UNIVERSITÀ DEGLI STUDI DI PISA

Facoltà di Medicina e Chirurgia

Corso di dottorato in
Esplorazione molecolare, metabolica e funzionale
del sistema nervoso e degli organi di senso

Tesi di Dottorato

Dai dati grezzi all'interpretazione biologica:

progettazione e analisi degli esperimenti di espressione genica

realizzati mediante microarray

Supervisore:

Dott.ssa Silvia Pellegrini

Candidata:

Ing. Erika Melissari

Anno Accademico: 2008/2009

"... Non vogliate negar l'esperienza
di retro al sol, del mondo sanza gente.
Considerate la vostra semenza
fatti non foste a viver come bruti
ma per seguir virtute e canoscenza"

Dante Alighieri, Divina Commedia, Inferno canto XXVI, 116-120

Indice

Abstract

Negli ultimi venti anni la genetica e la biologia molecolare hanno contribuito significativamente al progresso scientifico-medico, fornendo strumenti per isolare, clonare e studiare molti dei geni che compongono il genoma umano. E' ora possibile analizzare contemporaneamente l'espressione di migliaia di geni, ossia valutare quello che viene chiamato profilo genico, grazie all'uso di speciali supporti tecnologicamente avanzati denominati *microarray*. Un singolo esperimento di espressione genica realizzato con microarray produce migliaia di dati, per i quali è necessario un approccio rigoroso di tipo matematico e bioinformatico, sia nelle fasi di acquisizione e analisi che in quelle di interpretazione e archiviazione.

A differenza delle fasi di preparazione dei campioni e ibridizzazione dei vetrini, che ormai sono regolate da protocolli sufficientemente standardizzati, i passaggi che portano dall'estrazione dei dati all'interpretazione biologica dei risultati non possono essere riassunti in un protocollo unico.

Questo progetto di dottorato ha avuto lo scopo di studiare i metodi di progettazione di un esperimento di espressione genica mediante microarray e gli strumenti bioinformatici che servono a realizzare le fasi di estrazione e pre-trattamento dei dati, l'analisi statistica e l'interpretazione dei risultati. Tali metodi sono stati applicati a quattro esperimenti realizzati nel laboratorio presso il quale è stata svolta questa tesi.

Sono stati individuati, fra quelli disponibili, i metodi bioinformatici per l'estrazione, il pre-trattamento e l'analisi statistica dei dati più affidabili e versatili per l'eliminazione degli errori legati alla metodica e per l'acquisizione di un dato statisticamente robusto. Il confronto critico dei metodi analizzati ha messo in luce la necessità di mettere a punto una soluzione ottimale di analisi per ciascun esperimento.

La valutazione degli strumenti utili per l'interpretazione biologica dei risultati ha messo, invece, in evidenza profonde limitazioni legate essenzialmente all'assenza di informazioni ordinatamente catalogate e alla incompleta modellazione dei processi di co-regolazione genica nelle banche dati.

Introduzione

Nel corso degli ultimi anni, la rapida evoluzione delle metodiche e degli strumenti a disposizione della biologia molecolare, ha fatto sì che il sequenziamento del DNA divenisse una tecnica sempre più efficiente e raffinata. La prima sequenza genomica ad essere stata pubblicata, nel 1995, è stata quella di *Haemophilus influenzae*, un piccolo batterio gram-negativo con un genoma di circa 1,8 milioni di basi. Successivamente, nel 1996, è stato completato il sequenziamento del primo genoma eucariotico, quello del lievito *Saccharomyces cerevisiae*, che comprende circa 13 milioni di basi organizzate in sedici cromosomi.

Nel 2001 è stato raggiunto l'obiettivo primario del Progetto Genoma Umano, vale a dire la pubblicazione della prima bozza del genoma, completata in maniera definitiva nel 2003 [1]. Questo evento ha dato un grosso impulso alla bioinformatica e alla moltiplicazione delle informazioni biologiche accessibili in modo più o meno libero sulla rete informatica.

Due aspetti rendono peculiari e complesse le informazioni relative alle sequenze di genomi. Il primo aspetto è che la quantità e la varietà dei dati ottenuti da queste ricerche non hanno precedenti nella storia della biologia e probabilmente della scienza in generale. Il secondo aspetto, non meno importante del primo, è che si tratta di problemi nuovi, mai affrontati prima d'ora, che richiedono lo sviluppo di nuovi strumenti di analisi.

La bioinformatica trova dunque nell'analisi di dati genomici un'area di indagine veramente innovativa e stimolante. Determinare la sequenza di un genoma, infatti, non significa comprendere automaticamente il programma genetico che essa racchiude. Anche con i più sofisticati sistemi attualmente disponibili si riescono ad interpretare solo parzialmente ed approssimativamente gli elementi funzionali contenuti in un genoma e, ancor meno, si riesce a comprendere il significato dell'informazione genomica nella sua globalità.

Il problema principale consiste, quindi, nell'identificare le sequenze di DNA che sono trascritte in RNA messaggero (mRNA) per essere poi tradotte in proteine. L'analisi del trascrittoma, cioè dell'insieme degli RNA trascritti, consente di mettere a fuoco la questione indagando direttamente il livello di espressione di vari trascritti in cellule diverse e in condizioni fisiologiche e patologiche diverse.

Anche nei più semplici procarioti molti geni si accendono o si spengono o modificano la frequenza di trascrizione rispetto alla loro espressione di base in risposta a particolari situazioni. Il profilo trascrizionale riflette quindi lo stato funzionale di una cellula; di conseguenza, capire in quali circostanze un gene si è espresso è spesso un presupposto essenziale per comprenderne la funzione.

La regolazione dell'espressione genica assume un'ulteriore dimensione negli organismi multicellulari dove tipi diversi di cellule sono caratterizzati da profili trascrizionali diversi. Lo studio sistematico del livello di espressione dei trascritti è quindi di grande importanza per almeno due distinte ragioni: in primo luogo per il fatto che il genoma di qualsiasi cellula esprime in ogni determinato momento solo una parte dei suoi geni; in secondo luogo perché non esistono ancora dei validi metodi predittivi che, in base alla sequenza genomica, siano in grado di dare indicazioni sulle condizioni in cui un gene viene espresso.

E' quindi importante essere consapevoli dell'esistenza di un gene, ma è altrettanto importante capire il contesto in cui esso viene espresso.

Gli acidi nucleici offrono un metodo di indagine diretta basato sulla specificità di ibridizzazione di due eliche complementari, che possono fungere da sonde per l'identificazione e la quantificazione di specifici mRNA.

L'ibridizzazione è stata per decenni utilizzata in biologia molecolare come principio base di metodiche quali il Southern blotting e il Northern blotting. I microarray a DNA rappresentano l'applicazione più avanzata di queste tecnologie di ibridazione, essendo in grado di ospitare molte migliaia di sonde diverse, corrispondenti ad altrettanti geni, su un unico supporto di vetro grande quanto un vetrino da microscopio ottico.

La complementarità di due filamenti di DNA riflette la regola secondo la quale l'adenina si lega alla timina e la citosina si lega alla guanina. Uno o entrambi i filamenti di DNA ibridizzati possono essere sostituiti con RNA che, pur differendo per la presenza dell'uracile al posto della timina, va incontro ugualmente al fenomeno dell'ibridizzazione (Fig. I1).

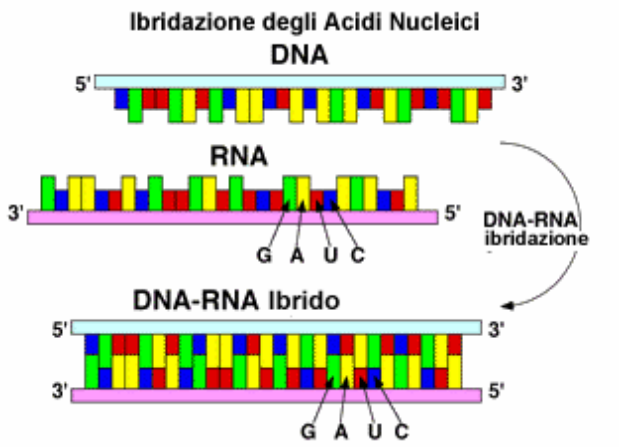


Figura I1: *Ibridizzazione degli acidi nucleici*

Mentre tecniche quali il Southern blotting e il Northern blotting forniscono risposte solo semi-quantitative sull'espressione genica, gli esperimenti che utilizzano i microarray sono totalmente quantitativi, cioè

riescono a dare una stima dell'espressione differenziale tanto più precisa quanto più tale espressione è differente nelle due condizioni che si stanno comparando. Il blocco sperimentale di base di un esperimento di espressione genica è, infatti, costituito da due campioni di mRNA, di cui uno è assunto come controllo, ossia la sua espressione genica è relativa a condizioni "normali", mentre l'altro rappresenta la condizione di espressione alternativa da studiare. I due campioni possono essere ibridizzati sullo stesso vetrino (two-color protocol) oppure ciascuno su un vetrino diverso (one-color protocol); in ogni caso l'informazione che verrà ricavata sarà l'espressione relativa di un campione rispetto all'altro. Tale rapporto fra i livelli di espressione nelle due condizioni viene denominato "fold-change". L'esperimento di espressione genica realizzato mediante microarray è, quindi, quantitativo e comparativo.

I microarray possono essere paragonati a microprocessori biologici poiché abilitano l'analisi parallela di profili di espressione genica. Inoltre, come i microprocessori elettronici, che vengono appositamente costruiti per realizzare funzioni generiche o più specifiche dell'apparecchiatura sulla quale verranno montati, esistono microarray capaci di analizzare simultaneamente l'intero trascrittoma di un organismo oppure dedicati all'indagine di porzioni più piccole, per esempio i geni trascritti in uno specifico organo, o disegnati per osservare l'espressione di specifiche reti di geni, i cosiddetti "pathway".
La scelta opportuna del vetrino sul quale si andrà successivamente a valutare l'espressione genica è il primo di sei passaggi attraverso i quali si articola un tipico esperimento microarray "two-color".

Queste fasi sono:

1. scelta del tipo di vetrino da utilizzare per l'esperimento; il microarray può essere acquistato o direttamente costruito se è disponibile l'apparato di stampa dei supporti;

2. progettazione dell'esperimento di espressione genica: definizione degli obiettivi e scelta del disegno sperimentale;

3. preparazione dei campioni da ibridizzare: estrazione dell'RNA dai campioni, verifica di integrità e purezza dell'RNA, retrotrascrizione in cDNA e marcatura con due diverse molecole fluorescenti;

4. ibridizzazione dei campioni fluorescenti sul microarray; le sequenze marcate si ibridizzano con le loro complementari sul chip e generano un segnale di fluorescenza d'intensità proporzionale al numero di copie trascritte del gene;

5. estrazione dei dati grezzi: lettura dei valori di fluorescenza, effettuata con uno speciale scanner a due canali che genera due immagini indipendenti relative ai due fluorocromi usati;

6. pre-trattamento, analisi statistica e interpretazione dei dati.

E' possibile cercare di schematizzare le fasi più prettamente bioinformatiche con un diagramma di flusso seppure, come si vedrà in seguito

in questa tesi, non esiste una vera e propria standardizzazione di esse. Un esempio estensivo di tale diagramma è riportato in figura I2.

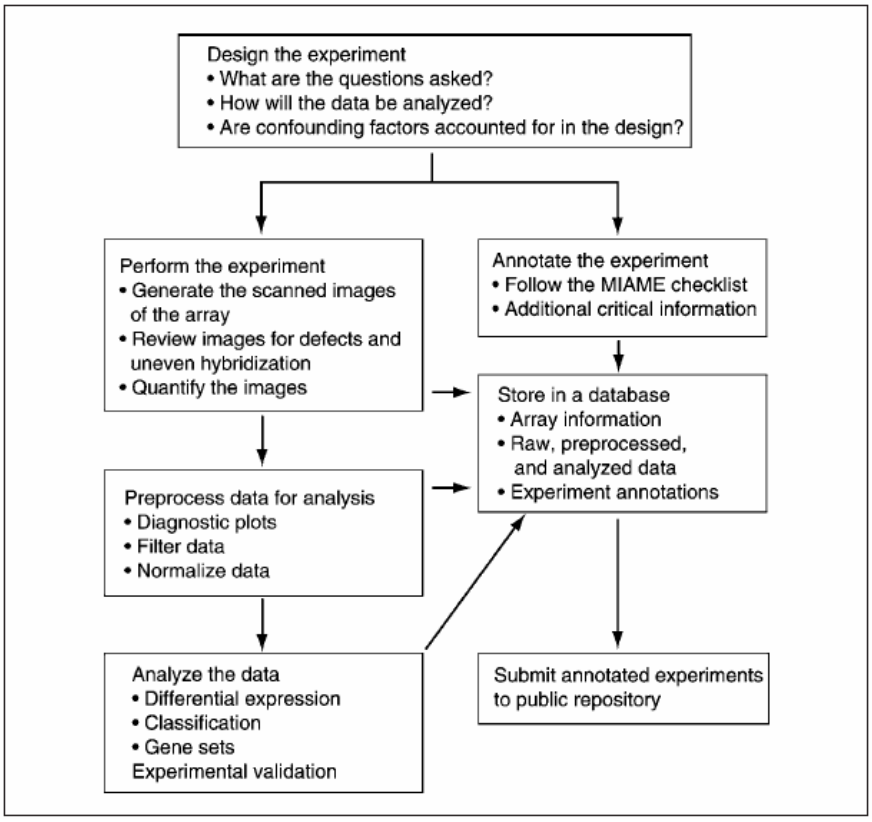


Figura I2: *Diagramma di flusso operativo di un esperimento di microarray*

La rapidità con la quale viene realizzato un esperimento microarray (dall'estrazione dei campioni di RNA all'ibridizzazione sui vetrini), rispetto all'esecuzione sequenziale con altre metodiche (analisi tradizionale con Southern o Northern blotting) di tanti esperimenti di espressione genica quanti sono i geni presenti su un vetrino, rende minimo il tempo che consente di ottenere i dati grezzi, cioè le informazioni che verranno successivamente analizzate per rispondere ad un quesito biologico iniziale.

Rispetto all'uso delle metodiche di analisi dell'espressione genica alternative ai microarray e non così parallelizzate, si è quindi invertito il diagramma che descrive lo svolgimento temporale di un progetto di analisi dell'espressione genica (Fig. I3). Infatti, mentre nell'analisi tradizionale la maggior parte del tempo era dedicato alla messa a punto e alla realizzazione dell'esperimento in laboratorio e una minima parte all'analisi dei pochi dati che potevano essere ricavati, adesso con l'uso dei microarray la maggior parte del tempo è dedicata alla progettazione del disegno dell'esperimento e alla successiva analisi dell'enorme quantità di dati ricavata.

