti nella lista di geni differenzialmente espressi identificata dal test SAM. Inoltre, si è verificato che l'ordinamento dei geni attribuito dal test SAM e quello derivante da Ebayes sono molto simili. Esaminando l'intersezione delle liste di tutti i geni ordinate secondo i valori dei q-value di SAM e di Ebayes si nota che tra i primi 304 geni le due liste ne hanno in comune il 92.8%, tra i primi 1000 ne hanno in comune 951 (95.1%) e tra i primi 5000 ce ne sono 4910 di comuni (98.2%).

Tenendo fede a quanto detto alla fine del §4.4.2 si è effettuata l'analisi di differenziale espressione anche sulla matrice ottenuta adottando la correzione proposta da RUV-4. I risultati, naturalmente, non sono esattamente gli stessi di quelli ottenuti sui dati corretti con COMBAT, dato che i due metodi di aggiustamento sono completamente differenti. In particolare, il test SAM su questi dati identifica ben 559 geni differenzialmente espressi, di cui 52 up-regolati e 507 down-regolati. Ebayes, invece, ne identifica 131. L'insieme risultante dall'intersezione delle liste ottenute con l'applicazione di SAM sui dati corretti con COMBAT e su quelli corretti con RUV-4 ha cardinalità 143; quello risultante dall'intersezione delle liste ottenute con Ebayes ha cardinalità 64. Questi risultati sono riportati nella sottostante Tabella 4.15, nella quale è evidente la differenza di percentuale di intersezione tra le due statistiche test. In particolare, il test Ebayes effettuato sui dati corretti con RUV-4 identifica l'80% dei geni identificati sui dati corretti con COMBAT. Con il test SAM, invece, solo il 47% dei geni identificati sui dati corretti con RUV-4 è presente anche tra quelli identificati sui dati corretti con COMBAT. Ciò è certamente dovuto alla forte asimmetria tra numero di geni identificati come sovra e sotto espressi da SAM nei dati corretti con RUV-4: sono infatti solo 52 i geni sovra espressi e 507 i sotto espressi; in netto contrasto con i 213 sovra espressi e i 92 sotto espressi identificati dallo stesso test con la correzione di COMBAT.

| Statistica Test | N° DEG COMBAT | N° DEG RUV-4 | Cardinalità Intersezione | % |
|---|---|---|---|---|
| SAM | 304 | 559 | 143 | 47% |
| Ebayes | 80 | 131 | 64 | 80% |

**Tabella 4.15: Numero di geni differenzialmente espressi per entrambi i metodi di correzione ed entrambe le statistiche test, con indicazione della cardinalità dell'intersezione tra le due liste e percentuale di geni identificati sui dati corretti con COMBAT che compaiono anche con la correzione di RUV-4**

| Simbolo | logFC | adj.P.Val | B | RUV-4 |
|---|---|---|---|---|
| C6orf62 | 0,653 | 0,0443 | 3,609 | ✓ |
| NUAK1 | -1,179 | 0,0443 | 3,529 | ✓ |
| BTN3A3 | 0,855 | 0,0491 | 2,917 | ✓ |
| MRS2 | 0,788 | 0,0491 | 2,630 | ✓ |
| CXCL11 | 1,974 | 0,0491 | 2,333 | ✓ |
| DEPTOR | 1,486 | 0,0491 | 2,268 | ✓ |
| AEBP1 | -1,322 | 0,0491 | 2,150 | ✓ |
| TDP2 | 0,669 | 0,0491 | 2,144 | ✓ |
| COL16A1 | -0,761 | 0,0491 | 2,018 | ✓ |
| C1orf109 | 0,515 | 0,0491 | 1,879 | ✓ |
| UROD | 0,534 | 0,0491 | 1,816 | ✓ |
| PPCS | 0,601 | 0,0491 | 1,652 | ✓ |
| VCAN | -1,730 | 0,0491 | 1,642 | ✓ |
| PDGFRB | -0,824 | 0,0491 | 1,614 | ✓ |
| AHSA1 | 0,634 | 0,0491 | 1,613 | ✓ |
| PCOLCE | -1,126 | 0,0491 | 1,604 | ✓ |
| FSTL3 | -0,528 | 0,0491 | 1,575 | ✓ |
| KIAA1324 | 0,921 | 0,0495 | 1,429 | ✓ |
| HSD17B1 | 0,614 | 0,0495 | 1,380 | ✓ |
| PRKCDBP | -0,696 | 0,0495 | 1,344 | ✓ |
| UBE2K | 0,555 | 0,0495 | 1,338 | × |
| ANO1 | 0,991 | 0,0495 | 1,277 | ✓ |
| FXYD5 | -1,397 | 0,0495 | 1,270 | ✓ |
| TMEM134 | 0,670 | 0,0495 | 1,266 | ✓ |
| C8orf33 | 0,876 | 0,0499 | 1,224 | ✓ |

**Tabella 4.16: Lista dei geni differenzialmente espressi identificati tramite il test Ebayes sui dati corretti con COMBAT con un FDR del 5%. La seconda colonna è il logaritmo del fold chance nelle due condizioni di alta e bassa sopravvivenza; le terza colonna è il p-value aggiustato con il metodo FDR; la quarta contiene i logaritmi degli odds ratio che il gene sia differenzialmente espresso (si veda la formula 2.31 del paragrafo 2.5.3). L'ultima colonna è un flag che mostra quali dei geni sono stati identificati con Ebayes anche nei dati corretti con RUV-4.**

Nel §A.6 sono riportati ulteriori risultati riguardo la relazione presente tra i valori delle statistiche test con i due metodi di aggiustamento.

Nel §A.7 è riportata la lista degli 80 geni differenzialmente espressi identificati con il test Ebayes sui dati corretti con COMBAT. Inoltre, per ognuno di questi, è presente un flag che indica l'appartenenza o meno alla lista dei geni differenzialmente espressi identificati con Ebayes nei dati corretti con RUV-4. Un estratto di quella tabella, contenente solo i geni aventi un p-value aggiustato minore di 0.05 è riportato nella sottostante Tabella 4.16.

### 4.4.5. Gene set analysis

Come spiegato nel §2.6 la fase successiva a quella di identificazione dei geni differenzialmente espressi è la gene set analysis. A seconda del metodo che si vuole utilizzare, questa può sfruttare i risultati ottenuti in fase di identificazione dei geni differenzialmente espressi (metodi competitivi) oppure può essere svolta indipendentemente da questi, utilizzando l'intera matrice dei valori di espressione (metodi indipendenti); inoltre può guadagnare potenza dalle informazioni contenute nei pathway (metodi topologici) oppure considerare questi come semplici liste di geni (metodi non topologici). In questo paragrafo si vogliono mostrare i risultati ottenuti dall'applicazione delle quattro tecniche di analisi di gene set spiegate nel §2.6, ossia l'analisi di arricchimento classica, il Global Test, la Signaling Pathway Impact Analysis (SPIA) e CliPPER. Questi sono stati scelti per le loro differenti caratteristiche, sapendo che alcuni riusciranno a cogliere aspetti non colti da altri e sperando che l'unione dei risultati ottenuti possa dare qualche indicazione sulle possibili classi funzionali o sui pathway coinvolti nei processi biologici che spiegano la diversa sopravvivenza delle pazienti affette da tumore all'ovaio.

La distinzione che genera la maggiore differenza tra tutte le analisi di gene set è quella tra metodi competitivi e metodi indipendenti, proprio a causa della diversa ipotesi nulla alla base dei test che queste analisi effettuano. L'analisi di arricchimento è l'unico metodo competitivo, dato che, in realtà, SPIA è un metodo misto. E' risaputo, però, che il punteggio di perturbazione del pathway

calcolato da SPIA (pPERT), il quale rappresenta la "componente indipendente" dell'analisi, è solitamente poco influente rispetto alla componente "competitiva" (pNDE) dello stesso; per questo motivo i risultati che SPIA produce non sono solitamente molto diversi da quelli dell'analisi di arricchimento. Global test e CliPPER sono, invece, metodi indipendenti. In definitiva, ci si attende che l'analisi di arricchimento e SPIA conducano a risultati abbastanza simili tra loro ma, probabilmente, differenti rispetto a quelli del Global Test e di CliPPER, i quali a loro volta dovrebbero essere abbastanza concordi tra loro nei risultati.

Per effettuare le suddette analisi si è utilizzato un web server pubblico chiamato Graphite Web, nato per l'analisi di dati di espressione genica e la visualizzazione di pathway biologici. Graphite Web permette di eseguire tutte le quattro gene set analysis nominate in questo elaborato, offrendo una visualizzazione dei pathway che mira a semplificare l'interpretazione dei risultati ottenuti. L'applicativo è nato grazie ad un gruppo di ricercatori del dipartimento di Biologia dell'università di Padova ed il suo funzionamento è descritto in Sales et. al. (2013).

L'interpretazione dei risultati prodotti dalle analisi di gene set non è per niente banale e prevede non solo approfondimenti sulle funzioni che sono rappresentate dai pathway individuati, ma anche la ricerca in letteratura di comportamenti analoghi riscontrati in altre condizioni simili. Per questi motivi si è voluto individuare, innanzitutto, i pathway comuni ai metodi utilizzati per le analisi di gene set, tenendo presente le considerazioni fatte riguardo le grosse differenze tra tipologie di metodi.

Per l'analisi di arricchimento e per SPIA, i quali sfruttano la lista dei geni identificati in precedenza come differenzialmente espressi, si è deciso di utilizzare quelli identificati dal test SAM, in quanto di numerosità più elevata. Inoltre si ricordi che SAM ed Ebayes, nei dati corretti con il metodo COMBAT, danno risultati tra loro concordanti, dato che tutti i geni dichiarati differenzialmente espressi con SAM, anche se non significativi per Ebayes, hanno un ranking alto nella lista dei geni ordinata per valore della statistica test.

L'analisi di arricchimento conduce all'individuazione di un solo pathway modificato significativamente tra le due condizioni di sopravvivenza alta e bassa delle pazienti, il quale è l'unico rilevato anche da SPIA. Il Global Test ne

identifica 50, mentre CliPPER ne identifica 67. L'unico pathway individuato dall'analisi di arricchimento e da SPIA è presente anche tra quelli individuati dal Global Test, ma non tra quelli trovati da CliPPER. Tutte le analisi sono state svolte considerando un q-value del 10%. La Figura 4.30 mostra tramite un diagramma di Venn i pathway comuni tra quelli identificati dal Global Test e quelli individuati da CliPPER. Questi, unitamente a quello rilevato dall'analisi di arricchimento, da SPIA e dal Global Test, ma non da CliPPER, sono tutti riportati nella Tabella 4.17.

Come fatto in fase di identificazione dei geni differenzialmente espressi, si è voluto verificare se i pathway rilevati nei dati corretti con COMBAT fossero circa gli stessi di quelli rilevati sui dati corretti con RUV-4. Per questo è stata aggiunta una colonna con un flag. Come accade per i dati corretti con COMBAT, il pathway ECM-receptor interaction viene rilevato da tutti i metodi tranne che da CliPPER, mentre per decidere se il flag doveva essere positivo o negativo si sono considerati solo i pathway derivanti dall'intersezione tra Global Test e CliPPER, esattamente come si era proceduto sui dati corretti con COMBAT.

Sui pathway riportati nella suddetta tabella sono stati fatti degli approfondimenti per verificare se, in letteratura, siano presenti studi con caratteristiche simili a quello svolto in questo elaborato e che abbiano portato ad identificare qualcuno di questi.

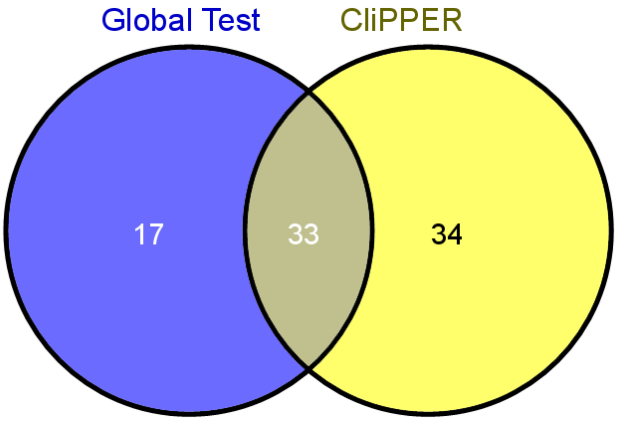



**Figura 4.30: Diagramma di Venn degli insiemi di geni identificati dal Global Test e da CliPPER.**

| Nome Pathway | q-value Global Test | α-Mean | α-Var | RUV-4 |
|---|---|---|---|---|
| ECM-receptor interaction | 0,036 | - | - | ✓ |
| Axon guidance | 0,036 | 0,01 | 0,50 | ✓ |
| beta-Alanine metabolism | 0,036 | 0,00 | 0,72 | ✓ |
| Small cell lung cancer | 0,036 | 0,02 | 0,69 | ✓ |
| Toll-like receptor signaling pathway | 0,036 | 0,00 | 0,70 | ✓ |
| Carbohydrate digestion and absorption | 0,039 | 0,00 | 0,93 | ✓ |
| Epithelial cell signaling in Helicobacter pylori infection | 0,039 | 0,01 | 0,87 | ✓ |
| Glycolysis / Gluconeogenesis | 0,039 | 0,01 | 1,00 | ✓ |
| NF-kappa B signaling pathway | 0,039 | 0,00 | 0,32 | ✓ |
| Terpenoid backbone biosynthesis | 0,039 | 0,00 | 0,09 | ✓ |
| TGF-beta signaling pathway | 0,039 | 0,00 | 0,57 | ✓ |
| Toxoplasmosis | 0,048 | 0,03 | 0,89 | ✓ |
| Gap junction | 0,056 | 0,01 | 0,61 | ✓ |
| Fructose and mannose metabolism | 0,058 | 0,01 | 0,59 | ✓ |
| Legionellosis | 0,058 | 0,02 | 0,12 | ✓ |
| Nicotinate and nicotinamide metabolism | 0,058 | 0,00 | 0,64 | ✓ |
| Prostate cancer | 0,058 | 0,00 | 0,96 | ✓ |
| Type II diabetes mellitus | 0,058 | 0,04 | 0,67 | ✓ |
| Bacterial invasion of epithelial cells | 0,062 | 0,00 | 1,00 | ✓ |
| Folate biosynthesis | 0,064 | 0,00 | 0,39 | ✗ |
| Complement and coagulation cascades | 0,065 | 0,03 | 0,67 | ✓ |
| Glycosylphosphatidylinositol (GPI)-anchor biosynthesis | 0,065 | 0,02 | 0,77 | ✗ |
| Jak-STAT signaling pathway | 0,065 | 0,00 | 1,00 | ✓ |
| Leukocyte transendothelial migration | 0,065 | 0,00 | 0,28 | ✓ |
| Pancreatic cancer | 0,065 | 0,01 | 0,96 | ✓ |
| Shigellosis | 0,065 | 0,00 | 1,00 | ✓ |
| Wnt signaling pathway | 0,065 | 0,00 | 0,08 | ✓ |
| VEGF signaling pathway | 0,072 | 0,01 | 0,85 | ✓ |
| Melanoma | 0,075 | 0,00 | 1,00 | ✓ |
| Pantothenate and CoA biosynthesis | 0,075 | 0,01 | 0,62 | ✗ |
| Apoptosis | 0,090 | 0,00 | 0,45 | ✓ |
| p53 signaling pathway | 0,092 | 0,00 | 0,68 | ✓ |
| Non-small cell lung cancer | 0,096 | 0,05 | 0,98 | ✓ |
| Tryptophan metabolism | 0,101 | 0,02 | 0,13 | ✗ |

**Tabella 4.17: Lista dei 34 pathway risultati significativamente modificati nelle due condizioni di alta e bassa sopravvivenza. Il primo è il pathway comune a 3 metodi: analisi di arricchimento, SPIA e Global Test. Tutti gli altri risultano dall'intersezione tra la lista dei pathway trovati con il Global Test e quelli trovati con CliPPER. La seconda colonna è il q-value del Global Test, mentre la terza è il p-value relativo alla differenza media di espressione tra i geni del pathway nelle due condizioni; la quarta colonna è relativa alla differenza nelle correlazioni tra i geni del pathway nelle due condizioni.**

Le liste di tutti i pathway risultati significativi per ognuno dei metodi utilizzati sono riportate in appendice (§A.8).