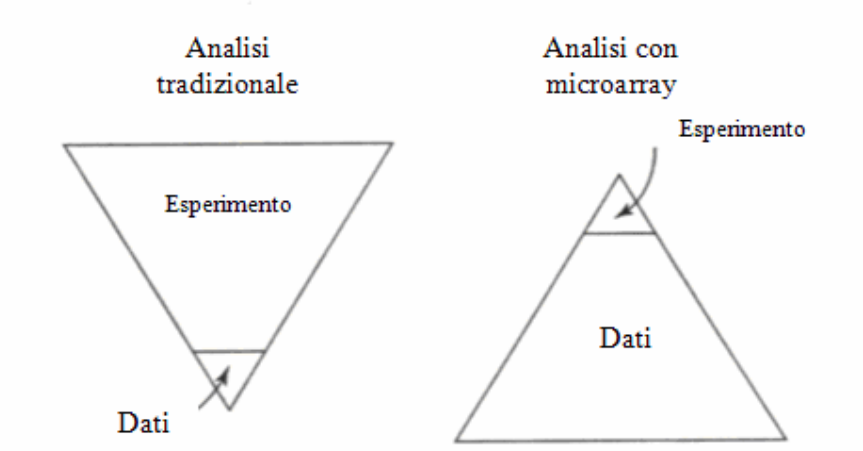


Figura I3: Diagramma temporale di un esperimento di microarray

I metodi per la progettazione efficace ed efficiente degli esperimenti di espressione genica realizzati mediante microarray sono stati illustrati in molte pubblicazioni dal 2000 ad oggi [2-14].

Infatti, malgrado l'apparente semplicità nella formulazione di un'ipotesi biologica da verificare attraverso un esperimento di espressione genica con microarray, la complessità delle problematiche coinvolte nella sua realizzazione pratica ha fornito la spinta allo sviluppo di una sezione dedicata della biostatistica ed alimenta la ricerca continua di soluzioni appositamente studiate.

In particolare, la strutturazione un po' naif dei primi esperimenti di espressione è stata ormai sostituita da una vera e propria fase di progettazione che richiede a più livelli competenze biologiche e informatiche e che è fondamentale per la corretta realizzazione dell'esperimento.

Inoltre, la teoria alla base della quantizzazione dei segnali acquisiti, del calcolo del rapporto delle loro intensità, della normalizzazione del dato e dell'estrazione del risultato, è estremamente sofisticata e pone grossi problemi, soprattutto dal punto di vista dell'analisi statistica di dati generati con esperimenti simultanei su migliaia di geni.

Per far fronte alla quantità enorme di dati prodotti in un esperimento di espressione genica, la bioinformatica produce, ormai quasi quotidianamente, soluzioni sempre più sofisticate per l'estrazione dell'informazione e l'interpretazione dei dati ottenuti [15-19].

In questa tesi verrà fornita un'ampia panoramica delle tecniche di disegno degli esperimenti di espressione genica che utilizzano i microarray, con particolare attenzione ai metodi di progettazione degli esperimenti di confronto dell'espressione globale fra classi. Verranno, inoltre, illustrate le metodologie di estrazione e visualizzazione dei dati, sottrazione del rumore e normalizzazione dei dati. Successivamente, saranno esposte tre tecniche di analisi statistica per la determinazione della lista dei geni differenzialmente espressi fra i gruppi di campioni a confronto, denominate *SAM* (Significance Analysis of Microarrays) [20], *LIMMA* (LInear Model of MicroArray data) [21] e *MAANOVA* (MicroArray ANalysis Of VAriance) [22]. Verranno, infine, presentati gli strumenti bioinformatici che consentono di analizzare la lista dei geni differenzialmente

espressi ricavata, al fine di individuare se in essa possono essere messi in luce percorsi di co-regolazione dell'espressione genica che spiegano il fenomeno che si sta studiando attraverso i microarray.

Per tutti i metodi illustrati verrà descritta l'applicazione pratica in alcuni esperimenti di espressione genica eseguiti nel Laboratorio Microarray del Dipartimento di Patologia Sperimentale, Biotecnologie Mediche, Infettivologia ed Epidemiologia dell'Università di Pisa, durante il periodo di svolgimento di questa tesi.

Capitolo 1

Metodi per la progettazione di esperimenti di “class comparison”

Per controllare le molteplici fonti di variabilità che si abbattono sui dati di espressione genicamicroarray è necessaria un’accurata pianificazione dell’esperimento per il raggiungimento degli obiettivi che lo studio si propone. La definizione preventiva e particolareggiata di questi obiettivi si ripercuote sulla progettazione e sull’analisi dell’esperimento.

Gli esperimenti di espressione genica possono essere catalogati in tre categorie:

- “class comparison”: lo scopo di questi studi è indagare se c’è differenza di espressione fra due o più classi di soggetti in condizioni sperimentali differenti e determinare quali sono i geni responsabili di questa differenza.
- “class discovery”: in questo caso i soggetti non vengono preventivamente catalogati in base al fenotipo, ma lo scopo dell’analisi è riuscire a rivelare attraverso i profili di espressione genica, se esistono raggruppamenti spontanei fra campioni e se essi hanno un significato biologico o correlano con altri dati disponibili sui soggetti analizzati.
- “class prediction”: lo scopo di questi esperimenti è lo sviluppo di profili di espressione genica, comunemente detti “signature”, formati da un numero limitato di geni che servono come classificatori di soggetti con fenotipo ignoto o di geni con funzione sconosciuta in classi di soggetti con fenotipo noto o in gruppi di geni con funzione assegnata.

Gli schemi di confronto fra campioni che possono essere adottati devono considerare la tipologia di esperimento che si vuole realizzare.

Esistono, tuttavia, alcuni principi generali che restano validi per tutti i tipi di esperimenti: sufficiente replicazione dell’informazione, randomizzazione e bilanciamento per contenere i fattori di disturbo.

Nel seguito si illustreranno le tecniche di disegno sperimentale relative agli esperimenti di “class comparison”, poiché in questa categoria sono classificati gli esperimenti condotti durante lo svolgimento di in questa tesi.

1.1. Necessità di replicare le osservazioni

La variabilità intrinseca degli esperimenti di espressione genica realizzati con microarray impone la replicazione delle osservazioni a più livelli.

La replicazione delle osservazioni, infatti, è l'unico metodo che consente di contenere, purtroppo senza eliminare, la componente *random* del rumore dei dati.

InoltreOltre che dal rumore, il livello di espressione rilevato in un esperimento è influenzato, oltre che dal rumore, dalla variabilità biologica tipica di ciascun campione. Per variabilità biologica si intende l'insieme delle peculiari differenze di espressione genica che ciascun organismo, dal più complesso al meno complesso, può mettere in atto nella risposta allo stesso stimolo.

La variabilità biologica si somma al rumore *random*; tuttavia, anche se la metodica microarray riuscisse a produrre un segnale ideale, generato senza alcuna componente di rumore, la replicazione biologica sarebbe in ogni caso necessaria. Infatti, con una sola osservazione, derivante dal confronto fra i livelli di espressione di due soggetti appartenenti alle due classi che si vogliono studiare, non c'è modo di fare inferenza statistica sui dati e di determinare se l'espressione differenziale rilevata sia dovuta alla personale risposta di espressione dei due specifici soggetti analizzati o, piuttosto, al fenomeno obiettivo oggetto dello studio microarray.

Maggiore è la variabilità biologica associata alle classi che si stanno studiando e tanto più ampio dovrà essere il campione collezionato per riuscire a stimare con un buon livello di confidenza statistica il valore di espressione genica differenziale fra le due classi, ossia la media campionaria dell'espressione differenziale.

Negli esperimenti "two-color" questo valore, detto "fold-change", è ottenuto facendo la media o la mediana dei rapporti fra i valori di espressione nei soggetti in cui è osservabile il fenomeno da studiare e quelli nei soggetti in condizioni normalidi controllo.

Se, inoltre, si vuole ottenere un sufficiente grado di affidabilità sui singoli "fold-change" che contribuiscono alla media campionaria o sul livello di espressione assoluta di ciascun soggetto sarà necessario realizzare più misure sullo stesso mRNA (repliche tecniche).

Le repliche tecniche non sono replicati indipendenti dell'espressione di ciascun gene, quindi contribuiscono alla varianza dei dati in maniera differente rispetto alle osservazioni provenienti da repliche biologiche e devono essere statisticamente combinate per ottenere un unico valore rappresentativo dell'espressione genica per il soggetto considerato.

Le repliche tecniche, dunque, producono un valore più affidabile dell'espressione nei soggetti replicati, ma non forniscono nessun contributo alla valutazione della varianza biologica. L'unico modo di fare inferenza su quest'ultima resta collezionare un adeguato numero di soggetti per classe da ibridizzare su altrettanti microarray.

Una tecnica che può consentire di abbassare il contributo della varianza biologica senza dover necessariamente ibridizzare tutte le copie biologiche è il "pooling" degli mRNA. Per "pooling" si intende la creazione di un campione di mRNA dall'unione di diversi mRNA provenienti da singoli soggetti della stessa

classe: l’effetto di questa operazione è di mitigare le differenze biologiche di espressione fra soggetti.

Questa tecnica può consentire di risparmiare sul numero di array da ibridizzare. Tuttavia bisogna tener presente che per fare inferenza statistica sui dati sono necessarie molte osservazioni provenienti da “pool” indipendenti, quindi l’osservazione dell’espressione su “pool” di campioni comporta di dover collezionare molti più campioni di quanti ne servirebbero per ottenere la stessa informazione con esperimenti realizzati su singoli soggetti [23].

Un altro elemento, di natura più puramente statistica, che contribuisce all’innalzamento del numero di copie da collezionare è il contenimento dell’errore di tipo I sui risultati, cioè del numero di falsi positivi. Qualunque tecnica statistica utilizzata per fare inferenza su dati produce una certa quota di risultati falsi positivi e nel caso dei microarray questo corrisponde a dichiarare un gene differenzialmente espresso quando in realtà non lo è. Per mantenere questa quota sotto una soglia definita accettabile dallo sperimentatore è necessario che il test statistico adoperato sia sufficientemente conservativo. La conservatività di un test è inversamente proporzionale alla potenza, cioè quanto più si vogliono contenere i falsi positivi tanto meno il risultato sarà esente da falsi negativi. Questo implica che ripetendo l’esperimento con un campione più ampio si potrebbero trovare risultati differenti. L’unico metodo per ottenere un risultato con il giusto livello di significatività statistica e potenza è avere un numero adeguato di soggetti sui quali effettuare l’esperimento.

Esistono diverse formule per determinare il numero opportuno di copie biologiche e di repliche tecniche utili a determinare una prefissata variazione di espressione genica con adeguata robustezza statistica [2-5, 8, 10, 11, 24, 25].

Queste modalità di calcolo sfruttano i concetti generali di calcolo della numerosità campionaria, rimodellandoli sulla base degli schemi di confronto fra campioni che è possibile realizzare in un esperimento microarray, come verrà illustrato successivamente.

Come regola generale è possibile assumere che siano necessarie almeno sei copie biologiche per classe [26] e tre repliche tecniche [2] per soggetto per realizzare un esperimento microarray che presenti un minimo di robustezza per le procedure di inferenza statistica.

Purtroppo, dal punto di vista pratico esistono limitazioni di varia natura alla corretta applicazione della teoria appena esposta e lo sperimentatore si ritrova frequentemente a dover accettare dei compromessi dettati dal fatto che spesso non è possibile eseguire un esperimento disegnato con il massimo livello di replicazione.

Una prima limitazione può essere rappresentata dalla difficoltà nel collezionare i campioni, che non sempre sono facilmente reperibili. Se, per esempio, si volesse realizzare uno studio per confrontare l’espressione genica di porzioni di tessuto cerebrale umano sede di scariche epilettichead attività epilettogena, rispetto a quellea di tessuto cerebrale normale, sarebbe necessario collezionare alcune decine di campioni di tessuto cerebrale per ciascuna delle due classi. Un così alto numero di campioni serve per moderare l‘elevata variabilità biologica presente in un organismo così complesso come è l’uomo.

Il reperimento deiottenere campioni umani, inoltre, presenta sempre spesso presenta notevoli problemi a causa sia della scarsa disponibilità sia degli inevitabili problemi di disponibilità e di ordine eticoetici coinvolti.

Un'altra limitazione deriva dalla necessità di replicare l'osservazione relativa a ciascun soggetto a livello sperimentale, realizzando più ibridizzazioni per lo stesso mRNA (repliche tecniche) e , innalzando notevolmenteaumentando così il costo di esperimenti economicamente già molto onerosi.

1.2. Fattori di confondimento

Gli esperimenti microarray sono estremamente sensibili a diversi fattori sperimentali diversi, come ad esempio. Alcuni di essi sono: le condizioni ambientali di stabulazione degli animali utilizzati per lo studio, le procedure di semina o di trasfezione delle colture cellulari, la fase di estrazione dell'RNA, la marcatura o l'ibridizzazione dell'mRNA.

Se, per esempio, per un esperimento i campioni appartenenti ad una classe verranno preparati tutti insieme in una sessione sperimentale, mentre gli altri in una separata sessione e ciascuna sessione verrà effettuata in giorni differenti, non sarà possibile separare l'effetto di un eventuale fattore di confondimento (temperatura, peggior resa di un "kit", etc) dall'espressione differenziale.

L'ideale procedura per eliminare il fattore di confondimento prevede di processare tutti i campioni insieme, ma, siccome questo principio teorico non è fisicamente realizzabile, allora è necessario adottare uno schema di lavoro randomizzato che sia progettato per bilanciare il tipo e il numero di campioni rispetto alla procedura sperimentale che si deve realizzare. Se, per esempio, si devono estrarre otto campioni per classe di RNA (16 campioni di mRNA totali) e durante una giornata si programma di estrarne quattro, allora la procedura di randomizzazione consiglierebbe di estrarre due campioni appartenenti ad una classe e due appartenenti all'altra. Allo stesso modo non è consigliabile suddividere gli otto campioni in gruppi sbilanciati (per esempio 4+4+3+2+3).

La stessa regola vale per la marcatura o l'ibridizzazione: l'ideale sarebbe marcare tutti i campioni in un'unica sessione sperimentale o ibridizzare tutti i vetrini insieme, ma è possibile suddividerli in gruppi avendo l'accortezza di randomizzare il più possibile la distribuzione.

1.3. Schemi di confronto dei campioni

L'ibridizzazione dei campioni appartenenti alle classi che si vogliono confrontare può essere progettata seguendo diversi schemi: ciascuno di essi deve essere scelto in relazione agli aspetti dell'espressione genica che lo studio si è prefissato di indagare.

Gli schemi maggiormente utilizzati sono tre: il "Reference Design" (RD), il "Balanced Block Design" (BBD) e il "Loop Design" (LD). Ciascuno di essi può essere progettato con o senza inversione della marcatura dei campioni ("Dye-Swap" Design (DSD)). Tutti e tre gli schemi sono ugualmente efficaci nel realizzare uno studio di "class comparison", ma la scelta di uno schema a scapito degli altri deve avvenire sulla base della determinazione di quale risorsa è maggiormente disponibile: i microarray da acquistare o i campioni da collezionare.