I pathway individuati sono coerenti con i risultati noti in letteratura sulla progressione tumorale. Un coinvolgimento del pathway relativo alla Matrice Extracellulare (ECM) è stato dimostrato anche nello sviluppo del tumore alla mammella (Emery, et al., 2009), mentre sono già note sregolazioni di molti geni coinvolti nel pathway NF-kappa B durante lo sviluppo e la progressione del tumore all'ovaio. Ci sono cinque proteine della famiglia NF-kB che controllano diversi processi chiave richiesti per lo sviluppo dei tumori, come l'attivazione dei geni anti-apoptotici, per cui la cellula tumorale non muore ed è libera di riprodursi (Alvero, 2010). Il coinvolgimento del pathway TGF-beta, poi è compatibile con ciò che avviene solitamente nelle cellule tumorali; questa famiglia di proteine, infatti, è responsabile della proliferazione cellulare ed è stata rilevata in casi di tumore all'ovaio da Yeung et. al. (2013).

# Capitolo 5

# 5. Potere discriminante dei geni differenzialmente espressi

In ambito medico lo studio dei marcatori genici è di grande importanza. I marcatori sono geni che hanno un elevato potere predittivo e prognostico. A questo scopo vengono usati spesso modelli di regressione. In ambito di analisi di microarray, nelle quali il numero di variabili (geni) supera enormemente il numero di osservazioni, si rende necessario l'utilizzo di tecniche particolari che permettono di ottenere delle stime affidabili dei parametri di regressione di ogni variabile. In quest'ottica sono state sviluppate delle tecniche di regressione che sono basate sulla penalizzazione della funzione obiettivo da minimizzare per l'ottenimento delle stime. Se l'introduzione di una penalità, da un lato, comporta una distorsione delle stime dei parametri, da un altro, però, porta ad una diminuzione della varianza delle stesse che può tradursi in una diminuzione dell'errore quadratico medio (MSE). Una di queste tecniche, denominata lasso, è in grado anche di fare una selezione automatica delle variabili da includere nel modello.

In questo paragrafo sarà introdotto il metodo lasso e la sua generalizzazione al modello logistico. Poi si procederà a stimare il modello sui dati a disposizione, con l'obiettivo di identificare i geni, tra i differenzialmente espressi, che meglio predicono la diversa sopravvivenza delle pazienti.

## 5.1. Regressione lasso

Il lasso (Tibshirani, 1996) è una tecnica di regressione che penalizza la grandezza dei coefficienti, con l'obiettivo di ottenere una selezione automatica dei

predittori da includere nel modello. Ha avuto un notevole sviluppo soprattutto perché permette di ottenere stime dei parametri di regressione anche quando il numero di variabili è maggiore del numero di osservazioni a disposizione, assegnando ad alcuni dei coefficienti di regressione un valori pari a zero.

Per un modello di regressione lineare, la soluzione lasso è definita nel modo seguente:

$$\hat{\beta}^{lasso} = \text{argmin}_{\beta} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$$
$$\text{soggetto a:} \quad \sum_{j=1}^{p}|\beta_j| \leq t \tag{5.1}$$

Dove $y_i$ rappresenta l'i-esimo valore della variabile risposta e $x_{ij}$ è l'i-esimo valore del j-esimo predittore.

La precedente formula può essere riscritta in forma di Lagrange come:

$$\hat{\beta}^{lasso} = \text{argmin}_{\beta} \left\{\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{5.2}$$

La penalità adottata rende impossibile l'ottenimento di un risultato lineare in *y* e, dunque, la soluzione deve essere ottenuta tramite algoritmi di programmazione quadratica.

Un metodo che permette di ottenere la stima cercata è detto Pathwise Coordinate Descent Optimization. L'idea è di fissare il parametro di penalizzazione λ della (5.2) e ottimizzare ciclicamente per ogni parametro, mantenendo gli altri fissati al loro valore corrente. Supponendo che tutte le variabili esplicative siano state standardizzate con media nulla e norma unitaria, sia $\tilde{\beta}_k(\lambda)$ la stima corrente di $\beta_k$ fissato il parametro di penalizzazione λ. La (5.2) può essere riscritta come:

$$R\left(\tilde{\beta}(\lambda), \beta_j\right) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{k\neq j} x_{ik}\tilde{\beta}_k(\lambda) - x_{ij}\beta_j\right)^2 + \lambda\sum_{k\neq j}|\tilde{\beta}_k(\lambda)| + \lambda|\beta_j| \tag{5.3}$$

che può essere visto come un problema di lasso univariato dove la variabile risposta è data dai residui parziali:

$$y_i - \tilde{y}_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\tilde{\beta}_k(\lambda) \tag{5.4}$$

Questo problema ha soluzione che porta all'aggiornamento:

$$\tilde{\beta}_j(\lambda) = S\left(\sum_{i=1}^{n} x_{ij}\left(y_i - \tilde{y}_i^{(j)}\right), \lambda\right) \tag{5.5}$$

dove $S(t, \lambda) = sign(t)(|t| - \lambda)_+$ è chiamato operatore *soft-thresholding*.

Iterando ripetutamente la (5.5) per ogni variabile fino a convergenza, si ottiene la stima lasso.

L'algoritmo viene utilizzato per la stima del lasso in una griglia di valori di λ. Si parte dal più piccolo valore $\lambda_{max}$ tale per cui $\hat{\beta}(\lambda_{max}) = 0$; poi si fa decrescere λ e si ripete il ciclo su tutte le variabili fino a convergenza. Poi λ viene ridotto nuovamente e il processo è ripetuto utilizzando le soluzioni precedenti di $\hat{\beta}$ per il nuovo valore di λ.

## 5.2. Regressione logistica regolarizzata

Quando la variabile risposta è binaria, il modello lasso descritto nel paragrafo precedente può essere generalizzato.

Sia $G$ la variabile risposta, la quale assume valori $g \in \{0,1\}$. Definito $\pi(x) = \Pr(G = 1|x)$, il modello logistico utilizza la funzione legame *logit* per collegare $\pi$ al predittore lineare $\beta_0 + x^T\beta$. Si ha dunque:

$$\log\frac{\pi(x)}{1-\pi(x)} = \beta_0 + x^T\beta \tag{5.6}$$

o, equivalentemente:

$$\pi(x) = \frac{\exp(\beta_0+x^T\beta)}{1+\exp(\beta_0+x^T\beta)} \tag{5.7}$$

Supponendo di disporre di *n* osservazioni, detta $\pi(x_i)$ la probabilità (5.7) per l'i-esima osservazione, si massimizza la log-verosimiglianza penalizzata:

$$\max_{\beta_0,\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\pi(x_i) + (1-y_i)\log\big(1-\pi(x_i)\big)\right] - \lambda P(\beta)\right\} \tag{5.8}$$

nella quale la parte relativa alla log-verosimiglianza può essere riscritta come:

$$l(\beta_0,\beta) = \frac{1}{n}\sum_{i=1}^{n}[y_i \cdot (\beta_0 + x_i^T\beta) - \log(1+\exp(\beta_0 + x_i^T\beta))] \tag{5.9}$$

Supponendo che le stime dei parametri siano $(\tilde{\beta}_0, \tilde{\beta})$, per ogni valore di λ, viene calcolata un'approssimazione quadratica della verosimiglianza, $l_Q$, e viene utilizzato l'algoritmo Coordinate Descent Optimization introdotto nel paragrafo precedente per risolvere il problema di minimo:

$$\min_{\beta_0,\beta}\{-l_Q(\beta_0,\beta) + \lambda P(\beta)\} \tag{5.10}$$

In pratica, la determinazione della soluzione richiede tre cicli diversi:

*Ciclo esterno*: decrementa λ.

*Ciclo intermedio*: aggiorna l'approssimazione quadratica $l_Q$ utilizzando le stime correnti $(\tilde{\beta}_0, \tilde{\beta})$.

*Ciclo interno*: esegui l'algoritmo Coordinate Descent sulla quantità (5.10).

## 5.3. Analisi sui DEG identificati da Ebayes

Il modello di regressione logistica regolarizzata delineato nel paragrafo precedente può tornare molto utile in un'ottica di individuazione dei geni con maggiore potere discriminante tra i differenzialmente espressi. Considerando gli 80 geni differenzialmente espressi identificati con la statistica test Ebayes sui dati corretti con il metodo COMBAT, è possibile definire un modello di regressione logistica regolarizzata per l'individuazione dei predittori maggiormente utili per la previsione della variabile risposta *sopravvivenza*. Il lasso, infatti, permette di fare una selezione automatica delle variabili esplicative, ponendo a zero le stime dei coefficienti meno influenti.

Il modello è stato stimato sui valori di espressione corretti disponibili per gli 80 geni differenzialmente espressi. Naturalmente, per ottenere le stime dei coefficienti di regressione, è necessario decidere il λ più adatto, il quale è stato scelto tramite cross validation leave one out. Inoltre, volendo ricavare una stima dell'errore di previsione, si è dovuto effettuare una seconda cross validation leave one out. L'utilizzo di questa procedura ha permesso di ottenere una matrice delle stime di tutti gli 80 coefficienti di regressione in 39 modelli diversi, ottenuti escludendo dall'insieme di stima una osservazione per volta. Così facendo si sono potuti individuare i geni maggiormente utilizzati nei modelli per la previsione della sopravvivenza, ossia quelli con coefficienti di regressione diversi da zero.

Le analisi hanno rivelato che i geni utili per la previsione della sopravvivenza sono 48. Questo risultato è stato ottenuto osservando, per tutti i 39 modelli costruiti, per quali geni sia stato stimato un coefficiente di regressione diverso da zero. Naturalmente, tra questi 48, alcuni vengono utilizzati in pochi modelli, altri in tutti o quasi. La distribuzione delle frequenze è riportata nel diagramma a barre della Figura 5.1.

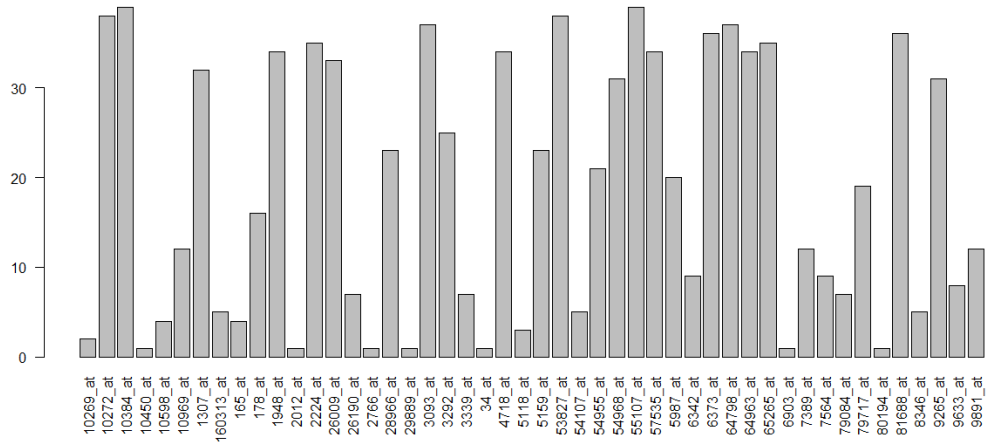



**Figura 5.1: Diagramma a barre delle frequenze osservate di apparizione dei geni differenzialmente espressi identificati con Ebayes sui dati corretti con COMBAT (solo quelli con frequenza maggiore o uguale a uno) nei 39 modelli ottenuti con CV leave one out.**

E' noto che, nei casi in cui vi sia multicollinearità tra le variabili esplicative, il lasso tende a portare a zero il coefficiente di regressione di tutti i predittori correlati, mantenendone solamente uno (Friedman, et al., 2010). Malgrado ciò, sembra sensato affermare che i geni utilizzati in un numero sufficientemente alto di modelli siano quelli con potere discriminante più alto. Per questo motivo si è deciso di fissare una soglia minima sul numero di utilizzi; i geni utilizzati in un numero di modelli inferiore a 30 verranno tralasciati. Così facendo si ottengono 18 geni con buon potere discriminate tra le due condizioni di sopravvivenza. La Tabella 5.1 riporta i nomi di questi geni, ordinati per frequenza di apparizione nei modelli. Riguardo l'errore calcolato con la procedura di cross validation bisogna considerare che, dato che il lasso fa una selezione automatica delle variabili esplicative da includere nel mdoello, l'errore è relativo all'utilizzo di tutti i predittori che si sono considerati, anche quelli utilizzati una sola volta. L'errore globale è risultato pari al 20.5% con un tasso di falsi positivi molto basso, pari al 7%.

Lo scopo dell'analisi, però, è di individuare dei possibili marcatori per la discrimanazione della sopravvivenza, per cui si è deciso di stimare un semplice modello logistico con variabili esplicative i geni risultati maggiormente discriminanti nel modello lasso. Si sono, dunque, considerati solamente i primi

due geni della Tabella 5.1, i quali sono gli unici che sono stati inseriti in tutti i 39 modelli logistici penalizzati. Le stime dei coefficienti di tale modello sono riportate nella Tabella 5.2. Si consideri che il modello è stato costruito considerando come successo l'evento "bassa sopravvivenza". Entrambi i coefficienti relativi ai geni risultano significativi ad un livello del 5%. In particolare, tutti e due sono negativi, ad indicare che l'aumento dell'espressione dei due geni in questione costituisce un fattore protettivo contro la bassa sopravvivenza; ossia l'aumento dell'espressione indica un aumento della probabilità di alta sopravvivenza. Potrebbe, dunque, essere sensato affermare che un'alta espressione di questi geni sia un sintomo di risposta alla malattia da parte della paziente e che, quindi, questi siano possibili marcatori per la sopravvivenza.

Per ottenere una stima dell'errore di classificazione di questo modello si è utilizzata una cross validation leave one out. Ciò permette anche di costruire un grafico molto usato nelle analisi di bontà dei modelli di classificazione, ossia la curva ROC. Quella ottenuta per questi dati è riportata in Figura 5.2.

| Simbolo | Frequenza |
|---|---|
| **BTN3A3** | 39 |
| **ANO1** | 39 |
| **FSTL3** | 38 |
| **FXYD5** | 38 |
| **UBE2K** | 37 |
| **DEPTOR** | 37 |
| **CXCL11** | 36 |
| **C6orf62** | 36 |
| **FDPS** | 35 |
| **C8orf33** | 35 |
| **EFNB2** | 34 |
| **NDUFC2** | 34 |
| **KIAA1324** | 34 |
| **MRPS11** | 34 |
| **ZZZ3** | 33 |
| **COL16A1** | 32 |
| **TMEM70** | 31 |
| **CYTH3** | 31 |

**Tabella 5.1: Simboli ufficiali dei 18 geni risultati maggiormente discriminanti tra le classi di sopravvivenza tramite modello logistico regolarizzato lasso.**

La soglia utilizzata sulla probabilità di successo stimata dal modello per la decisione della classe cui assegnare un'unità è stata scelta pari a 0.5. Gli errori commessi dal modello sono calcolabili dalla Tabella 5.3. Si nota che il tasso globale d'errore, ottenuto tramite cross validation leave one out, è pari al 10.3%; i tassi di falsi positivi e falsi negativi sono pari, rispettivamente, a 7.4% e 16.7%.

| Coefficienti | Stima | Std. Error | p-value |
|---|---|---|---|
| *Intercetta* | 44.9 | 15.00 | 0.0028 |
| *BTN3A3* | -4.29 | 1.67 | 0.0104 |
| *ANO1* | -2.15 | 0.88 | 0.0143 |

**Tabella 5.2: Stime dei coefficienti di un modello logistico per la spiegazione della sopravvivenza utilizzando come predittori i valori di espressione dei primi due geni della Tabella 5.1.**

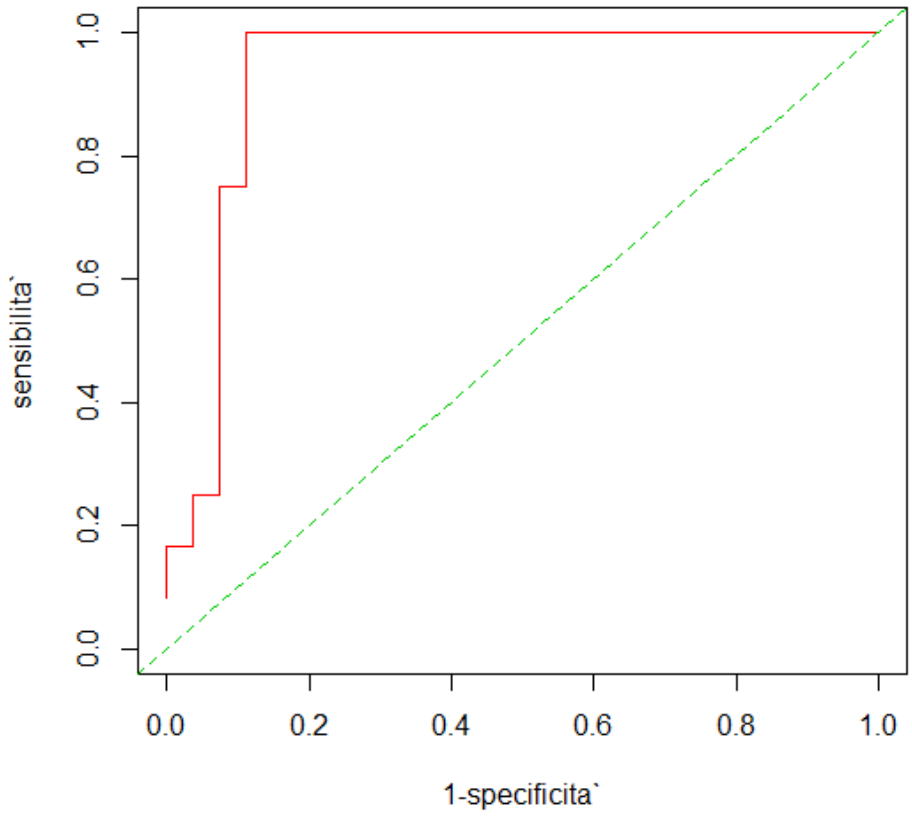



**Figura 5.2: Curva ROC ottenuta tramite cross validation per il modello logistico per la spiegazione della sopravvivenza utilizzando come predittori i valori di espressione dei primi due geni della Tabella 5.1.**

| **Previsti** | **Osservati** | |
|---|---|---|
| | *Bassa* | *Alta* |
| *Bassa* | 24 | 3 |
| *Alta* | 3 | 9 |
| *Tot* | 27 | 12 |

**Tabella 5.3: Tabella di errata classificazione ottenuta tramite cross validation per il modello logistico per la spiegazione della sopravvivenza utilizzando come predittori i valori di espressione dei primi due geni della Tabella 5.1.**

# Capitolo 6

# 6. Conclusioni

Questa tesi si è focalizzata su due aspetti fondamentali: (i) stima e rimozione della variabilità causata da fattori non biologici (batch effect) e (ii) studio dei profili di espressione di campioni biologici su delle pazienti affette da tumore all'ovaio.

Riguardo il primo aspetto, si sono passati in rassegna tre metodi: COMBAT, RUV-2 e RUV4, trattandone assunti, modalità di funzionamento e differenze. Ci si è poi concentrati sul secondo aspetto, ossia sull'analisi dei campioni biologici. Questi erano affetti da variazioni non biologiche causate dall'utilizzo di due piattaforme Affymetrix differenti e, dunque, per l'eliminazione di questa variabilità si sono dovuti applicare i metodi di correzione sopra citati. L'obiettivo delle analisi biologiche è stato quello di identificare le differenze in termini di espressione genica tra pazienti affette da tumore all'ovaio con sopravvivenza alta e con sopravvivenza bassa.

L'analisi ha potuto trarre beneficio dall'avere a disposizione osservazioni rilevate su entrambe le piattaforme per un piccolo sottoinsieme di campioni. In particolare, per cinque pazienti si disponeva dei valori di espressione sia per la piattaforma Affymetrix GC U133A che per la piattaforma Affymetrix GC U133 Plus 2. Ciò ha permesso di condurre delle analisi approfondite sulla qualità del funzionamento dei metodi di rimozione del batch effect che nella maggior parte degli studi di questo tipo non possono essere eseguite.

Analizzando solamente i dati ripetuti si è dapprima stimata e rimossa, con tutti e tre i metodi di correzione, la componente di variabilità non biologica presente, procedendo poi alla verifica della bontà degli aggiustamenti tramite cluster analysis. Si è potuto notare che, nei dati non corretti, la variabilità non biologica è talmente presente da oscurare qualsiasi variabilità biologica di

interesse, tanto che con una cluster analysis si sono individuati due gruppi ben distinti di campioni esattamente corrispondenti alla suddivisione delle osservazioni nelle due piattaforme. Dopo l'applicazione di ognuna delle tre correzioni, invece, la variabilità non biologica viene rimossa e lo stesso tipo di cluster analysis è in grado di identificare gruppi maggiormente corrispondenti alle coppie di osservazioni relative alla stessa paziente. In particolare, COMBAT commette un solo errore nella creazione delle coppie mentre RUV-2 e RUV-4, oltre a commettere il medesimo errore, non identificano un gruppo univoco per una delle coppie di osservazioni ma le aggregano ad un altro gruppo a distanze diverse.

Dalle analisi esplorative effettuate per la verifica della qualità degli esperimenti, però, si è rilevato che proprio i campioni mal identificati dalle analisi dei cluster sono quelli aventi qualità più bassa e ciò può spiegare la causa degli errori riscontrati negli appaiamenti. Si sono, quindi, valutati altri metodi per la verifica della bontà delle correzioni effettuate, come il miglioramento dei grafici RLE, la distribuzione delle varianze dei geni di controllo e la distribuzione dei p-value risultanti da test Ebayes per campioni appaiati. Tutte queste analisi condotte sui dati ripetuti in doppia piattaforma hanno portato alla conclusione che, per questi, il metodo di aggiustamento migliore sia COMBAT, seguito da RUV-4 e, poi, da RUV-2.

Eseguite le suddette analisi sui dati duplicati, avendo identificato il metodo di correzione migliore, si è potuto procedere con lo studio di tutti i campioni che si hanno a disposizione. Si tratta di 39 esperimenti: 25 effettuati con la piattaforma U133A e 14 con la piattaforma U133 Plus 2. Anche per questi si è condotta un'analisi della qualità, non rilevando nessun campione con qualità scarsa che potesse in qualche modo compromettere i risultati. Inoltre, sono state condotte specifiche analisi per l'identificazione del miglior metodo di aggiustamento del batch effect per l'insieme di tutti i dati, le quali hanno portato alla conclusione che sia COMBAT che RUV-4 producono buone correzioni. RUV-2, invece, per questi dati, sembra non essere adeguato.

Il metodo di correzione che si è quindi deciso di utilizzare sui dati completi è COMBAT, considerando anche le migliori performance ottenute sui campioni ripetuti nelle due piattaforme. Malgrado ciò, si sono sempre ottenuti tutti i risultati

anche con la correzione fatta da RUV-4, mettendo in luce differenze e somiglianze rispetto ai risultati ottenuti con COMBAT. Ciò che si è rilevato è una discreta concordanza nella maggior parte dei risultati. In particolare, nella fase di ottenimento delle liste dei geni differenzialmente espressi tra le condizioni di sopravvivenza alta e bassa di pazienti affette da tumore ovarico, nella quale si sono utilizzate due diverse statistiche test (SAM ed Ebayes), i risultati ottenuti dai due metodi sono stati abbastanza concordi rispetto alla statistica Ebayes, ma meno rispetto alla statistica SAM. Malgrado ciò, nella fase di gene set analysis si sono ottenuti risultati molto simili.

Per dare una valenza più clinica ai risultati identificando i geni con potere predittivo più alto, si è deciso di concludere le analisi effettuando anche una regressione logistica penalizzata sui geni differenzialmente espressi individuati dal test Ebayes nei dati corretti con COMBAT. Per fare ciò si è utilizzato il modello lasso, il quale ha la caratteristica di effettuare una selezione automatica delle variabili esplicative (geni) da includere nel modello di regressione, ponendo a zero i coefficienti relativi a variabili non predittive. L'utilizzo di un modello di questo tipo è stato dettato dal fatto che si disponeva di un numero di variabili maggiore rispetto al numero di osservazioni e, dunque, i modelli classici di regressione logistica non potevano essere utilizzati.

Per la stima del modello appena introdotto e per avere una misura dell'errore si è utilizzata la tecnica della cross validation leave one out. I risultati sono stati molto buoni ed hanno permesso di individuare due geni con alto potere predittivo della sopravvivenza.

Da un punto di vista biologico, più che all'analisi dei singoli geni risultati differenzialmente espressi nelle due condizioni biologiche, si è proceduto ad un'analisi dei pathway risultati alterati. Tutti i pathway individuati sono coerenti con i risultati noti in letteratura sulla progressione tumorale. Un coinvolgimento del pathway relativo alla Matrice Extracellulare (ECM) è stato dimostrato anche nello sviluppo del tumore alla mammella (Emery, et al., 2009), mentre sono già note sregolazioni di molti geni coinvolti nel pathway NF-kappa B durante lo sviluppo e la progressione del tumore all'ovaio. Ci sono cinque proteine della famiglia NF-kB che controllano diversi processi chiave richiesti per lo sviluppo dei tumori, come l'attivazione dei geni anti-apoptotici, per cui la cellula tumorale

non muore ed è libera di riprodursi (Alvero, 2010). Il coinvolgimento del pathway TGF-beta, poi è compatibile con ciò che avviene solitamente nelle cellule tumorali; questa famiglia di proteine, infatti, è responsabile della proliferazione cellulare ed è stata rilevata in casi di tumore all'ovaio da Yeung et. al. (2013).

Il risultato più interessante riguarda la lista di geni con valore predittivo. Tra questi, i due geni con massimo score sono BTN3A3 e ANO1. La proteina BTN3A3 fa parte della famiglia delle butirofiline e l'espressione della isoforma BTN3A2 in cellule tumorali è stata dimostrato molto recentemente in una piccola coorte di pazienti di stadio avanzato. Questa ulteriore conferma in una coorte indipendente ne rafforza la sua capacità prognostica. L'anoctamina 1 (ANO1), invece, è un gene localizzato nell'amplicone 11q13, una delle regioni cromosomiche maggiormente alterate durante i processi di tumorigenesi, il cui ruolo nei processi tumorali non è ancora chiaro. Recentemente è stata notata una correlazione tra espressione di ANO1 e prognosi in tumore al seno (che si sa avere molte similarità col tumore all'ovaio), ma il meccanismo per cui questo avviene non è stato dimostrato. Entrambi questi geni sono attualmente in fase di validazione.

Questi interessanti risultati biologici confermano la validità dei metodi utilizzati e delle scelte adottate durante la tesi.

# Appendice

## A.1. L'analisi fattoriale

L'analisi fattoriale è una tecnica simile a quella dell'analisi delle componenti principali, utilizzata quando si hanno a disposizione moltissime variabili esplicative per la spiegazione di un fenomeno. L'obiettivo dell'analisi fattoriale è quello di descrivere le molte variabili osservate in funzione di pochi fattori non osservabili (latenti), ottenendo una riduzione della complessità del problema e un'interpretazione della struttura di correlazione tra le variabili.

In relazione a quanto trattato in questa tesi, l'analisi fattoriale torna utile come strumento per la stima, nel metodo RUV, dell'inosservata matrice $W$, di dimensione *n*x*k* (*n* numero di array e *k* numero di ignoti fattori non biologici che dovranno essere rimossi). Questa viene stimata utilizzando il solo sottoinsieme di geni che fungono da controlli negativi, ossia non sono associati con il fattore biologico di interesse e, dunque, ogni loro variazione può essere imputabile a fattori non biologici che si vuole eliminare. Svolgere l'analisi su questo insieme di geni serve a prevenire la possibilità di raccogliere variabilità biologica (di interesse) e imputarla erroneamente a variabilità non voluta.

Si dispone, dunque, di una matrice $Y_c$ di valori di espressione avente *n* righe (array) e $G^{'}$ colonne (geni di controllo negativi). Come visto al §3.2.1, la relazione che si presume per questo gruppo di geni è:

$$Y_c = W\alpha_c + \varepsilon_c$$

Ciò che si vuole è ottenere una qualche stima di $W$ che possa essere poi utilizzata nel modello per tutti i geni e per stimare i parametri $\alpha$ e $\beta$.

Assumendo un termine d'errore indipendente e identicamente distribuito, $\varepsilon_j \sim N(0, \sigma_\varepsilon^2 I_m)$, $j \in c$, è stato dimostrato che la matrice $W\alpha_c$ che massimizza la verosimiglianza del modello è data da:

$$argmin\|Y_c - W\alpha_c\|_F^2$$

e la soluzione è $\widehat{W\alpha_c} = U\Lambda_k V^T$, dove $Y_c = U\Lambda V^T$ è la scomposizione a valori singolari di $Y_c$ e $\Lambda_k$ è la matrice diagonale con i *k* più grandi valori singolari come primi *k* termini e zero nel resto della diagonale. Per ricavare la stima di $W$ si può, ad esempio, utilizzare $\widehat{W} = U\Lambda_k$ (Jacob, et al., 2012).

### A.1.1. Scomposizione a valori singolari (SVD)

La scomposizione a valori singolari è una delle più importanti proprietà delle matrici.

Detta $A$ una matrice di dimensioni *m*x*n* a valori reali (vale anche per valori complessi), esistono:

- $U$ matrice *m*x*n* unitaria;
- $\Lambda$ matrice *m*x*n* diagonale con elementi diagonali non negativi e decrescenti $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, $p = \min(m, n)$;
- $V$ matrice *m*x*n* unitaria;

tali che

$$A = U\Lambda V^T$$

Questa fattorizzazione esiste sempre e viene denotata con la sigla *SVD* (Singular Value Decomposition). $U$ e $V$ sono matrici ortogonali e i valori $\lambda_i$ sono chiamati *valori singolari* di $A$. Questi sono pari alla radice degli autovalori di $A^T A$ (oppure di $AA^T$). Le colonne di $U$ (di $V$) rappresentano gli autovettori di $AA^T$ (di $A^T A$). I vettori $u_i$ e $v_i$ sono definiti, rispettivamente, *vettori singolari sinistri* e *destri* di $A$.

In questa tesi, nel metodo RUV (§3.2.1), la matrice da scomporre è $Y_c$, di dimensioni *n*x$G^{'}$, dove, pur essendo $G^{'}$ un numero molto inferiore rispetto a *G*, solitamente, si ha $n < G'$.

## A.2. COMBAT non parametrico

La versione di COMBAT descritta nel §3.1 non è l'unica proposta dai due autori Evan Johnson e Cheng Li. Come si ricorderà il metodo, al primo passo,

prevede la standardizzazione dei valori di espressione, creando così delle nuove variabili, $Z_{ijg}$ per le quali è possibile affermare che $Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2)$, dato l'assunto di distribuzione Normale per il termine d'errore del modello di posizione/scala. Il terzo passo, poi, prevede la stima dei parametri del batch effect utilizzando distribuzioni a priori empiriche. Nello specifico, si assume che $\gamma_{ig} \sim N(\gamma_i, \tau_i^2)$ e $\delta_{ig}^2 \sim Inv - Gamma(\lambda_i, \theta_i)$. Un assunto di questo tipo, però, non è sempre verificato ed è quindi necessario disporre di un metodo alternativo per la stima dei parametri del batch effect nel caso in cui non sia rispettato. Per questo motivo è stata creata la versione non parametrica di COMBAT, la quale permette di ricavare le stime dei parametri del batch effect senza assunzioni sulle distribuzioni a priori, ma semplicemente stimando i valori attesi, a posteriori, dei parametri del batch effect.