Gli obiettivi di uno studio possono essere essenzialmente di due tipi: valutare accuratamente l'intensità relativa a ciascun campione oppure valutare accuratamente la media campionaria delle differenze fra le due classi.

L'efficienza di un disegno è misurata come la sua capacità di produrre stime precise di questi obiettivi ed è, quindi, inversamente proporzionale alla loro varianza.

1.3.1. "Reference Design"

Il RD (Figura 1.1) è il disegno più semplice e più diffuso e prevede l'uso di un campione di RNA di riferimento (R in Figura 1.1), marcato sempre con lo stesso fluorocromo, da ibridizzare su ciascun array insieme ad un campione "non-reference", cioè un soggetto appartenente a una delle due classi, marcato sempre con l'altro fluorocromo.

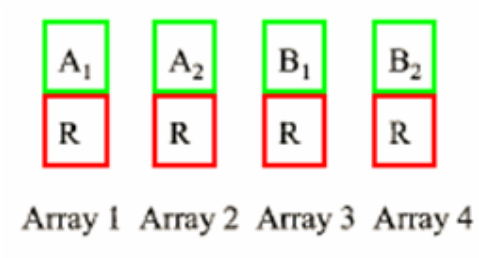


Figura 1.1: Schema del "Reference Design" per quattro campioni appartenenti a due classi

Il confronto fra campioni "non-reference" viene definito *indiretto* perché utilizza il campione di riferimento come tramite: il "fold-change" viene, quindi, ricavato come rapporto dei rapporti fra i campioni "non-reference" e il campione di riferimento. Grazie alla presenza del campione di riferimento, questo schema di ibridizzazione è particolarmente efficiente per misurare accuratamente le intensità dei singoli campioni (intensità assolute), quindi è il disegno più utilizzato per gli esperimenti di "class discovery" e "class prediction" Inoltre, viene generalmente adottato negli esperimenti di "class comparison" quando il numero di campioni è una risorsa limitata, ma non ci sono limitazioni all'acquisto degli array che servono per realizzare l'esperimento.

Una volta che è stato deciso di utilizzare un RD è necessario determinare quanti campioni servono per rilevare il "fold-change" desiderato con un livello opportuno di significatività statistica e di potenza. La formula che generalmente viene utilizzata per effettuare questo calcolo è:

$$n = \frac{4\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2}{\left(\frac{\delta}{\sigma}\right)^2} \qquad (1.1)$$

dove:

- n = numero totale di soggetti da collezionare, quindi n/2 per ogni classe;
- z = percentili della distribuzione normale standard;
- α = livello di significatività prefissato;
- β = livello di potenza prefissato;

- δ = logaritmo in base due del "fold-change" da detettare;
- σ = deviazione standard (SD) dei rapporti logaritmici fra i valori di espressione di ciascuna classe e il campione di riferimento. Si tratta della varianza biologica di ciascuna classe (o intra-classe).

Nella pratica comune il valore di α è fissato a 0.001, cioè 1‰ dei geni sarà falso-positivo, mentre quello di β è fissato a 0.05, cioè il livello di espressione genica δ sarà detettato con il 95% di potenza.

Utilizzando questa formula e l'approssimazione t-Student della distribuzione normale è possibile determinare quanti soggetti servono in totale per realizzare l'esperimento in RD.

1.3.2. "Balanced Block Design"

Il BBD (Figura 1.2) confronta direttamente i campioni appartenenti alle classi, per cui è particolarmente efficiente nella stima delle loro differenze. Per questo motivo questo schema è adottato per realizzare gli esperimenti di "class comparison" nei quali è possibile collezionare tutti i campioni necessari, ma bisogna contenere i costi per l'acquisto dei microarray. Infatti, dal confronto fra Figura 1.1 e Figura 1.2 è possibile osservare che mentre nel RD sono stati ibridizzati quattro campioni "non-reference" su altrettanti array, adottando un BBD con lo stesso numero di array è possibile ibridizzare otto campioni "non-reference". Il BBD produce il doppio risultato di risparmiare sugli array da acquistare e stimare meglio il "fold-change" della popolazione.

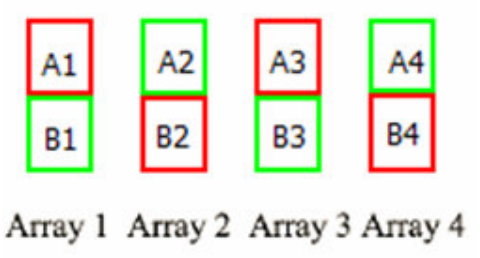


Figura 1.2: Schema del" Balanced Block Design" per otto campioni appartenenti a due classi

Il BBD è una versione ridotta del "Dye-Swap Design" (DSD). Lo scopo del DSD è quello di correggere un effetto di distorsione dell'intensità di segnale dovuta alla capacità di alcune sequenze di mRNA di incorporare una maggiore o minore quantità di uno dei due fluorocromi ("dye-effect" gene-specifico): questo modifica l'intensità rilevata, e di conseguenza il "fold-change", in maniera non dipendente dal fenomeno studiato attraverso l'esperimento.

Per correggere questo problema, nel DSD ciascuna coppia di campioni viene ibridizzata due volte su due array diversi invertendo la marcatura di ciascun campione per ogni ibridizzazione, per cui, per un numero fissato di array, si dimezza il numero di campioni "non-reference" osservato e si raddoppia il numero di osservazioni per ciascun campione.

Si supponga, per esempio, di poter acquistare soltanto 10 array. Se è stato scelto il DSD ci sarà la possibilità di ibridizzare solo cinque campioni per ciascuna classe, mentre utilizzando un BBD potranno essere ibridizzati 10 campioni per ciascuna classe.

Il DSD produrrà una più accurata valutazione del "fold-change" per ciascuna delle cinque coppie, anche se la media dei "fold-change" sarà meno simile alla media di popolazione poichè valutata solo su cinque osservazioni. Nel BBD invece, seppure a scapito dell'esatta valutazione dei singoli "fold-change" e del "dye-effect" gene-specifico, si otterrà una valutazione della media campionaria dell'espressione differenziale più vicina a quella che dovrebbe essere la media di popolazione proprio grazie al maggior numero di osservazioni a disposizione.

Per stabilire quanti soggetti è necessario collezionare bisogna utilizzare una versione modificata della formula 1.1:

$$n = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2}{\left(\frac{\delta}{\tau}\right)^2} \qquad (1.2)$$

dove:

- n = numero di soggetti per ciascuna classe;
- τ = SD dei rapporti logaritmici fra i soggetti appartenenti alle due classi. Essa è la variabilità biologica intra-classi.

La SD del BBD è maggiore di quella del RD perché è una combinazione della variabilità di entrambe le classi; infatti, mentre per il RD il rapporto delle intensità viene calcolato sempre rispetto allo stesso campione di riferimento, nel BBD esso coinvolge due campioni appartenenti a due classi diverse, quindi dipende dal particolare accoppiamento realizzato sul vetrino.

Tipicamente i pochi dati che si trovano in letteratura riguardo al valore della variabilità biologica si riferiscono alla σ. Per questo motivo sono state messe a punto delle formule matematiche che consentono di calcolare il valore da assegnare a τ sulla base della conoscenza di σ [7] in modo da poter fare il calcolo della numerosità campionaria correttamente.

Il BBD è definito "bilanciato" perché per ciascuna classe vi è lo stesso numero di soggetti marcati con ciascun fluorocromo. Questo implica il collezionamento di un numero pari di soggetti per ogni classe e l'esclusione di una coppia di soggetti se si verifica qualche problema sperimentale anche solo su un soggetto.

1.3.3. Disegno sperimentale a "Loop"

Esiste un terzo disegno sperimentale che cerca di mettere insieme i pregi del disegno con riferimento e quelli del confronto diretto in "dye-swap": il disegno a "loop" (Figura 1.3).

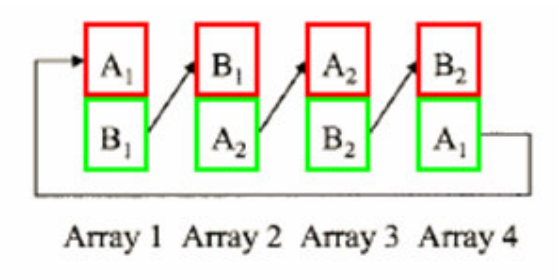


Figura 1.3: Schema del disegno a “loop”.

Questo tipo di disegno, introdotto da Kerr e Churchill [24], utilizza lo stesso numero di array del disegno con riferimento, ma supera il limite fondamentale di quest’ultimo, che consiste nel collezionare il maggior numero di misure sul campione di riferimento e non su quelli di interesse.

Il disegno sperimentale a “loop” realizza il doppio delle misure sulle varietà di interesse e compie un bilanciamento fra i marcatori e le varietà, marcando ogni varietà una volta con un fluorocromo e una volta con l’altro su due array diversi.

Un inconveniente pratico evidente di questo tipo di disegno è il fatto che bisogna realizzare il doppio delle reazioni di marcatura perché ogni campione deve essere marcato con entrambi i fluorocromi. Inoltre, se un array manifesta caratteristiche scadenti, il percorso chiuso che collega tutti i campioni generato con questo tipo di disegno si interrompe e non è possibile ricavare in maniera affidabile il dato di espressione genica differenziale.

Un ulteriore svantaggio del disegno a “loop” è l’impossibilità di ampliarlo con nuovi campioni nel caso in cui si debba proseguire l’esperimento con una fase di ampliamento.

Per evitare questo tipo di inconvenienti, a meno di dover utilizzare questo tipo di disegno per confrontare il livello di ciascun campione con quello di tutti gli altri in un confronto uno a uno, generalmente si preferisce utilizzare un DSD.

Capitolo 2

Metodi di estrazione dei dati grezzi

Il passaggio che traduce l'informazione di colore, contenuta nella fluorescenza emessa dalle molecole utilizzate per marcare i due campioni ibridizzati, in dato numerico viene denominato *quantizzazione*.

L'uso di particolari scanner e di software che applicano complessi algoritmi di manipolazione delle immagini e di digitalizzazione dell'informazione, consente di ricavare i dati preliminari relativi all'esperimento, detti anche dati grezzi. Questi dati dovranno essere successivamente corretti per eliminare le molteplici fonti di variabilità, tipicamente contenute in un esperimento che utilizza microarray, che li corrompono.

2.1 Il processo di quantizzazione del dato

Una volta che il vetrino è stato ibridizzato con i due campioni marcati è necessario sfruttare la capacità di questi marcatori, detti anche fluorofori o fluorocromi, di sviluppare fluorescenza per estrarre l'informazione.

La fluorescenza è una forma di energia prodotta da particolari molecole o materiali in risposta all'assorbimento di quanti energetici prodotti da una sorgente di energia stabile (per esempio un laser), con lunghezza d'onda fissa e specifica per ciascun materiale e intensità variabile e regolata dal voltaggio applicato alle piastre del tubo fotomoltiplicatore del laser.

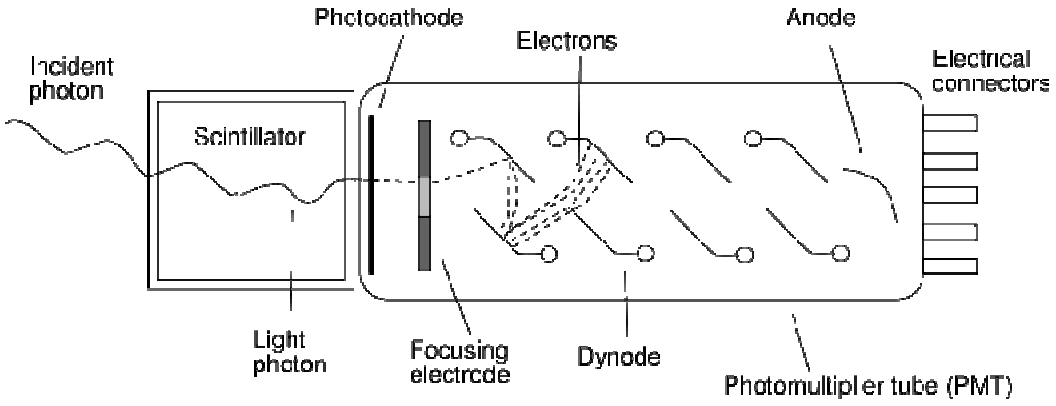


Figura 2.1: Schema del tubo fotomoltiplicatore di un laser

La molecola, a causa dell'assorbimento del fascio energetico, passa da uno stato di equilibrio energetico, ad energia più bassa, ad uno stato eccitato, che è fortemente instabile. Successivamente infatti, essa tenderà a ritornare spontaneamente allo stato di equilibrio emettendo l'energia accettata sottoforma di fluorescenza.

Per produrre l'energia necessaria all'eccitazione dei due fluorocromi vengono utilizzati appositi scanner che contengono due sorgenti di luce laser capaci di eccitare in maniera differente e specifica. Le lunghezze d'onda di assorbimento di fluorocromi tipicamente utilizzati negli esperimenti con microarray sono 532 nm e 635 nm, che corrispondono all'emissione in fluorescenza dei colori rispettivamente verde e rosso. Diverse molecole fluorescenti attualmente in commercio possono essere eccitate utilizzando queste lunghezze d'onda. Le più diffuse sono le Cianine, Cy3 per il verde e Cy5 per il rosso (Amersham Biosciences, Pittsburg, PA) e i fluorofori Alexa, Alexa 555 per il verde e Alexa 647 per il rosso (Invitrogen Corporation, Carlsbad, CA).

Gli scanner che producono l'energia per l'eccitazione dei fluorofori hanno anche la capacità di acquisire separatamente i segnali da essi prodotti e riescono a generare in questo modo due immagini, ciascuna relativa ad uno dei due campioni ibridizzati.

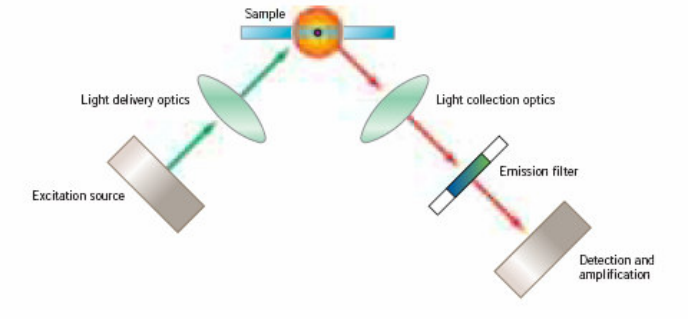


Figura 2.2: Catena di generazione del segnale per eccitazione del fluorocromo con scanner laser.

Le intensità di fluorescenza acquisite vengono salvate sottoforma di un'immagine in formato TIFF per i segnali acquisiti da ciascun canale; essa è una mappa d'intensità in due dimensioni della superficie del microarray e i segnali di fluorescenza sono digitalizzati nei suoi pixel. La sovrapposizione delle due immagini produce la tipica immagine formata da *spot* colorati con diverse gradazioni di giallo, se un gene risulta espresso in entrambi i campioni, o di rosso o di verde se un gene è espresso esclusivamente in uno dei due campioni. Con il termine *spot* si identifica sia la posizione fisica che l'insieme delle sonde relative ad un gene occupa sul vetrino che, per antonomasia, il gene stesso.