Come nel §3.1, si supponga che i valori siano stati standardizzati, che le nuove variabili standardizzate siano $Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2)$, che $\hat{\gamma}_{ig} = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ijg}$ e che $\hat{\delta}_{ig}^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(Z_{ijg} - \hat{\gamma}_{ig}\right)^2$. Si vuole stimare i parametri del batch effect, $\gamma_{ig}$ e $\delta_{ig}^2$, usando i valori attesi a posteriori dei parametri, definiti come $E[\gamma_{ig}]$ e $E[\delta_{ig}^2]$.

Sia $Z_{ig}$ il vettore di tutti i valori $Z_{ijg}$ per un gene $g$ in uno stesso batch $i$, con $j = 1,\ldots,n_i$. Data la distribuzione a posteriori congiunta dei dati $Z_{ig}$ e dei parametri del batch effect $\gamma_{ig}$ e $\delta_{ig}^2$, $\pi(Z_{ig}, \gamma_{ig}, \delta_{ig}^2)$, il valore atteso a posteriori di $\gamma_{ig}$ è:

$$E[\gamma_{ig}] = \int \gamma_{ig}\pi(Z_{ig}, \gamma_{ig}, \delta_{ig}^2)\, d(\gamma_{ig}, \delta_{ig}^2)$$

Sia inoltre $\pi(\gamma_{ig}, \delta_{ig}^2)$ l'ignota funzione di densità a priori per i parametri, $\gamma_{ig}$ e $\delta_{ig}^2$, e sia $L(Z_{ig}|\,\gamma_{ig}, \delta_{ig}^2) = \prod_j \varphi(Z_{ig}, \gamma_{ig}, \delta_{ig}^2)$, dove $\varphi(Z_{ig}, \gamma_{ig}, \delta_{ig}^2)$ è la funzione di densità di una Normale di media $\gamma_{ig}$ e varianza $\delta_{ig}^2$ calcolata in $Z_{ijg}$. Utilizzando il teorema di Bayes l'integrale precedente può essere riscritto come:

$$E[\gamma_{ig}] = \frac{1}{C(Z_{ig})}\int \gamma_{ig}L(Z_{ig}|\,\gamma_{ig}, \delta_{ig}^2)\pi(\gamma_{ig}, \delta_{ig}^2)\, d(\gamma_{ig}, \delta_{ig}^2)$$

dove $C(Z_{ig}) = \int\int \gamma_{ig}L(Z_{ig}|\,\gamma_{ig}, \delta_{ig}^2)\pi(\gamma_{ig}, \delta_{ig}^2)\, d(\gamma_{ig}, \delta_{ig}^2)$.

Per le stime di $C(Z_{ig})$ e dell'integrale presente in $E[\gamma_{ig}]$ si utilizza un metodo di integrazione Monte Carlo. Lo stesso viene fatto per $\delta_{ig}^2$ e si arriva a

definire le stime del batch effect non parametriche. Per un approfondimento maggiore della tecnica si veda il materiale aggiuntivo disponibile in Biostatistics del lavoro originale di Evan Johnson e Cheng Li (2007).

### A.2.1. Differenze nei risultati

Nel §4.3.2, durante la correzione del batch effect con il metodo COMBAT, si è deciso di utilizzare il metodo parametrico malgrado la difficoltà di accettazione degli assunti distributivi delle a piori. Per verificare la correttezza di tale scelta si è provato ad utilizzare anche il metodo di correzione non parametrico descritto nel paragrafo precedente. Ciò che si è ottenuto è una piccola differenza nei valori di espressione che non sembra comportare grossi cambiamenti per le analisi. D'altro canto, però, Il carico computazionale necessario per l'analisi dei 10 campioni in questione con il metodo parametrico è molto basso, mentre per il metodo non parametrico è decisamente molto alto; in particolare il tempo di calcolo dell'aggiustamento con il metodo non parametrico su un notebook con processore dual core Intel® Core™ i5 da 2.30GHz è stato pari a 1086.58 secondi (circa 18 minuti), mentre il metodo parametrico ha impiegato solamente 6.05 secondi. Dato che nelle analisi complete si sono poi utilizzati 39 campioni, se si fosse scelto l'aggiustamento COMBAT non parametrico il tempo di calcolo sarebbe stato molto maggiore.

In Figura A.1 sono riportati i dendrogrammi della cluster analysis compiuta sui valori aggiustati con entrambi i metodi. Quello di sinistra, relativo al metodo parametrico, è lo stesso della Figura 4.11 del §4.3.2. Si nota che la differenza è irrisoria; le distanze alle quali i gruppi vengono uniti sono sostanzialmente le stesse e i gruppi formati sono identici.

In Figura A.2 è riportata la distribuzione delle differenze tra i valori di espressione stimati con il metodo parametrico e quelli stimati con il metodo non parametrico. L'istogramma a destra è un dettaglio di quello a sinistra, fatto per evidenziare la frequenza delle differenze più alte e più basse. Quasi tutte le differenze sono contenute nella fascia tra -0.5 e 0.5; la maggior parte sta tra -0.1 e

0.1 (cfr. Figura A.3). Dato che, com'è noto, i valori di espressione sono riportati in logaritmo in base 2, ciò significa che per la maggior parte dei valori si ha:

$$-0.1 < \log_2 x_P - \log_2 x_{NP} < 0.1$$

cioè che:

$$0.93 < {}^{x_P}/_{x_{NP}} < 1.07$$

ossia che la maggior parte dei valori ottenuti con il metodo parametrico si discosta da quelli ottenuti con il metodo non parametrico di massimo il 7%.

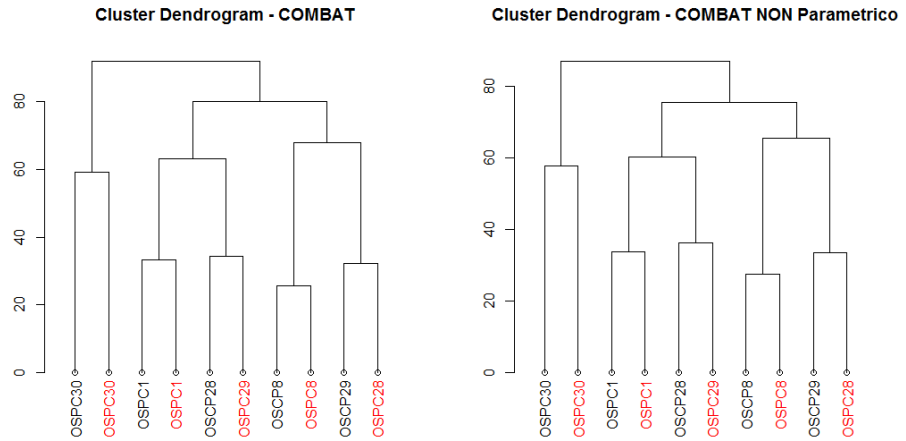



**Figura A.1: Confronto tra i dendrogrammi creati da una cluster analysis con correzione dei valori di espressione con il metodo COMBAT parametrico (a sinistra) e quello non parametrico (a destra).**

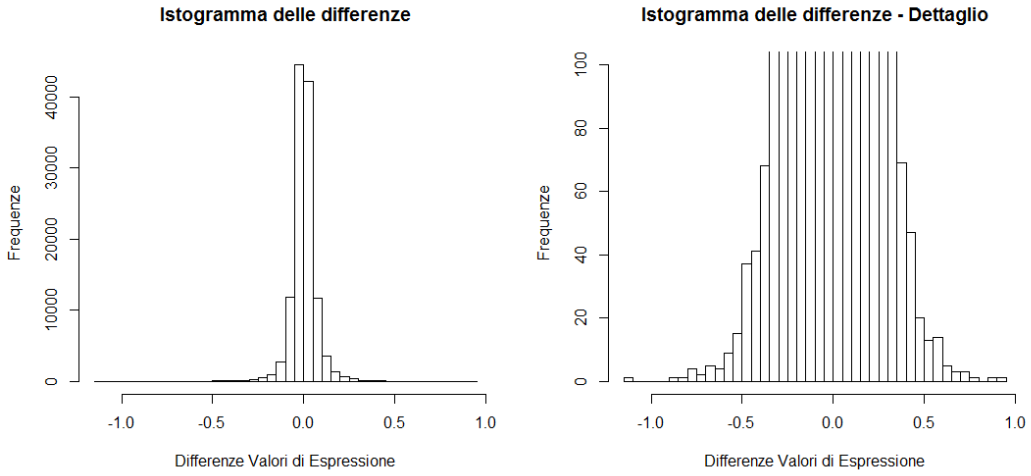



**Figura A.2: A sinistra, istogramma delle differenze tra valori stimati con i metodi COMBAT parametrico e COMBAT non parametrico. A destra è riportato un dettaglio dello stesso istogramma per mettere in evidenza le frequenze più basse.**

La Figura A.3 è relativa alle stesse differenze della Figura A.2, e mostra i boxplot delle differenze stratificate per esperimento.

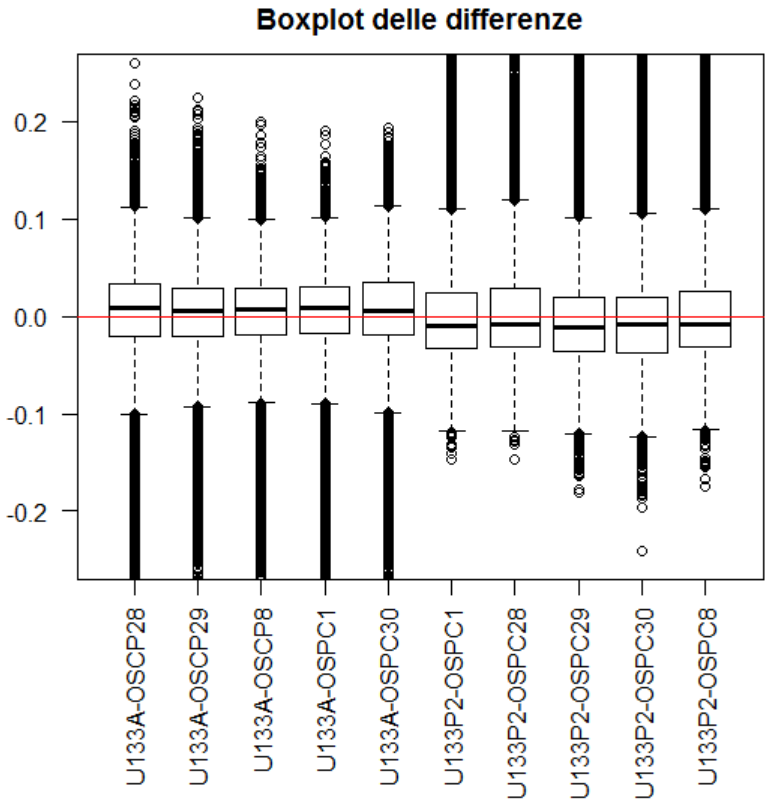



**Figura A.3: Boxplot delle differenze tra valori stimati con i metodi COMBAT parametrico e COMBAT non parametrico, stratificate per piattaforma.**

L'unica osservazione da fare riguardo questo grafico è sul fatto che la mediana delle differenze tra valori stimati con i metodi COMBAT parametrico e COMBAT non parametrico è sempre maggiore di zero per la piattaforma U133A, mentre è sempre minore di zero per la piattaforma U133 plus 2. In altre parole, il metodo non parametrico assegna, medianamente, valori più piccoli di quelli assegnati dal metodo parametrico nella piattaforma U133A e valori più grandi nella piattaforma U133 plus 2. Le differenze, però, sono valori molto piccoli e non sembrano poter comportare differenze nelle analisi.

A fronte di questi risultati si è deciso di utilizzare il metodo parametrico per ottenere le stime COMBAT del batch effect, dato che non comporta grosse

differenze nelle stime ma riduce in maniera drastica il costo computazionale per il loro ottenimento.

## A.3. Normalizzazione VSN

Come riportato nel §4.3.3 durante l'analisi critica della effettiva efficacia dei metodi di rimozione dell'effetto batch analizzati in questo elaborato, si è proceduto a svolgere le cluster analysis indicatrici della bontà dei metodi di aggiustamento anche modificando la tecnica di normalizzazione utilizzata. In particolare si è provata anche la normalizzazione VSN.

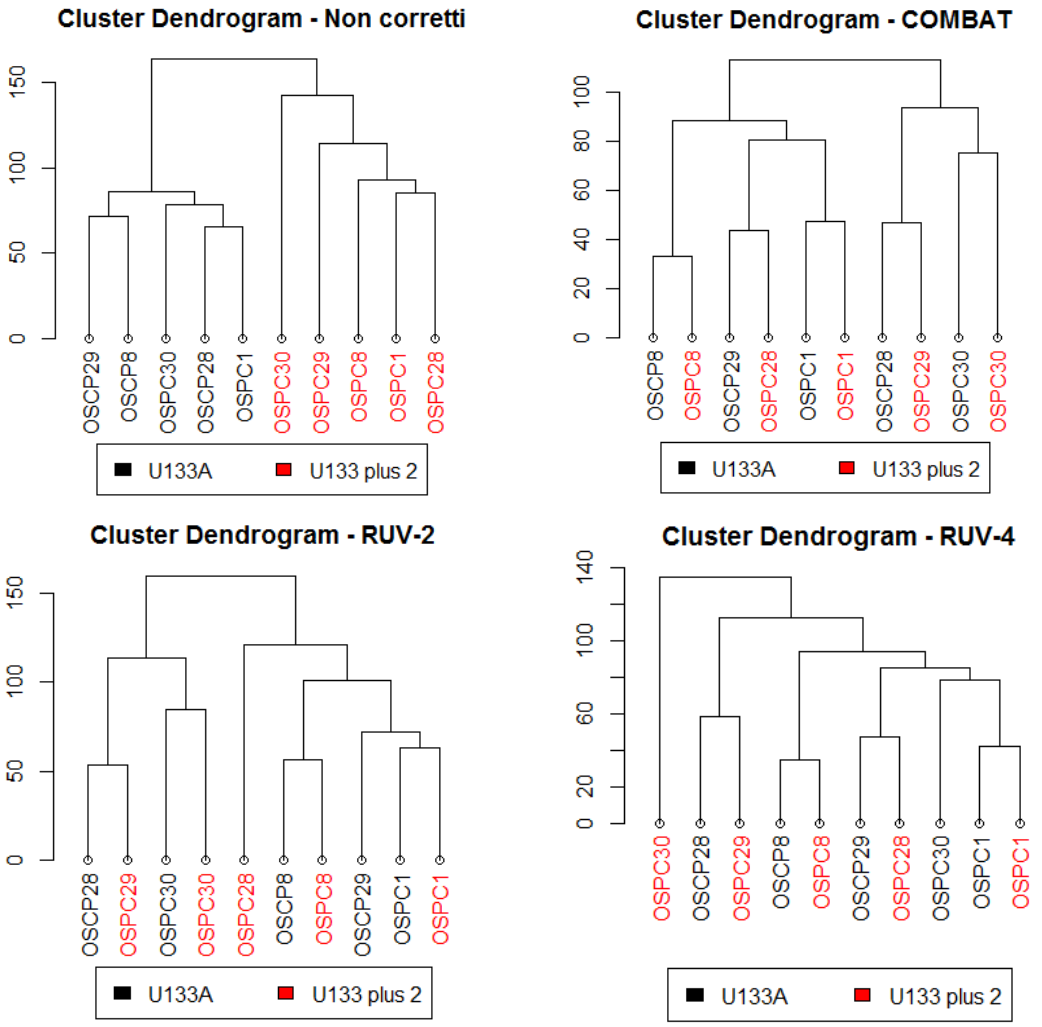
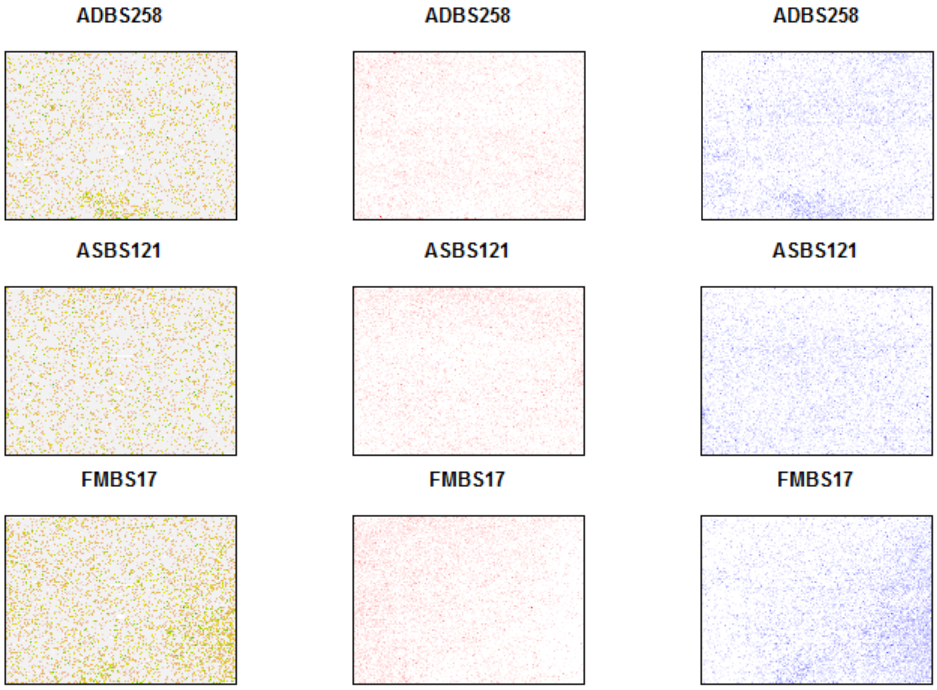


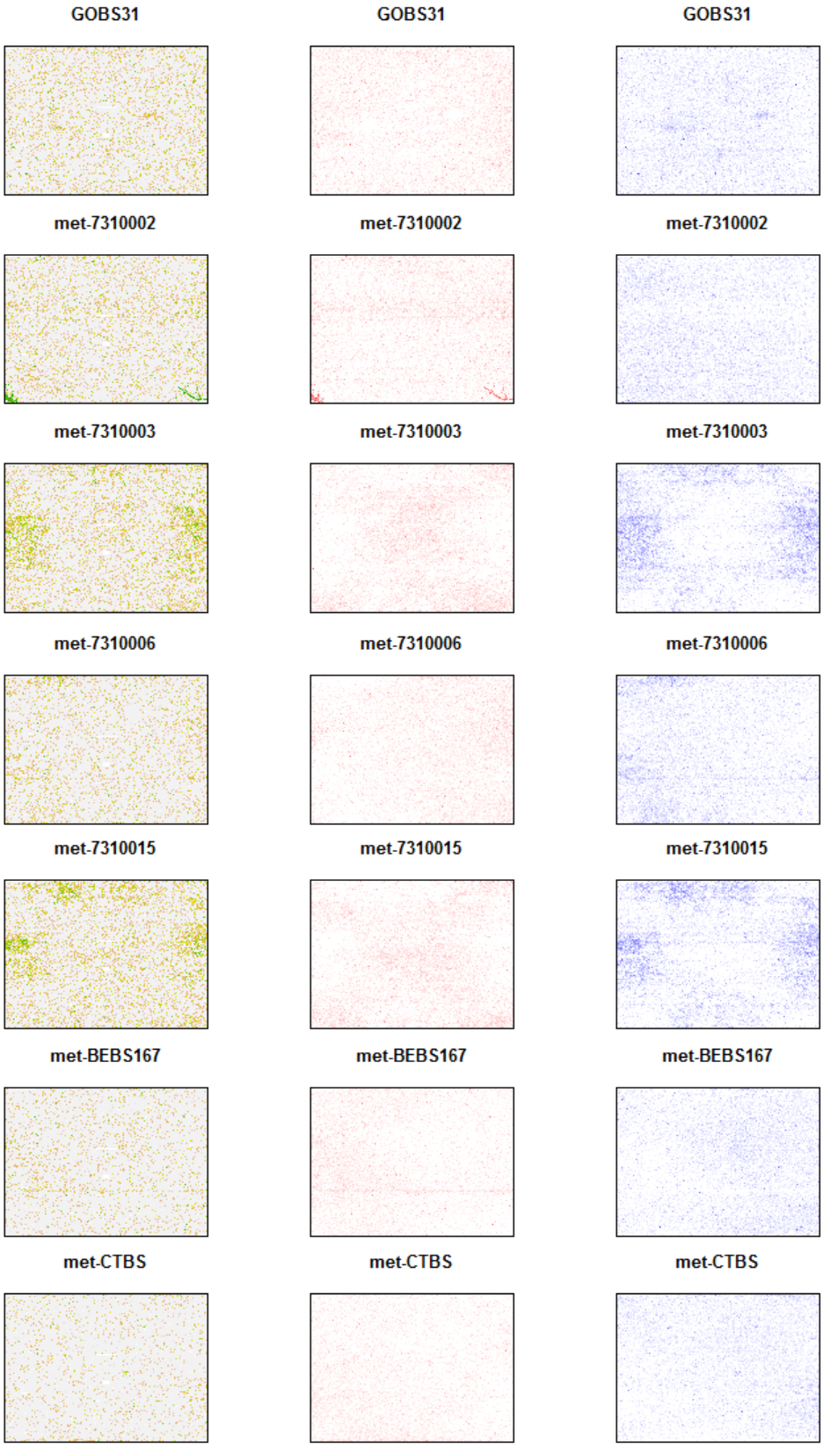

**Figura A.4: Dendrogrammi di cluster analysis eseguite sui dati non corretti (in alto a sinistra), sui dati corretti con metodo COMBAT parametrico (in alto a destra), sui dati corretti con RUV-2 (in basso a sinistra) e sui dati corretti con RUV-4 (in basso a destra).**

In Figura A.4 si riportano i dendrogrammi ottenuti senza alcun aggiustamento e a seguito degli aggiustamenti con i metodi COMBAT, RUV-2 e RUV-4, previa normalizzazione VSN e correzione del background tramite metodo RMA. I risultati sono del tutto analoghi, se non peggiori, a quelli ottenuti con la normalizzazione Quantile adottata dal metodo RMA utilizzato nelle analisi per ricavare i valori di espressione.

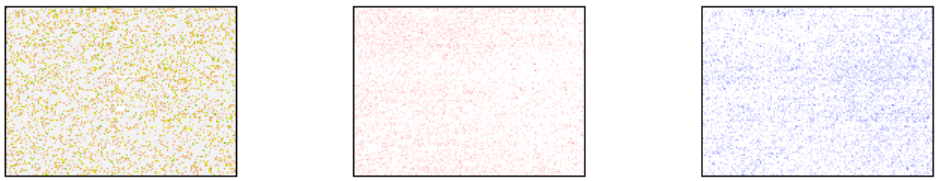
## A.4. Bontà dell'ibridazione

In questo paragrafo sono riportate le immagini relative all'analisi della bontà del processo di ibridazione. Le immagini a sinistra mostrano tutti i valori ricavati per ogni probe come gradazioni tra grigio e marrone. Le immagini centrali rappresentano i residui positivi del modello previsto dal metodo RMA e quelle di destra rappresentano i residui negativi. In Figura A.5 sono riportate le immagini relative alla piattaforma U133A, in Figura A.6 quelle relative alla piattaforma U133 plus 2.

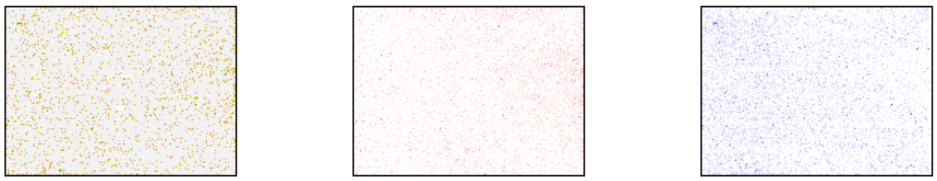



GOBS31
GOBS31
GOBS31
met-7310002
met-7310002
met-7310002
met-7310003
met-7310003
met-7310003
met-7310006
met-7310006
met-7310006
met-7310015
met-7310015
met-7310015
met-BEBS167
met-BEBS167
met-BEBS167
met-CTBS
met-CTBS
met-CTBS