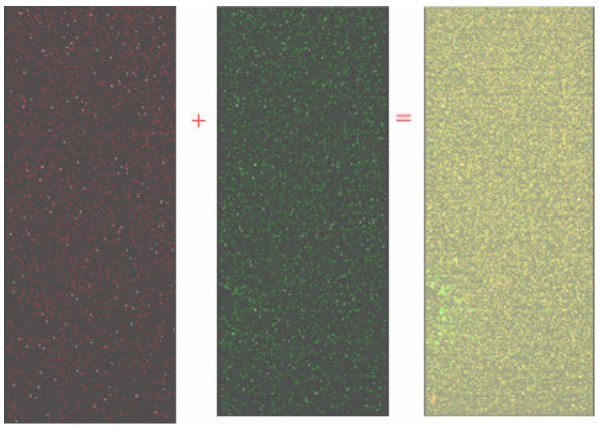


Figura 2.3: Immagini acquisite da ciascun canale dello scanner (sinistra e centro) e immagine prodotta dalla sovrapposizione dei due canali (destra)

L'estrazione dell'informazione contenuta nei pixel che formano ciascuno *spot* può essere suddivisa in tre fasi:

- posizionamento della griglia (gridding) sull'immagine;
- segmentazione;
- estrazione delle intensità del "foreground", ossia del segnale emesso dalle sonde marcate, e del "background", ossia del segnale emesso da fenomeni legati ad ibridizzazione aspecifica sul supporto o emissione impropria di fluorescenza da reagenti.

4.1.1 "Gridding" dell'immagine

Dopo aver portato a termine il protocollo di ibridizzazione dei campioni sul microarray e aver acquisito le immagini con lo scanner a doppio laser è necessario identificare la posizione di ogni *spot* sul supporto. Ciò avviene grazie all'allineamento sull'immagine di una griglia che viene generalmente fornita dal costruttore del microarray.

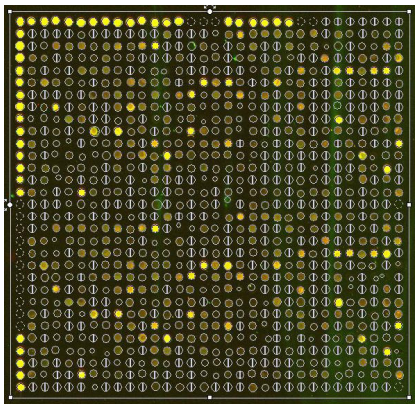


Figura 2.4: Sovrapposizione della griglia digitale all'immagine

Il centro di ciascun cerchietto della griglia identifica l'ideale posizione dell'insieme di sonde depositate sul vertino per detettare un gene, secondo quelle che sono le specifiche costruttive del microarray, e fornisce all'analista diverse informazioni sullo *spot* in esame grazie ad un file allegato che contiene, fra tante altre informazioni, anche il nome del gene corrispondente ad ogni *spot* e i suoi codici d'identificazione nelle banche dati genomiche.

Block	Column	Row	Name	GeneName	Description
1	1	1	BrightCorner	BrightCorner	
1	2	1	BrightCorner	BrightCorner	
1	3	1	NegativeControl	(-)3xSLv1	
1	4	1	AW523361	AW523361	AW523361 UI-R-BO0-aif-d-04-0-UI.s1 UI-R-BO0 Rattus norvegicus cDNA clone UI-R-BO0-aif-d-04-0-UI 3', mRNA
1	5	1	NM_017265	Hsd3b1	Rattus norvegicus hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 1 (Hsd3b1), mRNA
1	6	1	XM_223143	RGD1306921_predicted	PREDICTED: Rattus norvegicus similar to mKIAA1107 protein (predicted) (LOC305122), mRNA [XM_223143]
1	7	1	AI228129	AI228129	AI228129 EST224824 Normalized rat brain, Bento Soares Rattus sp. cDNA clone RBRCQ67 3' end, mRNA sequer
1	8	1	AF290213	Cacna1h	Rattus norvegicus calcium channel alpha-1-H subunit mRNA, complete cds. [AF290213]
1	9	1	AW143725	AW143725	AW143725 EST294021 Normalized rat embryo, Bento Soares Rattus sp. cDNA clone RGIBY81 5' end, mRNA seq
1	10	1	XM_223516	RGD1305329_predicted	Rattus norvegicus similar to mKIAA0232 protein (LOC305435), mRNA [XM_223516]
1	11	1	NM_001009275	RGD1310224_predicted	Rattus norvegicus similar to RIKEN cDNA 1810022C23 (predicted) (RGD1310224_predicted), mRNA [NM_001009
1	12	1	XM_234098	Nova1	PREDICTED: Rattus norvegicus neuro-oncological ventral antigen 1 (Nova1), mRNA [XM_234098]
1	13	1	NegativeControl	(-)3xSLv1	
1	14	1	BC071175	Ero1l	Rattus norvegicus ERO1-like (S. cerevisiae), mRNA (cDNA clone IMAGE:7099384), partial cds [BC071175]
1	15	1	NM_001000282	Olr440	Rattus norvegicus olfactory receptor 440 (predicted) (Olr440_predicted), mRNA [NM_001000282]
1	16	1	CB326764	CB326764	UI-R-DZ0-crq-p-04-0-UI.s1 UI-R-DZ0 Rattus norvegicus cDNA clone UI-R-DZ0-crq-p-04-0-UI 3', mRNA sequence |
1	17	1	A_44_P210961	A_44_P210961	Unknown
1	18	1	BE104359	BE104359	BE104359 UI-R-BX0-arn-h-07-0-UI.s1 UI-R-BX0 Rattus norvegicus cDNA clone UI-R-BX0-arn-h-07-0-UI 3', mRNA
1	19	1	TC559654	TC559654	Unknown
1	20	1	A_44_P312053	A_44_P312053	Unknown
1	21	1	TC564206	TC564206	Unknown
1	22	1	EQC	(+)eQC-37	
1	23	1	NM_001000446	Olr1234	Rattus norvegicus olfactory receptor 1234 (predicted) (Olr1234_predicted), mRNA [NM_001000446]
1	24	1	XM_218458	LOC292720	PREDICTED: Rattus norvegicus similar to ETS domain transcription factor ERF (Ets2 repressor factor) (LOC29272
1	25	1	XM_344397	LOC364374	PREDICTED: Rattus norvegicus similar to T cell receptor alpha chain (LOC498510), mRNA [XM_344397]

Figura 2.5: Parte del file contenente le informazioni relative a ciascun gene

Il corretto posizionamento della griglia permette di ricavare un dato consistente sugli *spot*; per questo motivo, spesse volte è necessario controllare l'allineamento *spot* a *spot* e intervenire manualmente su quegli *spot* che non vengono esattamente centrati o delimitati dalla griglia. E' fondamentale, come è facile intuire, che il processo di posizionamento delle sonde sul supporto

avvenga secondo uno schema preciso e ordinato, nel quale la posizione di ciascuno *spot* può essere identificata mediante due coordinate numeriche rispetto ad un punto di riferimento, in modo da agevolare l'identificazione degli *spot* nel processo di "gridding".

4.1.2 Segmentazione

Una volta che gli *spot* sono stati identificati, è necessario separare il contributo del "foreground" da quello del "background"; per questo motivo deve essere riconosciuta la forma di ogni *spot* attraverso una "*spot* mask".

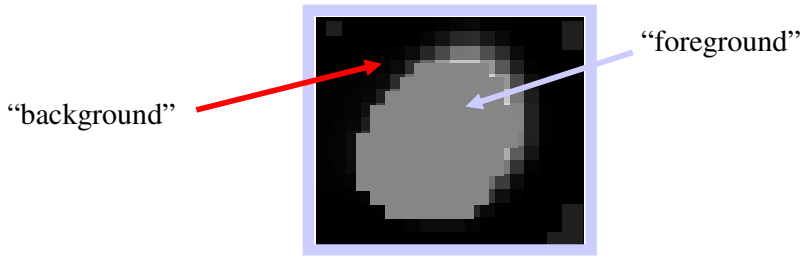


Figura 2.6: Separazione del "background" dal "foreground" attraverso una "spot mask".

Generalmente si assume che gli *spot* abbiano forma circolare di diametro costante; coerentemente con questa ipotesi si identifica come "foreground" tutto ciò che cade all'interno del cerchio e come "background" tutto quello che è all'esterno, operando una *segmentazione spaziale*.

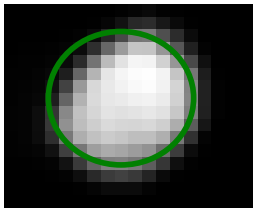


Figura 2.7: Segmentazione spaziale dello spot con griglia di forma prefissata.

Questa semplice assunzione non viene sempre rispecchiata dagli *spot* sul vetrino e ciò è riconducibile solitamente ad errori nella fase di deposizione delle sonde o a ibridizzazione non perfetta dei campioni marcati. Per questo motivo molti software di analisi dell'immagine includono la possibilità di fare una *segmentazione per intensità* dei pixel: in questo procedimento si sfruttano i valori di intensità dei pixel per delimitare l'area da attribuire al segnale, utilizzando algoritmi di "Seeded Region Growing" (SRG) comuni a molti software di manipolazione di immagini.

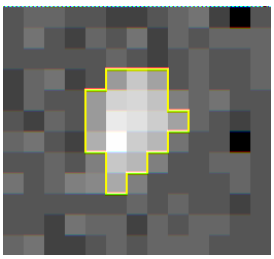


Figura 2.8: Segmentazione per intensità con algoritmo SRG.

4.1.3 Estrazione delle intensità di "foreground" e di "background"

I valori di "foreground" e di "background" possono essere calcolati in diversi modi, fra i quali il calcolo della media e della mediana sono fra i più comuni.

Il calcolo della media del segnale o "average intensity signal" consiste nel rapporto fra la somma delle intensità dei pixel identificati come segnale e il numero totale dei pixel che appartengono alla regione di demarcazione dello *spot*. Un calcolo analogo può essere fatto per la media del "background" prendendo in considerazione solo i pixel identificati come rumore dalla segmentazione. Per calcolare la mediana, invece, si ordinano per valore ascendente o discendente tutti i valori di intensità dei pixel della zona di demarcazione e si prende l'intensità del pixel che si posiziona a metà dell'ordinamento come rappresentativa dell'intera zona. Il valore di mediana di uno *spot* è generalmente più robusto di quello di media e ciò è dovuto al fatto che il suo procedimento di calcolo scarta in maniera automatica quei pixel che vengono definiti contaminanti, cioè quelli che non sarebbero dovuti entrare a far parte della zona di demarcazione che si sta considerando.

Nel calcolo della media viene assegnato uno stesso peso sia a pixel buoni che a pixel che dovrebbero essere scartati attraverso la segmentazione; per questo motivo la media dei pixel si configura come un parametro poco affidabile per stabilire il valore di intensità rappresentativo dello *spot*. Una verifica sulla eventuale discrepanza fra i valori di media e di mediana è un buon metodo per stabilire se la fase di segmentazione è stata condotta correttamente o per valutare i limiti del programma che si sta utilizzando.

Dal punto di vista del formato del dato, ogni canale viene generalmente acquisito in immagini a 16 bit o 20 bit, cioè è possibile discriminare rispettivamente 65.535 o 1.048.576 livelli d'intensità di segnale. Come regola generale i segnali che arrivano rispettivamente a livello 50.000 o 580.000 vengono considerati come limite superiore per una rilevazione del dato affidabile; al di sopra di questo livello il segnale inizia ad andare in saturazione e perciò può essere meno attendibile.

In realtà sarebbe consigliabile mandare in saturazione il minor numero di *spot* e ciò può essere fatto modulando opportunamente il guadagno del tubo fotomoltiplicatore dello scanner in fase di acquisizione dell'immagine. E' anche vero che mantenere un basso guadagno non permette di sfruttare a pieno la dinamica dei fluorocromi e impedisce la rilevazione di segnali deboli che spesso corrispondono a trascritti rari difficilmente identificabili.

Per coniugare la necessità di rivelare anche geni poco espressi, evitando di innalzare i livelli di saturazione del segnale, sono stati messi a punto algoritmi di doppia scansione degli array a due differenti guadagni del fotomoltiplicatore. La prima scansione viene effettuata al 100% del guadagno mentre la seconda al 10% del guadagno; infine, l'algoritmo di quantizzazione dei dati produce un file unico nel quale convergono le informazioni provenienti da entrambe le scansioni. Tale algoritmo è denominato eXtended Dinamic Range (XDR) ed è realizzato sugli scanner prodotti da Agilent Technologies (Agilent Technologies, Palo Alto, CA, USA).

Agilent è anche produttrice di uno dei più completi software per l'automatizzazione del processo di "gridding" allo scopo di eliminare l'errore utente-dipendente (per maggiori dettagli si rimanda al Capitolo 6).

Capitolo 3

Metodi di visualizzazione dei dati

Una volta che i dati grezzi sono stati estratti è consigliabile osservare le loro caratteristiche utilizzando diversi tipi di grafici.

La visualizzazione dei dati, infatti, può aiutare ad identificare artefatti che devono essere risolti utilizzando particolari tecniche nei passi successivi di analisi.

Questi strumenti di visualizzazione sono utili anche nel prosieguo dell'analisi per verificare l'effetto dei passaggi di pre-trattamento dei dati.

3.1 Scatterplot

Lo *scatterplot* è il grafico più semplice che si può utilizzare per visualizzare i dati.

Esso non è altro che un grafico cartesiano che presenta sull'asse delle ascisse i valori di intensità del canale verde e sull'asse delle ordinate i valori di intensità del canale rosso per ciascuno spot, o viceversa.

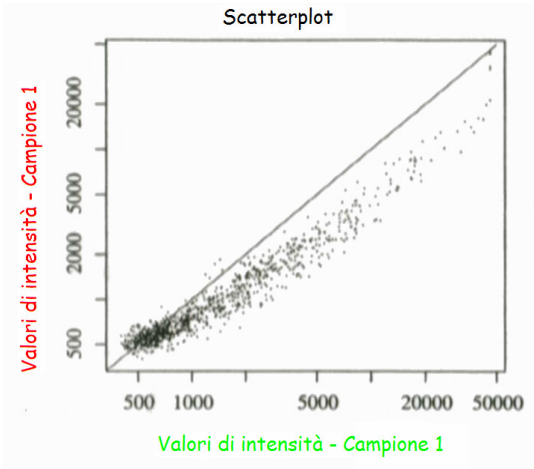


Figura 3.1: Scatterplot di due aliquote dello stesso mRNA ibridizzate sullo stesso vetrino.