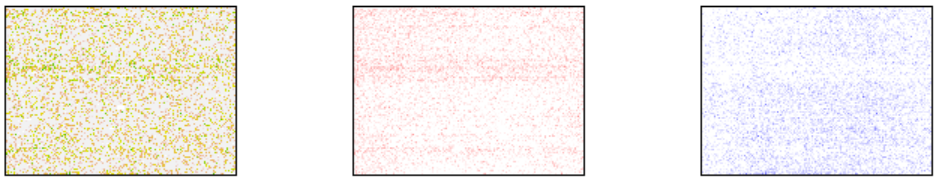
met-FFBS109 met-FFBS109 met-FFBS109



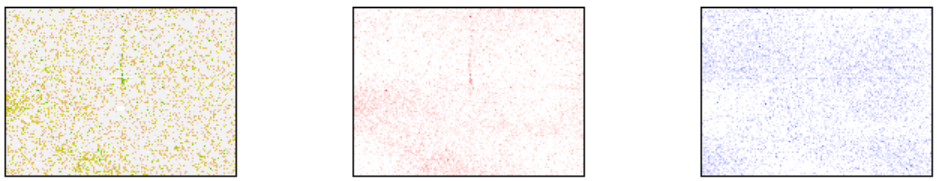
met-MRBS324 met-MRBS324 met-MRBS324



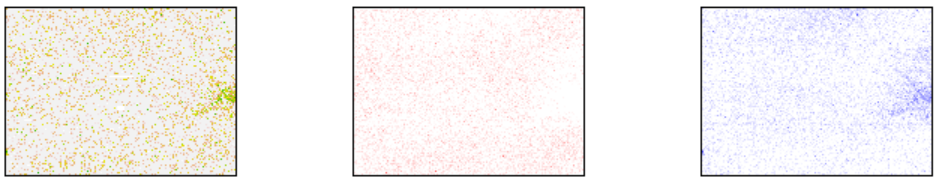
met-TFBS14 met-TFBS14 met-TFBS14



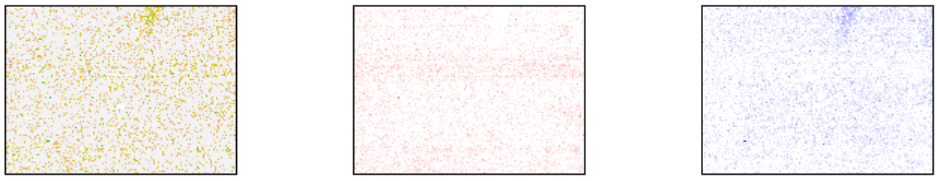
met-VEBS30 met-VEBS30 met-VEBS30



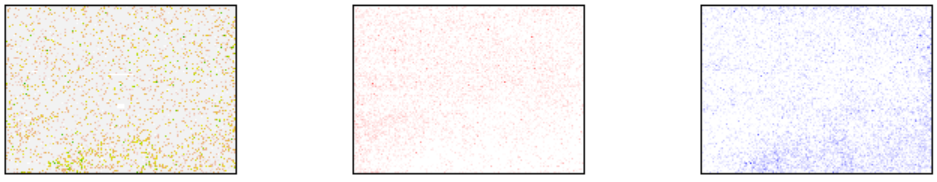
MMR8-6 MMR8-6 MMR8-6



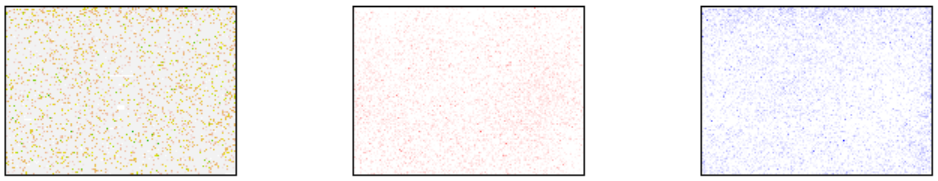
OSCP28 OSCP28 OSCP28



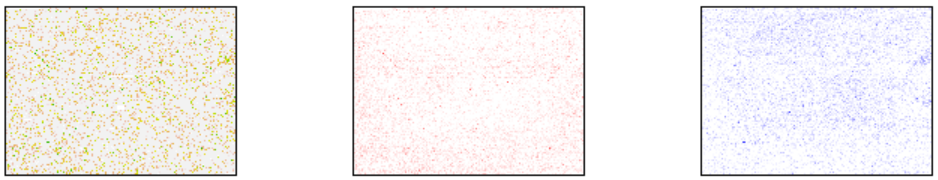
OSCP29 OSCP29 OSCP29



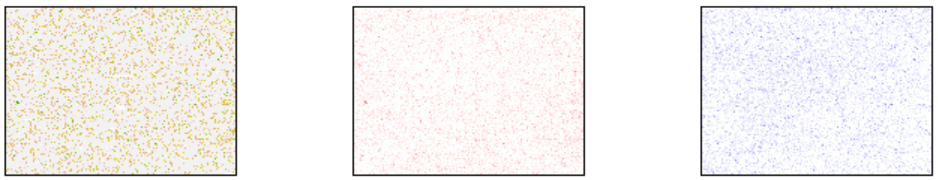
OSCP8 OSCP8 OSCP8



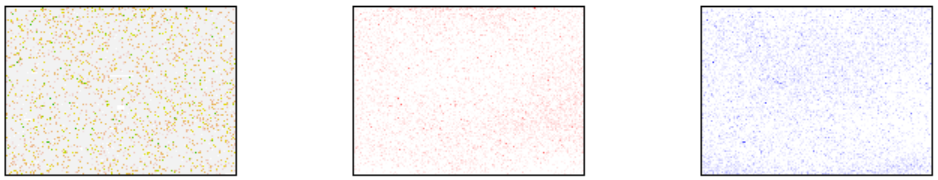
OSPC1 OSPC1 OSPC1



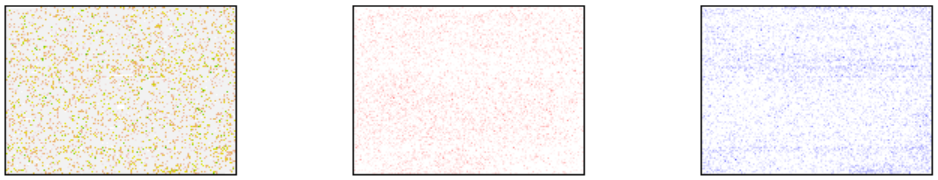
OSPC30 OSPC30 OSPC30



RDBS77 RDBS77 RDBS77



SBCBS92 SBCBS92 SBCBS92



SEBS97 SEBS97 SEBS97

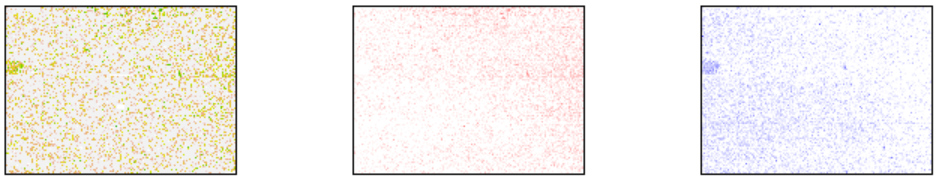
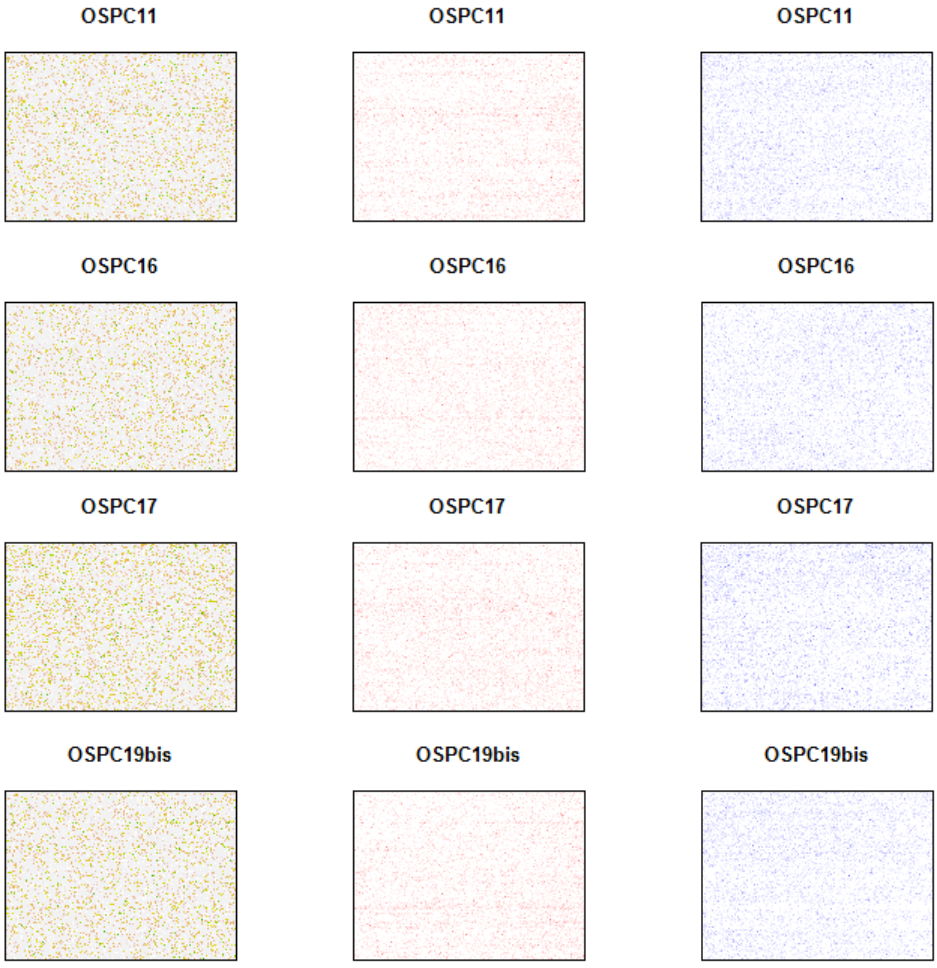
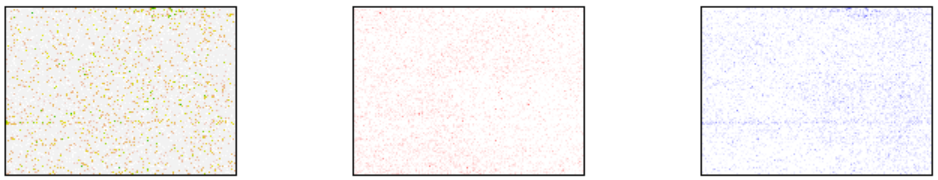


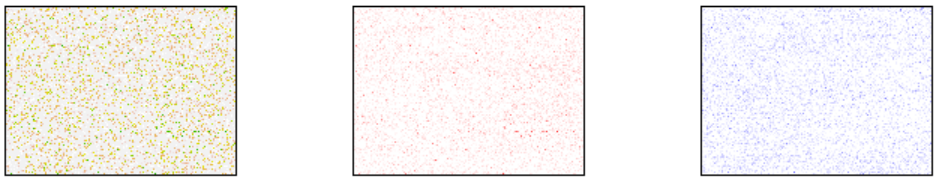

**Figura A.5: Analisi della bontà del processo di ibridazione nella piattaforma U133A.**



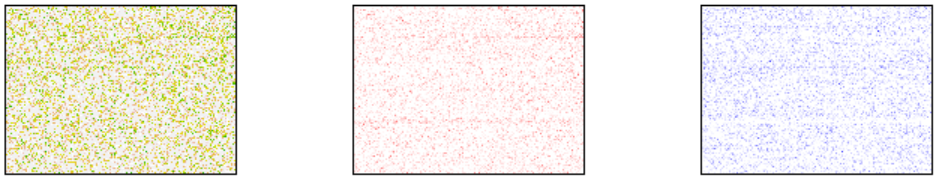
OSPC2 OSPC2 OSPC2



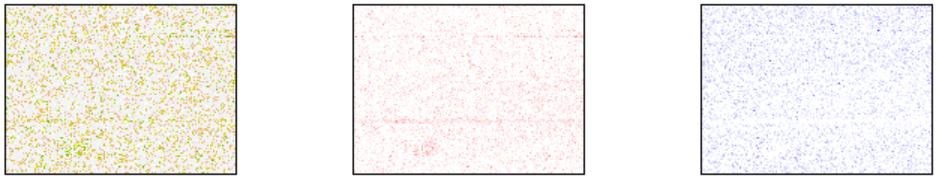
OSPC21 OSPC21 OSPC21



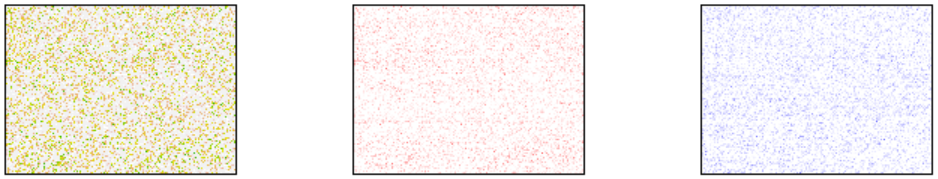
OSPC22 OSPC22 OSPC22



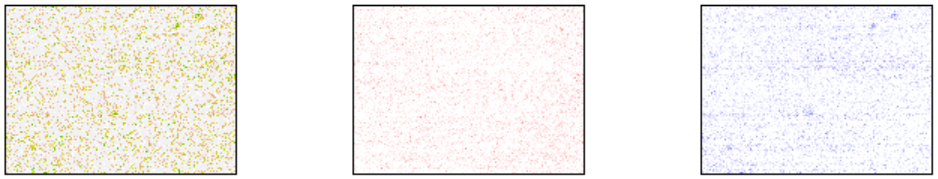
OSPC24 OSPC24 OSPC24



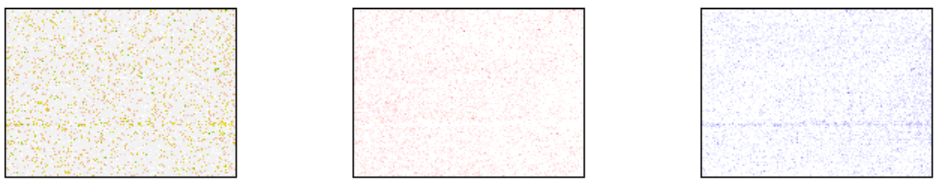
OSPC25 OSPC25 OSPC25



OSPC27 OSPC27 OSPC27



OSPC3 OSPC3 OSPC3



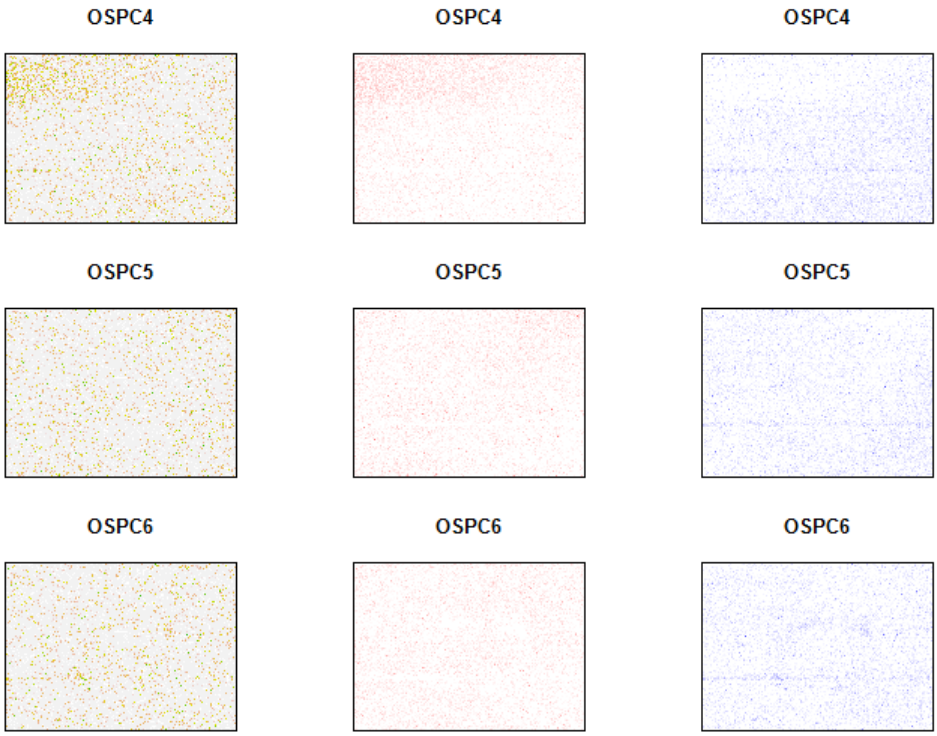



**Figura A.6: Analisi della bontà del processo di ibridazione nella piattaforma U133 plus 2.**

## A.5. Applicazione di RUV ai dati completi

Come spiegato nel §4.4.2 si è voluto provare ad utilizzare come metodi di correzione nell'insieme di tutti i campioni disponibili anche RUV-2 e RUV-4. Questi sono stati valutati, come fatto anche nel sopraccitato paragrafo, con l'utilizzo di una cluster analysis, l'analisi dei grafici RLE, dei boxplot delle varianze dei geni di controllo negativi e della distribuzione dei p-value. Tutti i grafici di queste analisi sono riportati nelle Figure da A.7 ad A.12.

I grafici vanno confrontati con quelli del §4.4.2 relativi ai dati normalizzati e a quelli corretti con COMBAT. E' evidente fin da subito l'inadeguatezza di RUV-2, mentre per RUV-4 i risultati sono migliori. Con l'utilizzo di questi metodi di valutazione delle correzioni non è possibile affermare che i risultati ottenuti con l'applicazione di COMBAT siano in assoluto i migliori, data la bontà dei segnali di adeguatezza anche del metodo RUV-4. Malgrado ciò, date le evidenti migliori prestazioni ottenute da COMBAT nelle analisi dei dati in doppia piattaforma,

questo sarà il metodo utilizzato in prima battuta nelle analisi e i risultati ottenuti con esso saranno considerati i più attendibili.

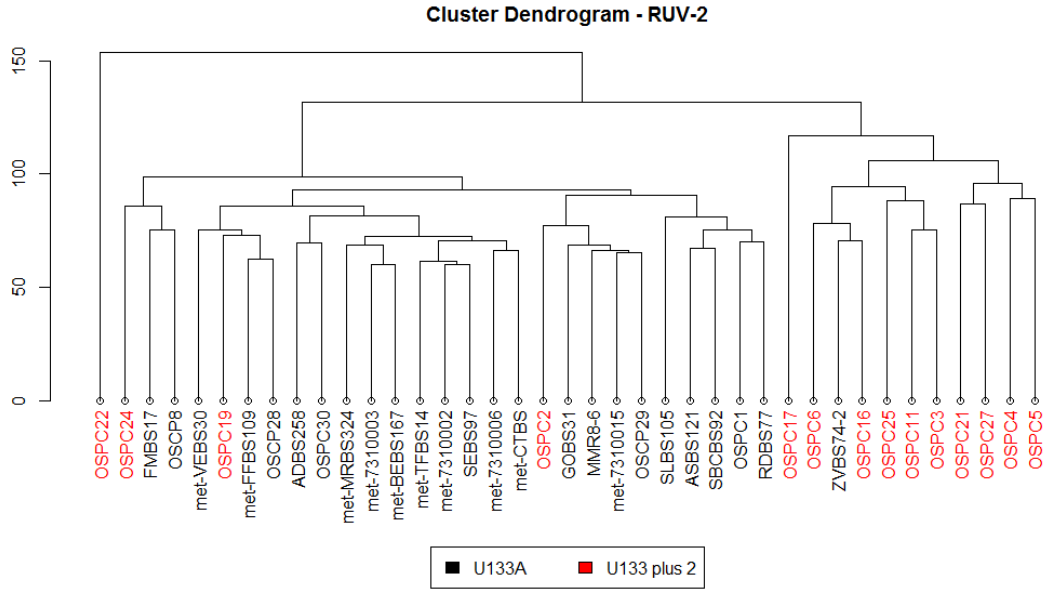



**Figura A.7: Dendrogramma della cluster analysis eseguita su tutti i dati (39 esperimenti) corretti con RUV-2.**

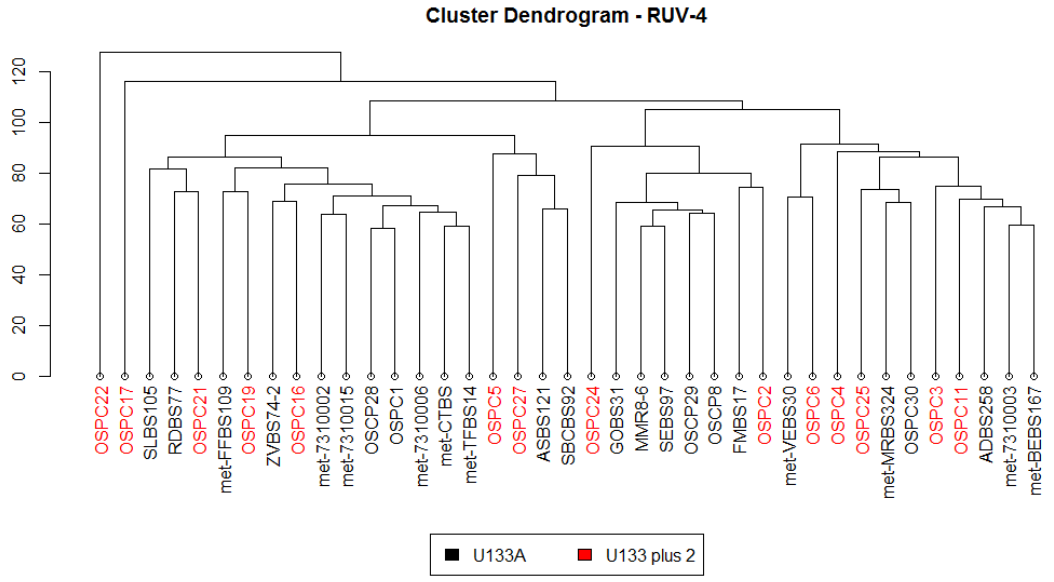



**Figura A.8: Dendrogramma della cluster analysis eseguita su tutti i dati (39 esperimenti) corretti con RUV-4.**

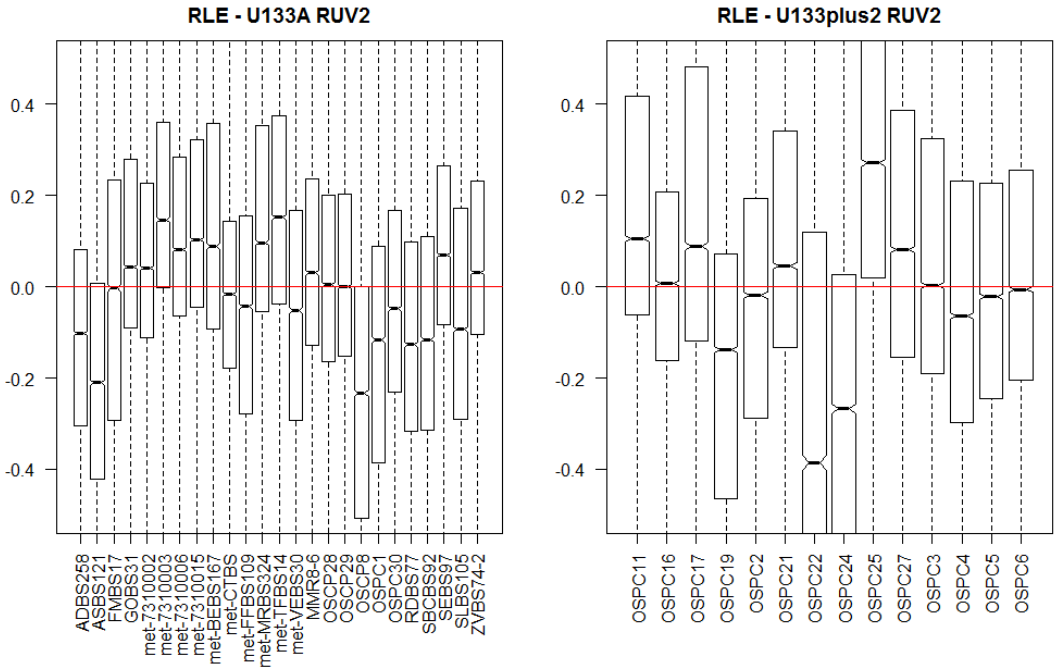



**Figura A.9: Grafici RLE per tutti i 39 campioni corretti con RUV-2. A sinistra quelli per la piattaforma U133A e a destra quelli per la piattaforma U133 plus 2.**