In questo grafico geni con livelli di espressione simile si posizionano nei pressi della bisettrice, mentre geni che hanno livelli di espressione differente nei due campioni visualizzati si distanzieranno da questa retta in maniera proporzionale alla differenza. In particolare, geni per i quali il livello di espressione è maggiore nel campione marcato con il fluorocromo rosso si posizioneranno nella porzione di piano al di sopra della bisettrice, mentre geni con livello di espressione più alto nel campione marcato di verde apparterranno al semi-spazio al di sotto della bisettrice. Dallo *scatterplot* è già possibile identificare se vi sono anomalie nelle distribuzioni delle intensità dovute, per esempio, alle differenti caratteristiche delle molecole utilizzate per marcare i campioni da ibridizzare. In figura 3.1 si può osservare la tipica forma definita "a banana" della distribuzione dei dati. In generale da questa figura si potrebbe erroneamente pensare che il campione marcato con il fluorocromo verde sia espresso in maggior quantità rispetto all'altro campione. Tuttavia, questo *scatterplot* si riferisce ad un esperimento nel quale due aliquote dello stesso campione di mRNA sono state marcate con i due fluorocromi e successivamente ibridizzate sul microarray (ibridizzazione self-self). La distribuzione dei dati dovrebbe essere localizzata intorno alla bisettrice, dunque ciò che si osserva è dovuto agli effetti lineari e non lineari di distorsione delle intensità originati dal sistema di rivelazione del segnale della metodica.

Lo stesso tipo di distribuzione viene generalmente osservata anche negli esperimenti nei quali vengono confrontati campioni provenienti da due classi sperimentali differenti. In questo caso lo *scatterplot* fornisce un chiaro indizio sulla necessità di utilizzare particolari metodi di correzione delle distorsioni, illustrati nel capitolo 5.

La limitazione principale dello *scatterplot* è quella di essere un grafico bidimensionale: esso consente quindi di visualizzare solo un array per volta e non abilita il confronto fra più array contemporaneamente.

3.2 MA plot ed RI plot

Il grafico MA deriva da una trasformazione matematica dello *scatterplot* e viene utilizzato per osservare l'abbattimento della variabilità dei dati sui "fold-change" in scala logaritmica.

I "fold-change" relativi ad un esperimento possono essere più facilmente osservati se si utilizza la trasformata logaritmica. Infatti, dopo questa semplice trasformazione si assegna un uguale intervallo di rappresentatività sia ai geni sotto-espressi che a quelli sovra-espressi, passando da 0→1 e 1→+∞ rispettivamente a -∞→0 e 0→+∞.

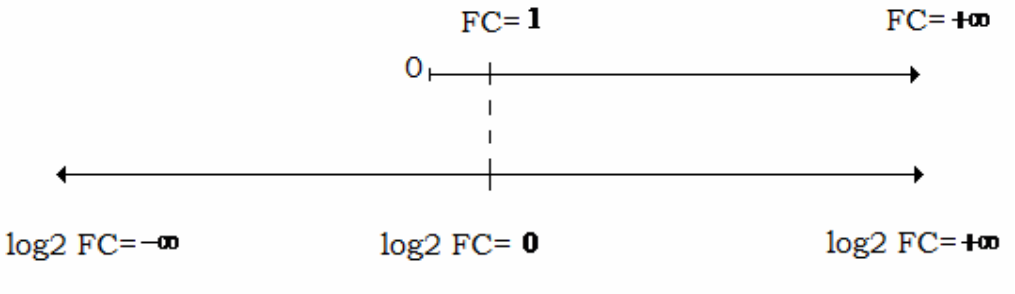


Figura 3.2: Intervalli di rappresentatività dei "fold-change" (FC) e dei log-"fold-change"

La trasformazione logaritmica ha anche l'effetto di trasformare la distribuzione "skewed" dei "fold-change" in una distribuzione gaussiana, molto più utile ed utilizzata negli strumenti statistici di analisi dei dati.

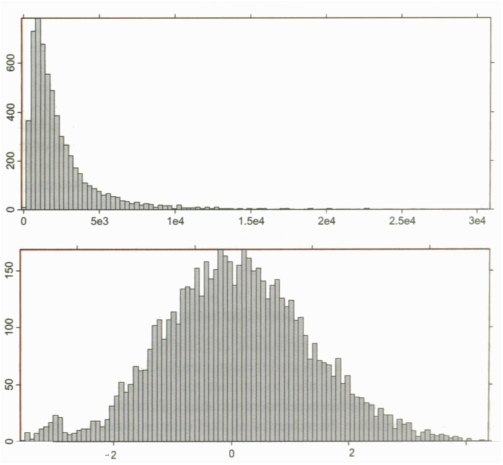


Figura 3.3: Distribuzione "skewed" dei fold-change (in alto) e distribuzione gaussiana dei log-fold-change (in basso)

Le trasformazioni matematiche dei dati che vengono utilizzate per generare un MA plot sono:

$$M = \log2(R/G) \qquad (3.1) \qquad A = \tfrac{1}{2} \log2(R\,G) \qquad (3.2)$$

Anche sul grafico MA è visualizzabile la tipica forma a banana provocata dalla distorsione dei dati. Su questo grafico gli spot con maggior espressione del campione marcato di verde sono posizionati al di sotto dell'asse delle ascisse, mentre quelli con maggior espressione nel campione marcato di rosso sono posizionati al di sopra dell'asse delle ascisse.

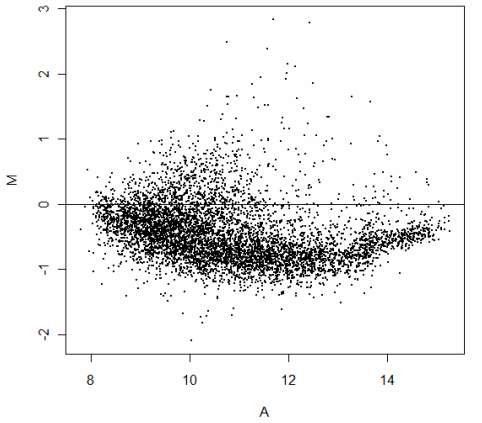


Figura 3.4: MA plot di un esperimento

Il grafico RI (Figura 3.5) è una variante del grafico MA. Anche in questo caso sull'asse delle ordinate è visualizzato il logaritmo del rapporto delle intensità dei due canali, mentre in ascissa vi è il logaritmo del prodotto, che è più intuitivamente collegato al segnale globale di uno spot rispetto alla media geometrica dei due canali espressa con il valore di A.

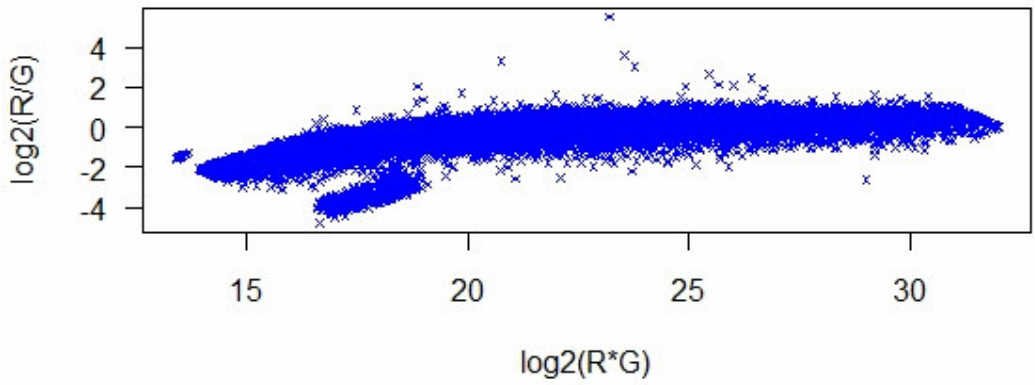


Figura 3.5: RI plot di un esperimento

3.3 M plot "diagnostici"

La trasformazione logaritmica dei "fold-change" di espressione consente di visualizzare l'effetto dellea tecniche di pre-trattamento dei dati.

I dati di espressione genica sono generalmente affetti da rumore che deriva da molteplici fonti legate sia ai passaggi dialla preparazione dei campioni e alla loro ibridazione sul vetrino, che al metodo di rivelazione deil datio attraverso la fluorescenza. Tale rumore, che si traduce in una inaffidabile quantificazione del segnale di intensità proveniente dalle sonde ibridizzate con i

campioni le sequenze marcatei, può essere costituito generato da attraverso due contributi: lo sporco presente sul vetrino, si parla in questo caso di "background", e gli effetti di distorsione lineari e non lineari o intensità-dipendenti dell'emissione di fluorescenza. Nel primo caso è possibile eliminare il "background" attraverso la sua quantificazione e la successiva sottrazione dal "foreground". Nel secondo caso è necessario utilizzare opportune tecniche di correzione delle distorsioni realizzando quella che si chiama normalizzazione dei dati (per maggiori dettagli si rimanda ai capitoli 4 e 5).

3.3.1 M-Mb plot

L'effetto dell'eliminazione dell'intensità del "background" dal "foreground" può essere osservato attraverso un grafico denominato M-M_b. I dati utilizzati per costruire questo grafico sono costituiti dai valori dei log-fold-change relativi al "foreground" senza aver effettuato la sottrazione del "background", ma avendo eliminato le distorsioni (M), e dai valori di "background" normalizzati (M_b).

Il grafico M-M_b ha lo scopo di visualizzare se esiste correlazione fra i valori di M ed M_b. Questa correlazione viene generalmente misurata utilizzando il coefficiente di correlazione di Spearman e si assume che sia consigliabile fare la sottrazione del "background" se il suo valore supera 0.2. La presenza di correlazione fra i valori di M e quelli di M_b per ciascuno spot fornisce una quantificazione matematica di eventuali effetti locali di innalzamento del "background". Infatti, se il "il background" fosse uniformemente distribuito su tutto il vetrino, cioè se si abbattesse in uguale quantità su tutti gli spot, l'operazione di sottrazione sarebbe inutile visto che il contributo del rumore rappresenterebbe una quota fissa del valore di "foreground" per ciascuno spot. La correlazione fra i due segnali è, in questo caso, completamente assente.

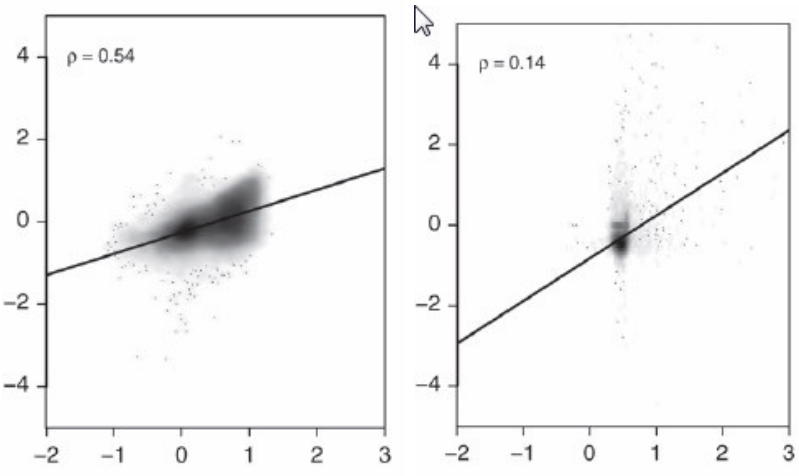


Figura 3.6: Grafici M- M_b di due array con alta (sx) e bassa (dx) correlazione fra M e Mb

Al contrario, la presenza di ampi effetti spaziali di rumore sul vetrino tende a far aumentare la correlazione fra i due segnali e rende fortemente consigliata l'operazione di sottrazione del "background [27], anche se essa produce un notevole incremento della varianza delle misure alle basse intensità, visualizzabile su un grafico MA con il tipico "effetto ventaglio".

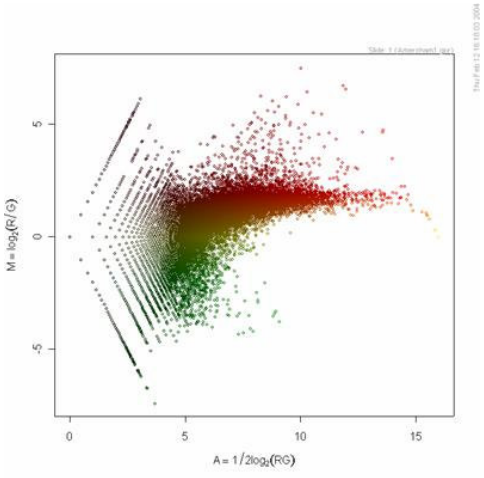


Figura 3.7: Effetto ventaglio visualizzato su un grafico MA

Il problema di aumento della varianza alle basse intensità generato dall'applicazione della sottrazione del "background" è ormai ampiamente superato con le nuove tecniche di sottrazione (vedi Capitolo 4). Infatti, l'uso di complicate tecniche di modellazione statistica della distribuzione del "foreground" e del "background" consente di procedere con la sottrazione senza introdurre una ulteriore componente di rumore a bassa intensità. La quantificazione della dipendenza del segnale dal rumore resta, tuttavia, una tecnica valida per consentire di identificare se il protocollo di ibridizzazione abbia prodotto qualche artefatto che può far pensare ad una revisione dello stesso.

3.4 Image-plot

La presenza di eterogeneità spaziale del "background" può essere visualizzata utilizzando una mappa bidimensionale in falsi colori dell'array, denominata *Image-plot*.

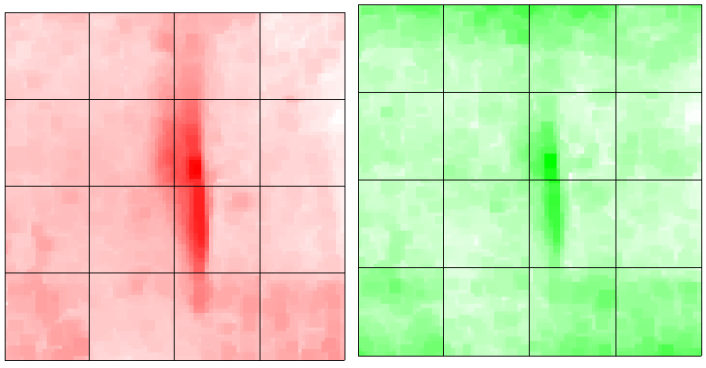


Figura 3.8: Image-plot del "background" di un array su canali separati

Sulle mappe in figura 3.8 è individuabile una zona centrale nella quale il "background" ha un livello più alto di intensità per entrambi i canali. Tuttavia,

la visualizzazione su due grafici separati non consente di quantificare se il contributo del "background" al "foreground" è superiore per uno dei due canali.

Dopo aver individuato l'area eterogenea è, quindi, molto più informativo rappresentare in un *Image-plot* entrambi i canali.

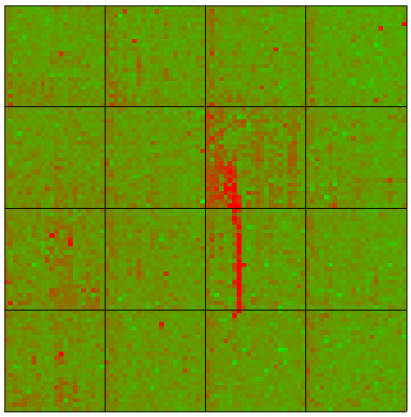


Figura 3.9: Image-plot dei due canali sovrapposti

Come è possibile osservare in figura 3.9 il contributo maggiore è dato dal segnale rosso, come dimostra la striscia rossa al centro dell'array in posizione identica alle due zone eterogenee individuate in figura 3.8.

L'*Image-plot* è anche utile per visualizzare l'effetto delle tecniche di eliminazione delle distorsioni lineari dei segnali di intensità. In particolare, se uno dei due canali mostra di avere un'intensità globale maggiore e ciò è riconducibile esclusivamente ad una maggior quantità di energia ricevuta dallo scanner piuttosto che ad una reale maggior espressione del campione, tale canale sarà predominante nell'*Image-plot* dei canali sovrapposti. Una volta che questo effetto è stato corretto l'*Image-plot* visualizzerà una mappa nella quale i due canali sono completamente bilanciati.

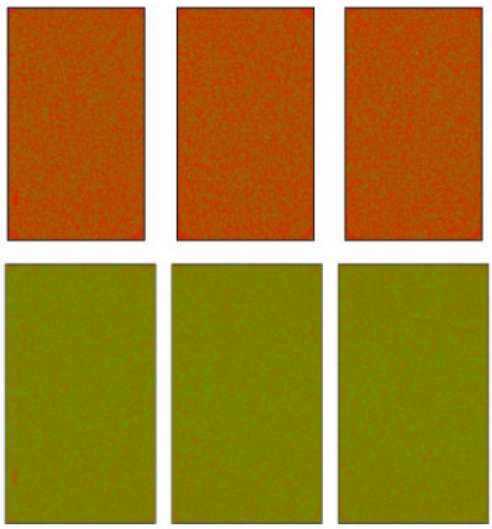


Figura 3.10: Image-plot di un set di tre array prima (sopra) e dopo (sotto) l'eliminazione delle distorsioni lineari

3.5 Boxplot

Nel grafico denominato *boxplot* vengono rappresentate alcune statistiche descrittive di un insieme di dati. Esso prende il suo nome dalla presenza della tipica scatola i cui estremi superiore e inferiore corrispondono alla posizione del quartile superiore, o Upper Quartile (UQ), e inferiore, o Lower Quartile (LQ), dei dati, che sono rispettivamente il 25° e il 75° percentile, mentre la linea centrale è la posizione della mediana dei dati, o 50° percentile. La scatola contiene, quindi, il 50% dei dati.

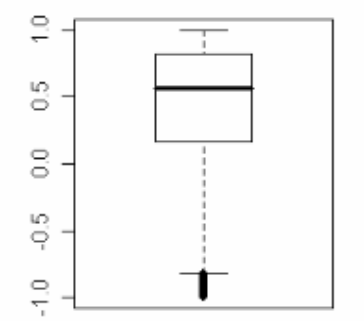


Figura 3.11: Boxplot di un insieme di dati

La distanza fra i quartili superiore e inferiore viene denominata *Distanza o Range Inter-Quartile* (IQD o IQR) e la distanza ricoperta dalla linea tratteggiata corrisponde a 1.5 IQD. Tutti i dati che cadono al di fuori della distanza ricoperta da UQ + 1.5 IQD o LQ – 1.5 IQD vengono considerati dati anomali o "outlier".

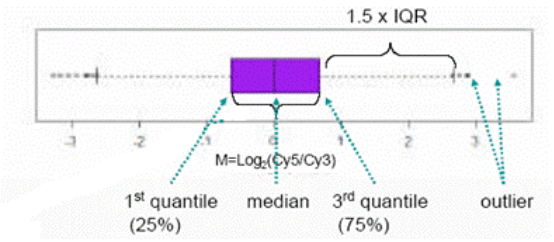


Figura 3.12: Posizione dei quantili nel boxplot e identificazione degli outlier

La IQD fornisce anche informazioni qualitative sull'ampiezza della dispersione dei dati. Tanto più sono distanti UQ e LQ tanto maggiore sarà la dispersione dei dati. Analogamente la posizione asimmetrica della mediana è indicativa di una dispersione differente fra i dati che appartengono al quartile superiore e quelli che appartengono a quello inferiore.

Il boxplot è estremamente utile nella visualizzazione dei dati microarray perché evidenzia sbilanciamenti fra array.

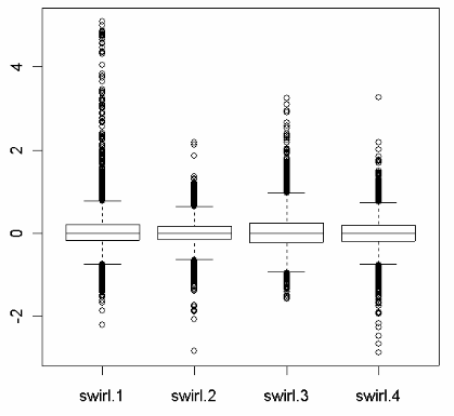


Figura 3.13: Boxplot di quattro array

In figura 3.13 si può notare come il primo e il terzo array presentino una dispersione maggiore del secondo e del quarto array, malgrado i set di dati utilizzati provengano da ibridizzazioni di campioni molto simili fra di loro. Questa disomogeneità individua la condizione per l'applicazione di specifiche tecniche di correzione dei dati.

3.6 Density plot

Questo grafico mostra il profilo delle densità empiriche dei dati per ciascun canale. I profili devono essere quasi completamente sovrappostio e se ciò non viene rilevato è a causa delle distorsioni lineari e non lineari dei segnali.

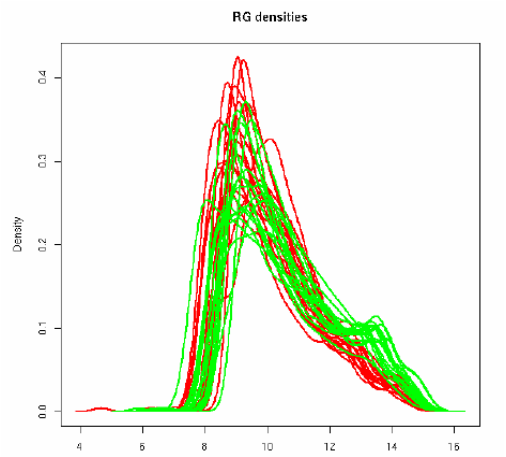


Figura 3.14: Density plot dei segnali di intensità di 10 array

I valori di intensità vengono automaticamente log-trasformati al fine di riconoscere in maniera intuitiva l'eventuale distribuzione statistica associata.

Sul grafico Density è possibile individuare qual è il contributo delle differenti intensità di segnale e avere un'idea sul livello globale di segnale dei singoli canali sui differenti array. In questo modo si possono individuare array con segnale particolarmente basso su uno o entrambi i canali e, dal confronto con gli altri, riuscire a capire se tale segnale debole possa essere dovuto ad un problema di ibridizzazione non ottimale.

Una volta che le distorsioni delle intensità di segnale sono state corrette attraverso opportune tecniche, le densità visualizzate nel grafico Density devono essere pressoché sovrapposte.

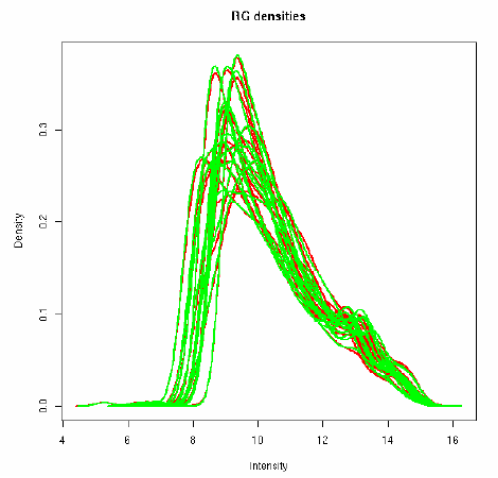


Figura 3.15: Density plot delle intensità corrette

Questo grafico è, dunque, un metodo estremamente efficace per osservare l'effetto che le tecniche di correzione hanno sui dati e fornisce anche un mezzo per comprendere l'entità della correzione.

3.7 Analisi delle Componenti Principali (PCA)

La visualizzazione contemporanea di tutte le osservazioni relative ad un esperimento microarray può aiutare ad avere una visione preliminare dell'andamento dell'espressione genica nell'esperimento. Essa è, tuttavia, estremamente difficoltosa a causa dell'alta dimensionalità di questi esperimenti.

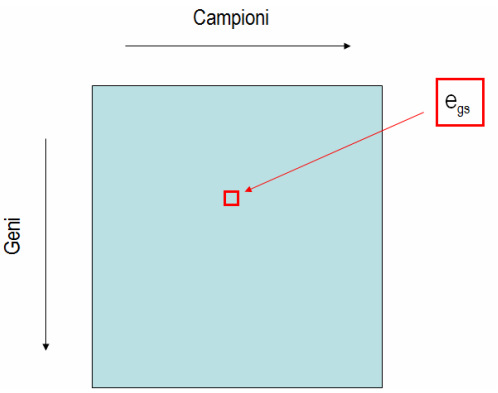


Figura 3.16: Matrice dei dati. Il quadrato rosso individua l'espressione del gene g nel campione s

I dati provenienti da un esperimento microarray vengono generalmente organizzati in una matrice in cui sulle righe sono posizionati i geni e sulle colonne gli esperimenti. In particolare ciascuna riga conterrà i valori di espressione oppure di "fold-change" di ciascun gene, cioè per ogni riga ci sarà il profilo di espressione di ciascun gene. Analogamente, a ciascun esperimento

sarà associato un profilo che è costituito dai valori di espressione di tutti i geni in quel particolare esperimento.

Le due dimensioni matriciali costituiscono altrettanti spazi vettoriali per la visualizzazione dei dati. Il primo di essi, detto spazio dei geni, visualizza ciascun gene come un punto nello spazio e le sue coordinate sono i valori di espressione che assume in tutti gli esperimenti, che rappresentano gli assi del sistema di riferimento. Questo spazio ha quindi dimensione pari al numero degli esperimenti.

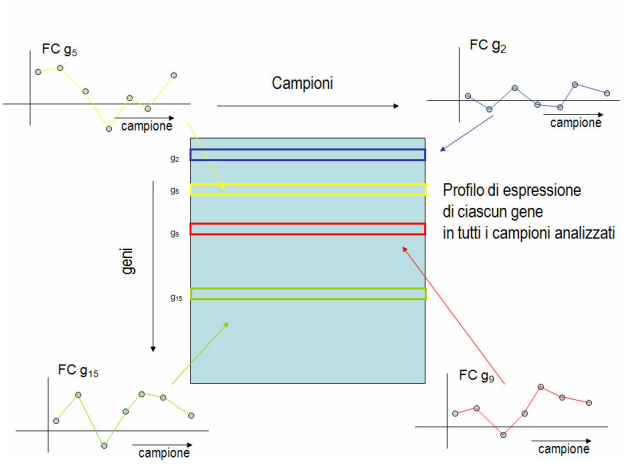


Figura 3.17: Spazio dei geni

Il secondo, detto spazio degli esperimenti, rappresenta il profilo di espressione di ciascun esperimento come un punto nello spazio i cui assi rappresentano ciascuno un gene differente. Questo spazio ha quindi dimensione pari al numero dei geni.

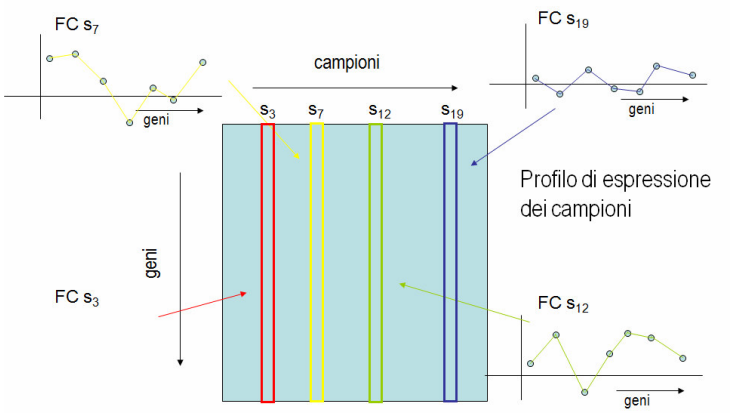


Figura 3.18: Spazio degli esperimenti

L'Analisi delle Componenti Principali (PCA) cerca di ridurre l'alta dimensionalità degli studi microarray individuando quali sono le componenti più importanti dell'informazione contenuta nell'espressione genica ed rappresentando ciascun gene o esperimento soltanto con quelle fra esse che sono realmente esplicative delle differenze fra i dati.

L'informazione è rappresentata dalla dispersione dei dati, per cui si cercano nello spazio di rappresentazione prescelto le direzioni rispetto alle quali i dati presentano la variabilità maggiore e si riduce la rappresentazione solo a quelle direzioni che la "spiegano" (rappresentano) maggiormente.

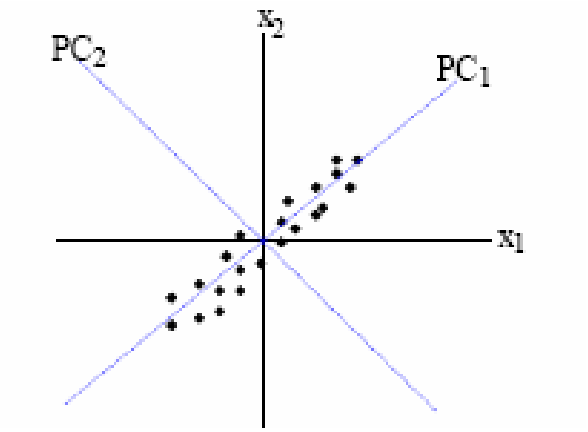


Figura 3.19: Informazione spiegata dalle prime due componenti principali.

Per esempio, in figura 3.19 la direzione che presenta la dispersione, quindi l'informazione, maggiore è quella individuata dalla prima componente principale PC1, mentre lungo la seconda componente principale PC2 i dati presentano una variabilità notevolmente più bassa. Si dice in questo caso che PC1 spiega la maggior parte della varianza.

La PCA è più comunemente utilizzata per visualizzare l'andamento globale dell'espressione genica sui campioni piuttosto che sui singoli geni, per osservare se campioni appartenenti alla stessa classe si posizionano nella stessa zona dello spazio delle osservazioni. Evidenziare somiglianze fra campioni è molto più semplice ed efficace in uno spazio basso-dimensionale, per esempio utilizzando le prime due o tre componenti principali, che generalmente spiegano la maggior parte della variabilità sui dati.

Nell'esempio di figura 3.20 viene misurata l'espressione differenziale in 3 campioni x1, x2 e x3. Le lettere dell'alfabeto indicano la posizione di 21 geni espressi in questo esperimento e le barre blu cercano di dare un aspetto 3D al grafico. E' più o meno possibile identificare una relazione fra x1 e x2, ma è meno chiaro se x3 sia correlato a x1 o x2 o entrambi.