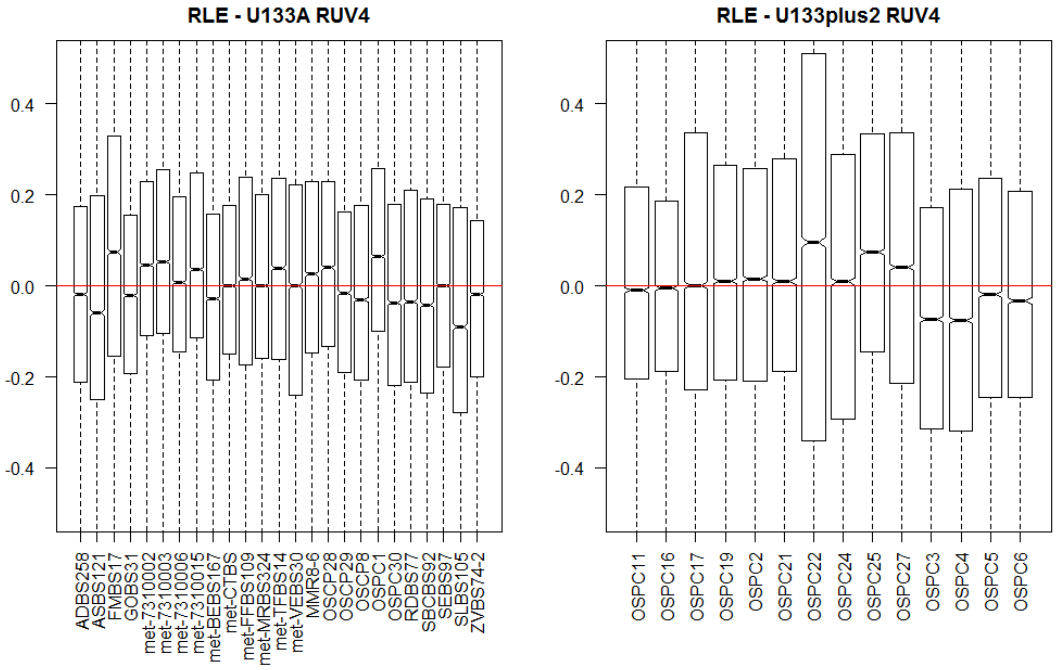



**Figura A.10: Grafici RLE per tutti i 39 campioni corretti con RUV-4. A sinistra quelli per la piattaforma U133A e a destra quelli per la piattaforma U133 plus 2.**

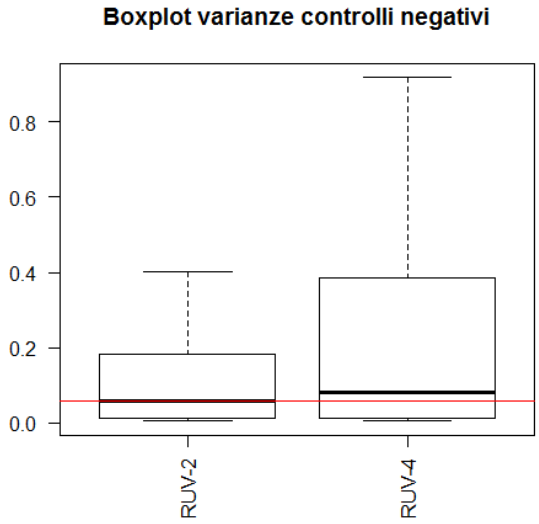



**Figura A.11: Boxplot delle varianze dei geni di controllo negativi per dati corretti conRUV-2 e RUV-4, rispettivamente.**

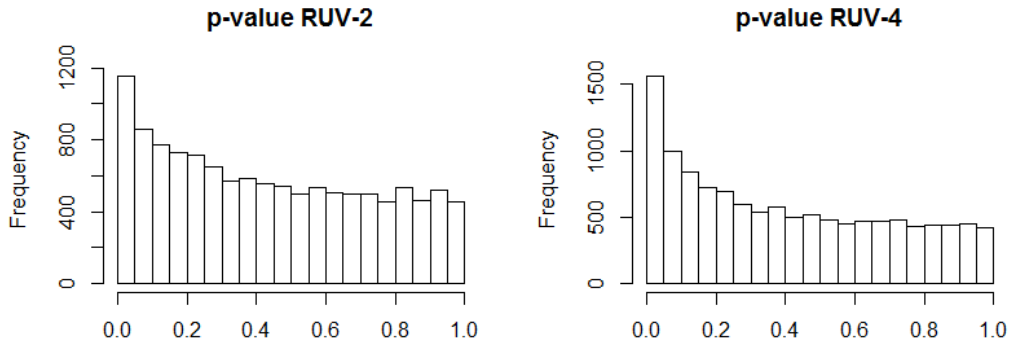



**Figura A.12: Istogrammi dei valori dei p-value ottenuti facendo un test Ebayes su ogni gene.**

## A.6. Statistiche test: COMBAT vs RUV-4

In questo paragrafo si vogliono confrontare i valori delle due statistiche test SAM ed Ebayes per ogni gene per le due correzioni del batch effect COMBAT e RUV-4. Lo scopo è di verificare se ci sono grosse differenze nell'identificazione dei geni differenzialmente espressi a seconda del metodo di correzione utilizzato. Nel caso in cui i metodi conducessero a risultati simili si avrebbe una maggiore fiducia nella correttezza degli stessi.

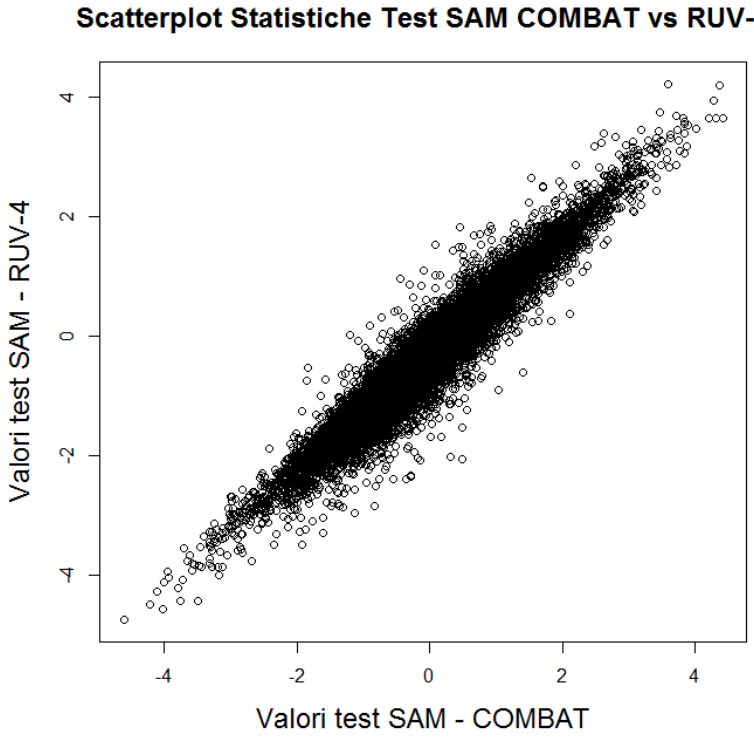



**Figura A.13: Scatterplot dei valori della statistica test SAM nei dati corretti con COMBAT e con RUV-4 per ogni gene.**

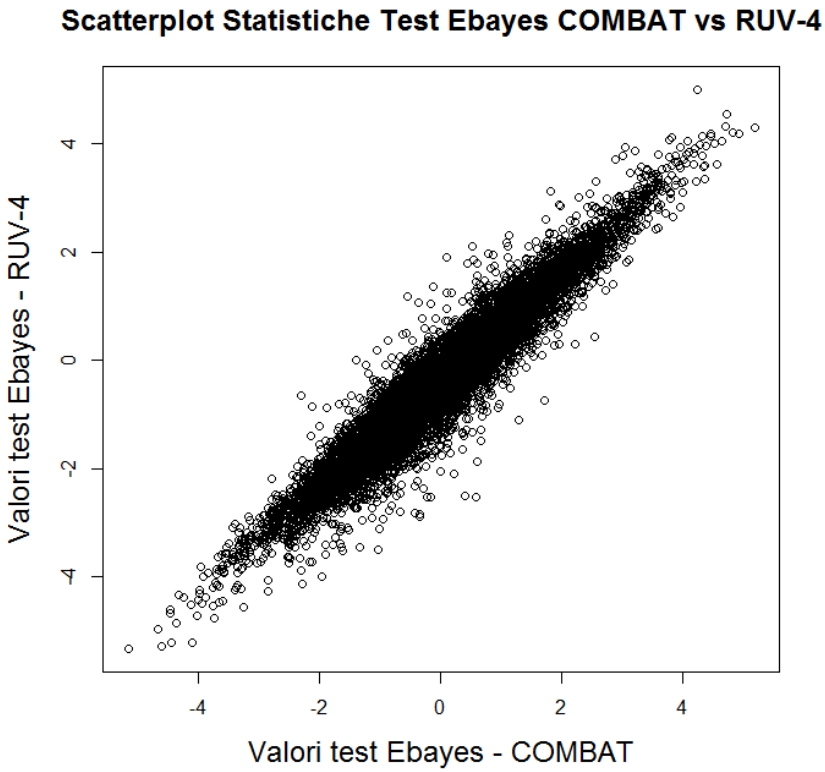



**Figura A.14: Scatterplot dei valori della statistica test Ebayes nei dati corretti con COMBAT e con RUV-4 per ogni gene.**

Nelle Figure A.13 e A.14 sono riportati gli scatterplot dei valori delle statistiche test SAM ed Ebayes nei dati corretti con COMBAT e con RUV-4 per ogni gene. E' evidente la forte relazione lineare tra i valori osservati. La correlazione lineare è, infatti, pari a 0.961 per i valori del test SAM e a 0.959 per il test Ebayes.

## A.7. Lista dei DEG

In questo paragrafo è riportata la lista dei geni differenzialmente espressi identificati tramite la statistica test Ebayes sui dati corretti con il metodo COMBAT. I geni sono riconoscibili tramite i simboli ufficiali e maggiori informazioni su essi possono essere trovate ricercando il simbolo nel database dell'NCBI (National Center for Biotechnology Information).

| Simbolo | logFC | adj.P.Val | B | RUV-4 |
|---|---|---|---|---|
| C6orf62 | 0,653 | 0,0443 | 3,609 | ✓ |
| NUAK1 | -1,179 | 0,0443 | 3,529 | ✓ |
| BTN3A3 | 0,855 | 0,0491 | 2,917 | ✓ |
| MRS2 | 0,788 | 0,0491 | 2,630 | ✓ |
| CXCL11 | 1,974 | 0,0491 | 2,333 | ✓ |
| DEPTOR | 1,486 | 0,0491 | 2,268 | ✓ |
| AEBP1 | -1,322 | 0,0491 | 2,150 | ✓ |
| TDP2 | 0,669 | 0,0491 | 2,144 | ✓ |
| COL16A1 | -0,761 | 0,0491 | 2,018 | ✓ |
| C1orf109 | 0,515 | 0,0491 | 1,879 | ✓ |
| UROD | 0,534 | 0,0491 | 1,816 | ✓ |
| PPCS | 0,601 | 0,0491 | 1,652 | ✓ |
| VCAN | -1,730 | 0,0491 | 1,642 | ✓ |
| PDGFRB | -0,824 | 0,0491 | 1,614 | ✓ |
| AHSA1 | 0,634 | 0,0491 | 1,613 | ✓ |
| PCOLCE | -1,126 | 0,0491 | 1,604 | ✓ |
| FSTL3 | -0,528 | 0,0491 | 1,575 | ✓ |
| KIAA1324 | 0,921 | 0,0495 | 1,429 | ✓ |
| HSD17B1 | 0,614 | 0,0495 | 1,380 | ✓ |
| PRKCDBP | -0,696 | 0,0495 | 1,344 | ✓ |
| UBE2K | 0,555 | 0,0495 | 1,338 | ✗ |
| ANO1 | 0,991 | 0,0495 | 1,277 | ✓ |
| FXYD5 | -1,397 | 0,0495 | 1,270 | ✓ |

| | | | | |
|---|---|---|---|---|
| **TMEM134** | 0,670 | 0,0495 | 1,266 | ✓ |
| **C8orf33** | 0,876 | 0,0499 | 1,224 | ✓ |
| **ALDH5A1** | 0,706 | 0,0545 | 1,113 | ✓ |
| **NDUFC2** | 0,483 | 0,0549 | 1,050 | ✓ |
| **LOXL2** | -0,609 | 0,0549 | 1,043 | ✓ |
| **FDPS** | 0,541 | 0,0569 | 0,980 | ✗ |
| **TBCC** | 0,616 | 0,0575 | 0,942 | ✓ |
| **TRIM27** | 0,647 | 0,0577 | 0,891 | ✓ |
| **TRIT1** | 0,531 | 0,0577 | 0,882 | ✓ |
| **EMP1** | -1,188 | 0,0684 | 0,686 | ✓ |
| **HSPG2** | -0,519 | 0,0684 | 0,677 | ✓ |
| **WDR77** | 0,507 | 0,0684 | 0,656 | ✓ |
| **GMPR** | 1,339 | 0,0718 | 0,589 | ✓ |
| **FBXW2** | 0,414 | 0,0721 | 0,554 | ✓ |
| **MAGOH** | 0,585 | 0,0721 | 0,538 | ✓ |
| **EBNA1BP2** | 0,617 | 0,0745 | 0,460 | ✗ |
| **MRPS11** | 0,489 | 0,0745 | 0,454 | ✓ |
| **KRT19P2** | -0,328 | 0,0745 | 0,442 | ✓ |
| **PALLD** | -0,911 | 0,0745 | 0,393 | ✓ |
| **CYHR1** | 0,521 | 0,0745 | 0,382 | ✗ |
| **PPL** | -0,876 | 0,0745 | 0,356 | ✓ |
| **ZNF75D** | 0,416 | 0,0745 | 0,320 | ✗ |
| **TUBB6** | -1,015 | 0,0745 | 0,313 | ✓ |
| **ZZZ3** | 0,547 | 0,0745 | 0,288 | ✓ |
| **COL11A1** | -2,544 | 0,0745 | 0,278 | ✓ |
| **HIST1H2BI** | 0,640 | 0,0745 | 0,271 | ✓ |
| **BTN3A2** | 0,593 | 0,0745 | 0,269 | ✗ |
| **TRMT12** | 0,667 | 0,0745 | 0,254 | ✓ |
| **DCHS1** | -0,450 | 0,0759 | 0,221 | ✓ |
| **THBS2** | -1,588 | 0,0781 | 0,169 | ✓ |
| **TMEM59** | 0,497 | 0,0781 | 0,164 | ✓ |
| **POLE3** | 0,721 | 0,0811 | 0,091 | ✓ |
| **ITGA3** | -0,778 | 0,0811 | 0,069 | ✓ |
| **EXOSC4** | 0,714 | 0,0811 | 0,069 | ✓ |
| **GNL2** | 0,584 | 0,0816 | 0,048 | ✗ |
| **TMEM70** | 0,364 | 0,0864 | -0,017 | ✗ |
| **GLT8D2** | -0,990 | 0,0872 | -0,039 | ✓ |
| **HIST1H2BK** | 1,104 | 0,0882 | -0,080 | ✓ |
| **C11orf67** | 0,647 | 0,0882 | -0,096 | ✓ |
| **MRPL13** | 0,899 | 0,0882 | -0,101 | ✗ |
| **AGL** | 0,702 | 0,0882 | -0,110 | ✗ |
| **HTATSF1** | 0,360 | 0,0882 | -0,117 | ✗ |
| **SCP2** | 0,493 | 0,0886 | -0,133 | ✗ |
| **MTL5** | 0,629 | 0,0940 | -0,205 | ✓ |
| **TAPBP** | 0,450 | 0,0940 | -0,211 | ✓ |
| **KRT17** | -1,570 | 0,0943 | -0,237 | ✓ |
| **CYTH3** | -0,288 | 0,0943 | -0,242 | ✓ |