La PCA aiuta ad evidenziare queste relazioni agendo sull'abbattimento della dimensione dello spazio di rappresentazione. In questo caso, lo spazio a 3D può essere ridotto almeno a due dimensioni, se tale riduzione comporta una perdita trascurabile di informazione.

E' intuibile che le direzioni nello spazio rispetto alle quali la nuvola di geni varia maggiormente sono quelle che contengono un'informazione maggiore e, rispetto ad esse, viene effettuato il cambio di sistema di riferimento, p. es. una rotazione, e l'abbattimento di quelle direzioni rispetto alle quali i geni variano di meno.

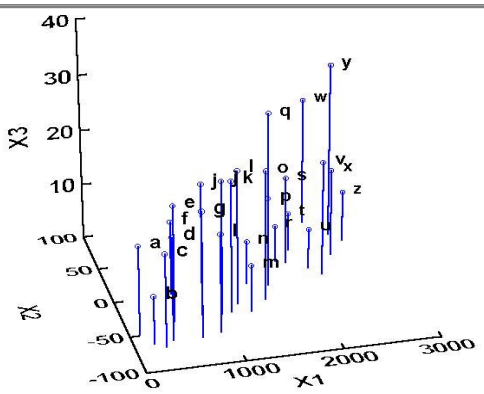


Figura 3.20: Rappresentazione dell'espressione di 21 geni nello spazio degli esperimenti

Prima di calcolare le direzioni delle componenti principali è necessario standardizzare i dati, cioè sottrarre ad essi la media o la mediana e dividerli per la deviazione standard.

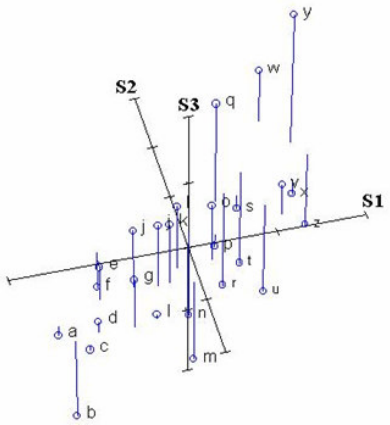


Figura 3.21: Rappresentazione 3D dei dati standardizzati

Le direzioni del nuovo sistema di coordinate PC sono gli autovettori della matrice di covarianza dei profili nello spazio delle osservazioni scelto. Questa matrice modella la distribuzione dei dati, dunque i valori assoluti dei suoi autovalori indicano come i dati si distribuiscono lungo queste direzioni. Tanto più un autovalore ha un valore assoluto alto quanto più i dati avranno una variabilità ampia lungo la direzione individuata dal suo autovettore. A ciascun autovalore corrisponde una quota di varianza spiegata. Nell'esempio di figura 3.20 alle tre PC corrispondono rispettivamente il 66%, il 33% e il 4% della varianza spiegata. Da queste percentuali si può facilmente intuire come la terza PC è ininfluente, cioè la minima componente di varianza spiegata da essa può essere tranquillamente ignorata senza che l'informazione contenuta nei dati venga in alcun modo ad essere corrotta o persa.

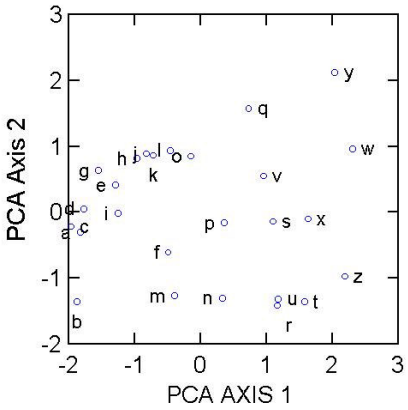


Figura 3.22: Rappresentazione dell'insieme dei dati sulle prime due PC

Per comprendere quale sia il contributo di ciascun campione alla varianza spiegata viene calcolata la matrice delle abbondanze.

	PCA1	PCA2	PCA3
S1	0.9688	0.0664	-0.2387
S2	0.9701	0.0408	0.2391
S3	-0.1045	0.9945	0.0061

Tabella 3.1: Matrice delle abbondanze

Gli elementi di questa matrice forniscono la quota di ciascun campione che viene rappresentata attraverso una determinata componente principale (S_i).

Per esempio la componente principale 1 rappresenta il 96.88 % della varianza di S1 e, mentre S1 e S2 sono positivamente correlate alla PC1, S3 è ad essa negativamente correlato.

3.8 "Heatmap": visualizzazione di somiglianze

I dati di espressione genica possono essere visualizzati utilizzando una scala di colori nella quale a ciascuna sfumatura corrisponde un diverso livello di espressione. Il risultato di questa codifica è una mappa in falsi colori, denominata "heatmap", dell'espressione dei geni coinvolti nell'esperimento.

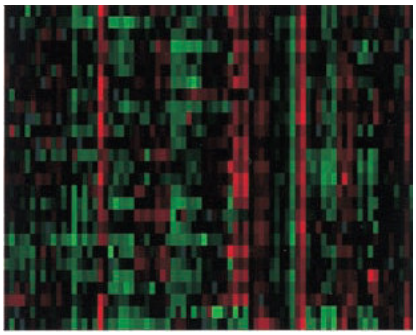


Figura 3.23: Mappa in falsi colori o "heatmap" dei dati di espressione genica

Per questa matrice di colori vale la stessa organizzazione illustrata per la PCA, cioè su ciascuna riga è posizionato il profilo di espressione di un gene mentre su ciascuna colonna vi è il profilo di espressione globale di un intero esperimento.

La visualizzazione dei dati su una "heatmap" diventa molto più informativa se essi vengono riordinati seguendo precisi criteri che hanno lo scopo di spostare in posizione adiacente quei profili che presentano la stessa evoluzione di espressione. A questo scopo è generalmente deputata l'analisi di somiglianze, più comunemente detta "clustering".

Capitolo 4

Metodi di sottrazione del “background”

La fluorescenza che deriva dal “background” contribuisce all’intensità proveniente dal “foreground” per cui è pratica comune sottrarre la sua intensità da quella del “foreground” al fine di ottenere il segnale netto. La necessaria applicazione dell’operazione di sottrazione è alquanto controversa e non esistono chiare linee guida per determinare le circostanze che possono essere favorevoli o sconsigliarne l’esecuzione. La sottrazione del “background” generalmente riduce l’errore nella valutazione dell’intensità, a scapito di un aumento della varianza dei “fold-change” [28-30].

E’ inoltre possibile che a causa della sottrazione i geni con bassa espressione, il cui segnale di intensità è più debole del “background”, vengano eliminati dalla successiva analisi con l’applicazione degli indicatori di qualità dei dati.

Per effettuare la sottrazione del “background” sono stati realizzati molti metodi. Alcuni di essi si basano semplicemente sulla determinazione del livello di sporco globale del vetrino [31]. Altri producono una valutazione locale del livello di sporco su un intorno più o meno ampio dello spot che si sta considerando [32]. Altri ancora modellano il “background” con una distribuzione statistica [33-35].

Tutti i metodi presentati in questo capitolo sono disponibili nel pacchetto di analisi statistica di dati per microarray denominato *LIMMA* [36] di *Bioconductor* [37], http://www.bioconductor.org).

4.1 Il "background"

La presenza di un segnale di fondo sul microarray è dovuta a diversi fattori [38] e la sua determinazione può variare sia in funzione dello scanner che del software utilizzati per quantizzare i dati [31].

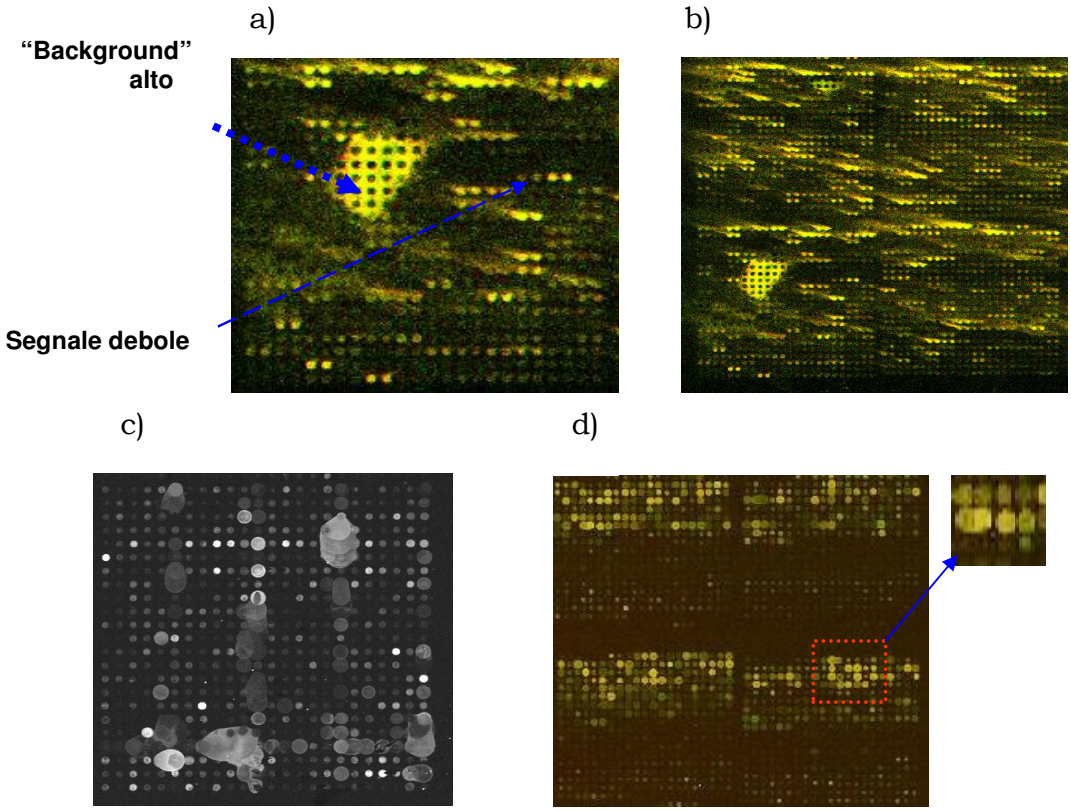


Figura 4.1: a) Microarray con "background" alto dovuto ad ibridizzazione fuori dallo spot.
b) Microarray con "comete" dovute a deposizione non precisa delle sonde.
c) Microarray con depositi irregolari della soluzione di buffer tampone per la deposizione.
d) Microarray con spot sovrapposti e di diametro irregolare.

Può accadere, per esempio, che parte della soluzione contenente le sonde da depositare sul vetrino abbia contaminato aree esterne allo spot consentendo, in questo modo, l'ibridizzazione del campione marcato anche dove non dovrebbe avvenire.-

Un esempio tipico è la rilevazione, in fase di scansione del microarray, delle cosiddette "comete", che possono essere osservate in figura 4.1 a e b. Nella stessa figura si osservano altri esempi di problemi riconducibili alla fase tecnologica di produzione del vetrino. Fortunatamente questo tipo di problemi sono stati superati con l'avvento delle nuove tecnologie costruttive dei vetrini.

Un altro fattore che contribuisce alla generazione di rumore è l'instaurarsi di legami aspecifici fra il supporto del microarray e il campione ibridizzato: in questo caso può essere utile sottoporre il vetrino ad un

procedimento che prende il nome di pre-ibridizzazione, allo scopo di saturare questi legami e non renderli disponibili in fase di ibridizzazione del campione marcato.

Può ancora succedere che i reagenti utilizzati nella soluzione di deposizione abbiano fluorescenza propria oppure siano riflettenti: anche in questo caso è possibile scambiare per “foreground” ciò che in realtà è esclusivamente “background”. Questi valori di intensità spuria possono essere esclusi dall’insieme di dati che vengono in prima istanza considerati come “foreground” attraverso la correzione o sottrazione del “background”.

4.2 Stima del “background”

4.2.1 “Background” locale

Nella stima locale del “background” viene identificato un intorno sufficientemente ampio centrato sullo spot e viene considerata la media o la mediana deil pixel esterni allo spot ma interni alla zona di demarcazione come valore locale del rumore.

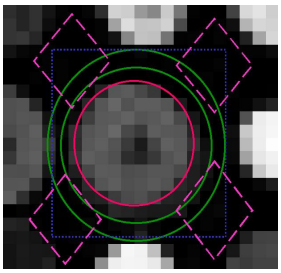


Figura 4.2: Intorno dello spot per il calcolo del “background” in diversi software di analisi.

Con questa operazione è possibile gestire la variabilità locale del rumore, tuttavia non è un procedimento privo di rischi; si pensi, per esempio, a spot con segnale debole: in questo caso la sottrazione di un livello locale di “background” che per qualche motivo risulti particolarmente alto porterebbe all’esclusione dello spot dall’insieme di quelli considerati accettabili. Inoltre, è poco agevole fare un calcolo del genere quando il microarray è particolarmente denso, per evidenti difficoltà che si creano nell’identificare la zona sulla quale impostare il valore di correzione.

4.2.2 “Background” da sotto-griglie

Per ovviare agli inconvenienti appena illustrati l’alternativa possibile è calcolare il valore di “background” su sotto-griglie del microarray; in questo modo si conduce il calcolo su un ambito meno locale e si può riuscire a ricavare una stima del rumore anche su array particolarmente densi di spot.

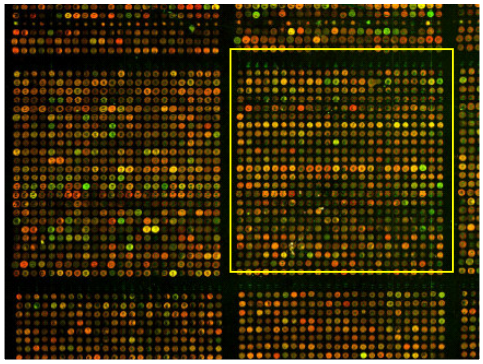


Figura 4.3: Sotto-griglia dell’array sulla quale calcolare il “background”.

4.2.3 “Background” da un intorno ampio dello spot

Una via di mezzo fra i due procedimenti appena illustrati fa uso di un’area centrata sullo spot di diametro tale da includere un gruppo di spot. Lo scopo di questo procedimento è quello di mantenere il computo del “background” su un ambito abbastanza locale ma non troppo ristretto, in modo da poter catturare anche la sua variabilità; grazie all’ampliamento dell’intorno è possibile applicare questo metodo anche su array densi.

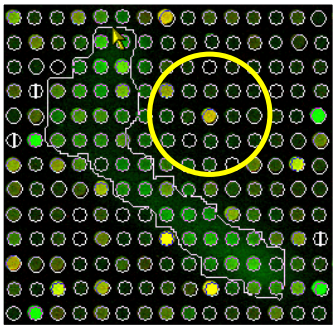


Figura 4.4: Calcolo del “background” su un intorno ampio dello spot.