| | | | | |
|---|---|---|---|---|
| **PPIE** | 0,442 | 0,0943 | -0,251 | ✗ |
| **SNX24** | -0,438 | 0,0949 | -0,268 | ✓ |
| **ZMPSTE24** | 0,580 | 0,0950 | -0,289 | ✓ |
| **PYCRL** | 0,589 | 0,0950 | -0,292 | ✗ |
| **PDPN** | -0,688 | 0,0969 | -0,321 | ✓ |
| **ACADM** | 0,487 | 0,0988 | -0,349 | ✗ |
| **ZNF16** | 0,525 | 0,0997 | -0,381 | ✗ |
| **MAFF** | -0,877 | 0,0997 | -0,397 | ✓ |
| **FSTL1** | -0,837 | 0,0997 | -0,399 | ✓ |
| **EFNB2** | -0,917 | 0,0997 | -0,401 | ✓ |

**Tabella A.2: Lista dei geni differenzialmente espressi identificati con Ebayes nei dati corretti con COMBAT. La seconda colonna è il logaritmo del fold chance nelle due condizioni di alta e bassa sopravvivenza; le terza colonna è il p-value aggiustato con il metodo FDR; la quarta contiene i logaritmi degli odds ratio che il gene sia differenzialmente espresso (si veda la formula 2.31 del paragrafo 2.5.3). L'ultima colonna è un flag che mostra quali dei geni sono stati identificati con Ebayes anche nei dati corretti con RUV-4.**

Tutti i geni riportati nella Tabella A.1 sono stati identificati come significativi anche con il test SAM. Si noti che l'ultima colonna della tabella è un flag per evidenziare se il gene è stato trovato anche nella rispettiva lista di differenzialmente espressi sui dati corretti con il metodo RUV-4.

## A.8. Liste dei pathway

In questo paragrafo sono riportate le liste dei pathway risultati alterati tra le condizioni di sopravvivenza alta e bassa delle pazienti. I metodi utilizzati sono l'analisi di arricchimento, SPIA, Global Test e CliPPER.

***Analisi di arricchimento:***

| Pathway | qvalue |
|---|---|
| **ECM-receptor interaction** | 5,37E-07 |

**Tabella A.3: Pathway alterato rilevato tramite l'analisi di arricchimento e relativo q-value.**

***SPIA:***

| Pathway | pNDE | pPERT | pG | pGFdr | Status |
|---|---|---|---|---|---|
| **ECM-receptor interaction** | 2,25E-05 | 0,000005 | 2,69E-09 | 2,39E-07 | Inhibited |

**Tabella A.4: Pathway alterato rilevato tramite Signaling Pathway Impact Analysis (SPIA). Sono riportati anche i p-value pNDE, relativo ad una analisi di arricchimento, pPERT, relativo alla perturbazione del pathway, e pG, ossia il p-value globale. La quinta colonna corrisponde al p-value globale corretto tramite FDR; la sesta è lo stato di inibizione o attivazione del pathway.**

***Global Test:***

| Pathway | qvalue |
|---|---|
| Axon guidance | 0,036 |
| beta-Alanine metabolism | 0,036 |
| ECM-receptor interaction | 0,036 |
| Small cell lung cancer | 0,036 |
| Toll-like receptor signaling pathway | 0,036 |
| Butanoate metabolism | 0,039 |
| Carbohydrate digestion and absorption | 0,039 |
| Epithelial cell signaling in Helicobacter pylori infection | 0,039 |
| Fatty acid metabolism | 0,039 |
| Glycolysis / Gluconeogenesis | 0,039 |
| NF-kappa B signaling pathway | 0,039 |
| Propanoate metabolism | 0,039 |
| Terpenoid backbone biosynthesis | 0,039 |
| TGF-beta signaling pathway | 0,039 |
| Toxoplasmosis | 0,048 |
| Gap junction | 0,056 |
| Fructose and mannose metabolism | 0,058 |
| HIF-1 signaling pathway | 0,058 |
| Influenza A | 0,058 |
| Legionellosis | 0,058 |
| Nicotinate and nicotinamide metabolism | 0,058 |
| Prostate cancer | 0,058 |
| Selenocompound metabolism | 0,058 |
| Steroid biosynthesis | 0,058 |
| Type II diabetes mellitus | 0,058 |
| Bacterial invasion of epithelial cells | 0,062 |
| Folate biosynthesis | 0,064 |
| Phenylalanine metabolism | 0,064 |
| Complement and coagulation cascades | 0,065 |
| Glutathione metabolism | 0,065 |
| Glycosylphosphatidylinositol(GPI)-anchor biosynthesis | 0,065 |
| Hepatitis C | 0,065 |
| Jak-STAT signaling pathway | 0,065 |
| Leukocyte transendothelial migration | 0,065 |
| Pancreatic cancer | 0,065 |
| Primary bile acid biosynthesis | 0,065 |
| Shigellosis | 0,065 |
| Tyrosine metabolism | 0,065 |
| Wnt signaling pathway | 0,065 |
| VEGF signaling pathway | 0,072 |
| Melanoma | 0,075 |
| Pantothenate and CoA biosynthesis | 0,075 |
| Vascular smooth muscle contraction | 0,075 |
| Graft-versus-host disease | 0,078 |
| Arrhythmogenic right ventricular cardiomyopathy (ARVC) | 0,080 |
| Apoptosis | 0,090 |
| p53 signaling pathway | 0,092 |

| | |
|---|---|
| **Pyrimidine metabolism** | 0,094 |
| **Non-small cell lung cancer** | 0,096 |
| **Tryptophan metabolism** | 0,101 |

**Tabella A.5: Pathway rilevati tramite Global Test e relativi q-value.**

***CliPPER:***

| Pathway | alphaMean | alphaVar |
|---|---|---|
| **Wnt signaling pathway** | 0,00 | 0,08 |
| **Terpenoid backbone biosynthesis** | 0,00 | 0,09 |
| **Legionellosis** | 0,02 | 0,12 |
| **Tryptophan metabolism** | 0,02 | 0,13 |
| **Basal cell carcinoma** | 0,00 | 0,22 |
| **African trypanosomiasis** | 0,00 | 0,27 |
| **Leukocyte transendothelial migration** | 0,00 | 0,28 |
| **Alanine, aspartate and glutamate metabolism** | 0,01 | 0,30 |
| **Cardiac muscle contraction** | 0,30 | 0,02 |
| **Dilated cardiomyopathy** | 0,00 | 0,32 |
| **NF-kappa B signaling pathway** | 0,00 | 0,32 |
| **Phototransduction** | 0,04 | 0,32 |
| **Porphyrin and chlorophyll metabolism** | 0,02 | 0,34 |
| **N-Glycan biosynthesis** | 0,03 | 0,35 |
| **Valine, leucine and isoleucine degradation** | 0,01 | 0,37 |
| **Folate biosynthesis** | 0,00 | 0,39 |
| **Glyoxylate and dicarboxylate metabolism** | 0,36 | 0,04 |
| **Natural killer cell mediated cytotoxicity** | 0,01 | 0,41 |
| **Apoptosis** | 0,00 | 0,45 |
| **Axon guidance** | 0,01 | 0,50 |
| **Pathogenic Escherichia coli infection** | 0,01 | 0,51 |
| **Dopaminergic synapse** | 0,00 | 0,55 |
| **Tight junction** | 0,03 | 0,52 |
| **TGF-beta signaling pathway** | 0,00 | 0,57 |
| **Melanogenesis** | 0,02 | 0,55 |
| **T cell receptor signaling pathway** | 0,01 | 0,56 |
| **Amino sugar and nucleotide sugar metabolism** | 0,01 | 0,58 |
| **Fructose and mannose metabolism** | 0,01 | 0,59 |
| **NOD-like receptor signaling pathway** | 0,01 | 0,60 |
| **Gap junction** | 0,01 | 0,61 |
| **Pantothenate and CoA biosynthesis** | 0,01 | 0,62 |
| **Nicotinate and nicotinamide metabolism** | 0,00 | 0,64 |
| **p53 signaling pathway** | 0,00 | 0,68 |
| **Prion diseases** | 0,02 | 0,66 |
| **Toll-like receptor signaling pathway** | 0,00 | 0,70 |
| **Complement and coagulation cascades** | 0,03 | 0,67 |
| **Antigen processing and presentation** | 0,00 | 0,71 |
| **Small cell lung cancer** | 0,02 | 0,69 |
| **Type II diabetes mellitus** | 0,04 | 0,67 |
| **beta-Alanine metabolism** | 0,00 | 0,72 |
| **Chagas disease (American trypanosomiasis)** | 0,05 | 0,67 |
| **Glioma** | 0,01 | 0,74 |
| **Fatty acid elongation** | 0,01 | 0,75 |

| | | |
|---|---|---|
| **Glycosylphosphatidylinositol(GPI)-anchor biosynthesis** | 0,02 | 0,77 |
| **Amyotrophic lateral sclerosis (ALS)** | 0,02 | 0,79 |
| **VEGF signaling pathway** | 0,01 | 0,85 |
| **Pentose phosphate pathway** | 0,04 | 0,83 |
| **Epithelial cell signaling in Helicobacter pylori infection** | 0,01 | 0,87 |
| **Hepatitis B** | 0,05 | 0,84 |
| **Measles** | 0,04 | 0,86 |
| **Fc epsilon RI signaling pathway** | 0,05 | 0,86 |
| **Toxoplasmosis** | 0,03 | 0,89 |
| **Carbohydrate digestion and absorption** | 0,00 | 0,93 |
| **Osteoclast differentiation** | 0,01 | 0,94 |
| **Prostate cancer** | 0,00 | 0,96 |
| **Pancreatic cancer** | 0,01 | 0,96 |
| **PPAR signaling pathway** | 0,02 | 0,95 |
| **Cell adhesion molecules (CAMs)** | 0,02 | 0,96 |
| **Renal cell carcinoma** | 0,02 | 0,96 |
| **Bacterial invasion of epithelial cells** | 0,00 | 1,00 |
| **Jak-STAT signaling pathway** | 0,00 | 1,00 |
| **Melanoma** | 0,00 | 1,00 |
| **Shigellosis** | 0,00 | 1,00 |
| **Synaptic vesicle cycle** | 0,02 | 0,98 |
| **Glycolysis / Gluconeogenesis** | 0,01 | 1,00 |
| **Herpes simplex infection** | 0,04 | 0,98 |
| **Non-small cell lung cancer** | 0,05 | 0,98 |

**Tabella A.6: Pathway alterati rilevati tramite il metodo CliPPER. La seconda colonna è il p-value relativo alla differenza media di espressione tra i geni del pathway nelle due condizioni; la terza è relativa alla differenza nelle correlazioni tra i geni del pathway nelle due condizioni.**

# Bibliografia

Alvero, A. B., 2010. Recent insights into the role of NF-κB in ovarian carcinogenesis. *Genome Medicine,* 2(56).

Benjamini, Y. & Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B,* 57(1), pp. 289-300.

Chen, C., Grennan, K. & Badner, J., 2011. Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods.. *PLoS ONE,* 6(2).

Cleveland, W. S. & Devlin, S. J., 1998. Locally-weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.,* Volume 83, pp. 596-610.

Dudoit, S., Yang, Y. H., Callow, M. J. & Speed, T. P., 2002. Statistical Methods For Identifying Differentially Expressed Genes In Replicated cDNA Microarray Experiments. *Statistica Sinca,* Volume 12, pp. 111-139.

Emery, L. A. et al., 2009. Early dysregulation of cell adhesion and extracellular matrix pathways in breast cancer progression. *The American journal of pathology,* 175(3), pp. 1292-1302.

Fare, T. L. et al., 2003. Effects of Atmospheric Ozone on Microarray Data Quality. *Analytical Chemistry,* 75(17).

Friedman, J., Hastie, T. & Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software,* 33(1), p. 1:22.

Gagnon-Bartsch, J. A. & Speed, T. P., 2012. Using control genes to correct for unwanted variation in microarray data. *Biostatistics,* 13(3).

Goeman, J. J., van de Geer, S. A., de Kort, F. & van Houwelingen, H. C., 2004. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics,* 20(1), pp. 93-99.

Hartemink, A. J., Gifford, D. K. & Jaakkola, T. S., 2003. Maximum likelihood estimation of optimal scaling factors for expression array normalization. *Bioinformatics,* 19(2), pp. 185-193.

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

Irizarry, R. A. et al., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics,* 4(2), pp. 249-264.

Jacob, L., Gagnon-Bartsch, J. & Speed, T. P., 2012. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *arXiv preprint.*

Johnson, W. E. & Li, C., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics,* 8(1), pp. 118-127.

Li, C. & Wong, W. H., 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *PNAS,* 98(1), pp. 31-36.

Martini, P. et al., 2013. Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic acids research,* 41(1), pp. e19-e19.

Massa, M. S., Chiogna, M. & Romualdi, C., 2010. Gene set analysis exploiting the topology of a pathway. *BMC systems biology,* 4(1), p. 121.

Oberaigner, W. et al., 2012. Survival for Ovarian Cancer in Europe: The across-country variation did not shrink in the past decade. *Acta Oncologica,* Volume 51, pp. 441-453.

Przybycin, C. G. & Soslow, R. A., 2011. Typing of ovarian carcinomas: an update. *Diagnostic Histopathology,* 17(4), pp. 165-177.

Sales, G., Calura, E., Martini, P. & Romualdi, C., 2013. Graphite Web: web tool for gene set analysis exploiting pathway topology. *Nucleic acids research,* 41(W1), pp. W89-W97.

Storey, J. & Tibshirani, R., 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences,* 100(16), pp. 9440-9445.

Tarca, A. L. et al., 2009. A novel signaling pathway impact analysis. *Bioinformatics,* 25(1), pp. 75-82.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 267-288.

Wiley, J., 2006. *Encyclopedia of Statistical Sciences.* s.l.:s.n.

Yeung, T.-L., Leung, C. S., Wong, K. K. & Samimi, G., 2013. TGF-β modulates ovarian cancer invasion by upregulating CAF-derived versican in the tumor microenvironment. *Cancer research,* 73(16), pp. 5016-5028.