4.2.4 “Background” da aree dedicate del vetrino

Esistono alcuni metodi di correzione che fanno uso di un fattore correttivo calcolato su aree nelle quali non sono presenti spot. Lo scopo è quello di stimare l’effetto dovuto all’interazione del substrato presente sulla superficie del vetrino con il campione marcato. Questo metodo, tuttavia, non è completamente affidabile, in quando queste aree non sono rappresentative di ciò che realmente avviene dove sono presenti le sonde. Per questo motivo è più utile ricavare il valore di “background” su aree specifiche sulle quali vengono appositamente depositate sonde di controllo, denominate controlli negativi, che si sa non essere complementari alle sequenze del campione in esame.

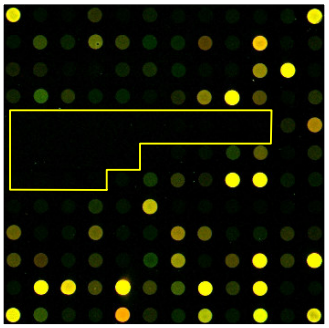


Figura 4.5: Calcolo del "background" su aree vuote del microarray.

4.3 Metodi di sottrazione del "background"

4.3.1 Metodo *subtract*

Il metodo di correzione del "background" denominato *subtract* assume che i segnali di intensità relativi al rumore nei due canali, indicati con R_b e G_b, dove R identifica il canale rosso, G il canale verde e b l'intensità di "background", siano additivi rispetto alle intensità del "foreground", indicate con R_f e G_f per i due canali e dove f identifica il "foreground". Data questa assunzione il segnale netto per ciascun canale e per ciascuno spot, può essere ricavato con:

$$R = R_f - R_b \tag{4.1}$$

$$G = G_f - G_b \tag{4.2}$$

Le stime di R_b e G_b possono essere state ricavate con uno qualunque dei metodi appena descritti. La sottrazione del "background" effettuata con questo metodo può produrre valori d'intensità corretti negativi o nulli che, quindi, daranno luogo ad un valore NA (Not-Avalilable) una volta log-trasformati.

In figura 4.6 si può osservare il grafico MA delle intensità nette per un vetrino realizzato ibridizzando due aliquote dello stesso campione marcate con i due fluorocromi (ibridizzazione self-self).

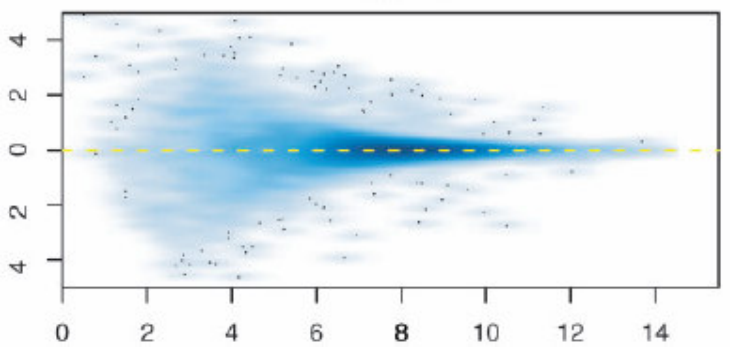


Figura 4.6: MA plot di dati grezzi a cui è stato applicato il metodo subtract

In questo grafico è evidente l'effetto di forte dispersione che la sottrazione del "background" realizzata con questo metodo ha sugli spot a bassa intensità di segnale, già illustrato come "effetto ventaglio".

4.3.2 Metodo *minimum*

Questo metodo è una modifica del metodo *subtract*. Il metodo impone un valore pari a metà del minimo dei valori positivi corretti per ciascun canale su tutto l'array a tutti quegli spot che presentano un valore nullo o negativo dopo la sottrazione. In figura 4.7 si può notare come la sostituzione del valore negativo o nullo prodotto dal metodo *subtract* con il valore *minimum* generi un lieve effetto di stabilizzazione della varianza alle basse intensità.

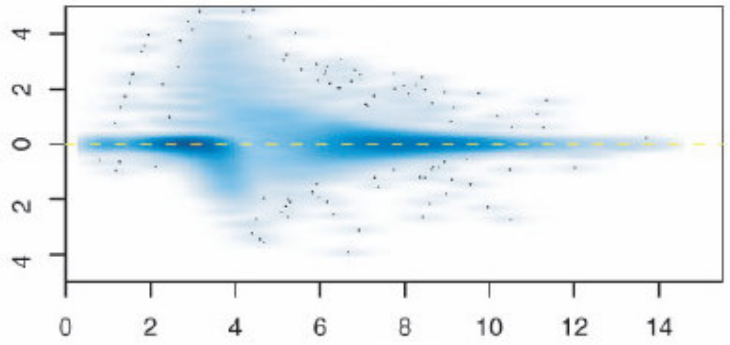


Figura 4.7: MA plot di dati grezzi a cui è stato applicato il metodo minimum

4.3.3 Metodo *Normexp + offset*

Il metodo *Normexp* effettua una modellazione statistica del "foreground" e del "background" e si basa su un modello di convoluzione già sperimentato per fare la correzione del "background" in microarray che utilizzano il protocollo di ibridizzazione one-color [39, 40]. *Normexp* modella le intensità dei pixel osservati come somma di due variabili random, una normalmente distribuita e una esponenzialmente distribuita, che rappresentano rispettivamente le intensità di "background" e di "foreground". La stima dei parametri del "kernel" di convoluzione avviene grazie ad uno stimatore di massima verosimiglianza, cui viene applicata l'approssimazione al punto-sella in modo da rendere la valutazione dello stimatore computazionalmente più leggera.

Il metodo *Normexp* può essere utilizzato con o senza l'aggiunta di un *offset*, cioè di un valore costante che viene sommato a tutti i valori di intensità corretti in modo da allontanarli dallo zero. Questo *offset* ha l'effetto di stabilizzare la varianza per gli spot a bassa intensità. Generalmente il valore predefinito di questa costante è 50, ma può essere modificato in funzione del maggiore o minore effetto di stabilizzazione della varianza che si vuole ottenere. In figura 4.8 si può osservare come questo metodo abbia un grosso effetto di stabilizzazione della varianza alle basse intensità.

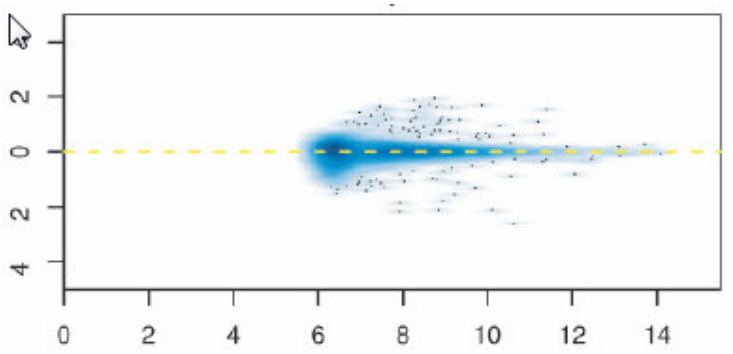


Figura 4.8: MA plot di dati a cui è stato applicato il metodo Normexp + offset

Il prezzo della notevole stabilizzazione è l'abbassamento dei "fold-change" che, tuttavia, non sembrerebbe sembra inficiare la capacità di ottenere la lista dei geni differenzialmente espressi e avrebbe sembra avere un effetto positivo anche sul contenimento dell'errore di tipo I sui dati [34].

4.4 Controllo di qualità dei dati

E' già stato illustrato l'effetto che producono alcuni metodi di sottrazione del "background" su un segnale debole. Ciò deve invitare ad utilizzare cautela nell'applicazione di questo procedimento. Non fare la correzione, tuttavia, può rendere i dati meno affidabili, dal momento che il contributo del rumore può modificare il dato relativo all'intensità dello spot generando falsi positivi o falsi negativi nella rilevazione dei geni differenzialmente espressi. L'approccio migliore consiste nel fare una correzione con un fattore che risulti da procedimenti globali, in modo da contenere al minimo gli errori introdotti da questa operazione.

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

SNR = Mediana del segnale netto / SD del rumore

dove al denominatore vi è la deviazione standard (SD) del rumore.

Molti software di analisi di microarray utilizzano il valore di SNR ricavato per ogni spot per escludere dal processo di normalizzazione quei dati che hanno un rumore troppo alto: in tal caso viene fissata una soglia per SNR, tipicamente pari a 3, in modo che agli spot con un valore di SNR più basso venga applicata una "flag", ossia un punteggio, che li esclude automaticamente dall'insieme degli spot utilizzati per la normalizzazione.

Sempre grazie all'applicazione di punteggi, che vengono assegnati in base alla rispondenza delle caratteristiche dello spot a quelle specificate dall'analista, è possibile escludere spot con forme irregolari, o con percentuale di pixel saturati superiore alla soglia definita come accettabile. Generalmente, in questo caso viene assegnato un punteggio che indica che lo spot è saturato se la percentuale dei suoi pixel, che superano una soglia differente per ciascun tipo

di scanner, è superiore al 10%, se si vuole ottenere una selezione particolarmente stringente, o al 50%, per una selezione più debole.

La selezione può escludere anche spot che vengono inseriti nel microarray esclusivamente per facilitare l’operazione di allineamento della griglia, oppure gli spot di controllo (spot che vengono lasciati appositamente vuoti, oppure controlli negativi per il “background” o ancora controlli positivi per il controllo di ibridizzazione).

Si effettua in questo modo un processo di eliminazione degli spot di bassa qualità che prepara il dato alla successiva operazione di normalizzazione.

Capitolo 5

Metodi di normalizzazione dei dati

Molte variabili possono influire e distorcere i risultati di un esperimento di microarray:

- disomogeneità del processo di deposizione delle sonde,
- quantità iniziali diverse di RNA,
- diversa efficienza del processo di retrotrascrizione dei due campioni;
- diversa efficienza di incorporazione dei due fluorocromi durante il procedimento di marcatura dei campioni,
- disomogeneità di ibridizzazione sul vetrino,
- diversa efficienza di emissione dei due fluorocromi,
- diversa efficienza dello scanner nell'eccitazione e nella lettura dei due canali di fluorescenza.

Tutti questi fattori possono influenzare pesantemente i dati causando spostamenti nelle distribuzioni dei rapporti delle intensità dei due fluorofori. E', quindi, necessaria, prima di ogni tipo di analisi statistica, una normalizzazione dei dati atta ad eliminare distorsioni sistematiche.

5.1 Normalizzazione dei dati

Un esempio di distorsione indotta dalla procedura sperimentale si può osservare nello scatterplot in figura 3.1, e riportato in figura 5.1 per comodità, in cui è mostrata un'ibridizzazione self-self.

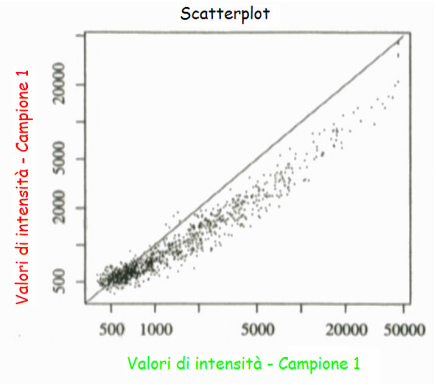
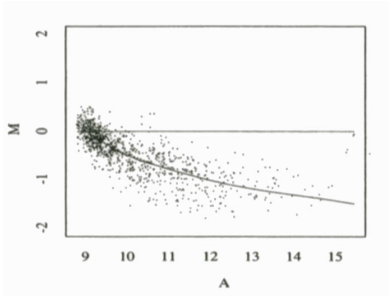


Figura 5.1: Scatterplot di due aliquote dello stesso mRNA ibridizzate su un vetrino.

Idealmente, i punti che individuano i valori di intensità sullo scatterplot si dovrebbero posizionare sulla diagonale: quando ciò non accade significa che sono presenti errori sistematici, se la deviazione della diagonale è tutta dalla stessa parte, come in questo esempio, oppure errori casuali (random), quando i punti si allontanano dalla diagonale in entrambe le direzioni.

Questo errore diventa ancora più evidente quando si visualizzano i dati su un MA plot.



A = ½ log (R*G)
M = log (R/G)

Figura 5.2: Grafico MA di un array su cui è stato ibridizzato lo stesso RNA marcato con entrambi i fluorocromi

La normalizzazione può essere di due tipi: *within-array* e *between-arrays*.

Si parla di normalizzazione *within-array* quando la tecnica scelta viene applicata ad ogni vetrino singolarmente, nell'intento di correggere gli errori sistematici su ogni array preso come unità a sé e indipendentemente dal disegno sperimentale, mentre si fa una normalizzazione *between-array*s quando si cerca di ottenere un dato confrontabile fra i differenti array considerando sia il disegno sperimentale applicato che le repliche biologiche o le eventuali repliche tecniche.

In ciascuna di queste situazioni è necessario scegliere un gruppo di geni da utilizzare per la normalizzazione. Questi possono essere:

- *tutti i geni sull'array.* L'ipotesi fondamentale che rende possibile l'applicazione di questo tipo di normalizzazione è che la maggior parte dei geni presenti sul vetrino sia non differenzialmente espressa. Questa ipotesi è automaticamente verificata su tutti quei microarray che analizzano sul vetrino l'intero trascrittoma di un organismo, dal momento che è abbastanza inverosimile che l'espressione genica sia differente, fra i due campioni, per la quasi totalità dei geni. Questo tipo di normalizzazione non è consigliabile, invece, su tutti quegli array che sono stati appositamente creati per osservare l'andamento dell'espressione genica su un numero ristretto di geni. In questo caso, infatti, è possibile che l'espressione genica sia differente, fra le due classi, su un buon numero di geni.

- *geni espressi in maniera costante.* Invece di utilizzare tutti i geni presenti sul vetrino, per la normalizzazione si può scegliere di usare un piccolo sottoinsieme rappresentato dai geni "housekeeping", cioè quei geni il cui livello di espressione non è influenzato da condizioni sperimentali differenti. Non è facile identificare questo sottoinsieme, ma spesso è possibile trovare un gruppo di geni che si comportano da "housekeeping" nelle condizioni sperimentali considerate. Una limitazione nell'utilizzo dei geni "housekeeping" è che essi tendono ad essere molto espressi e quindi potrebbero non essere rappresentativi delle intensità di altri geni di interesse. Questo tipo di normalizzazione può essere utilizzata anche nel caso di array di piccole dimensioni.

- *controlli.* Un'alternativa alla normalizzazione con geni housekeeping è l'utilizzo di controlli positivi, detti anche *spiked,* o di una serie di sequenze di controllo a concentrazione scalare *(titration).* Nel metodo dei controlli *spiked*, sequenze sintetiche di DNA o selezionate da organismi differenti da quello studiato sono depositate sull'array, mentre sequenze ad essi complementari sono aggiunte in concentrazione identica ai due campioni di mRNA in esame. Queste sequenze di controllo possono essere utilizzate per la normalizzazione perché daranno origine a segnali di uguale intensità nei due canali. Nell'approccio *titration*, si utilizzano spot dello stesso gene depositati sul vetrino in concentrazioni scalari. Anche in questo caso le sequenze complementari a questi spot vengono mescolate in uguale concentrazione ai due campioni di mRNA in esame e i segnali prodotti devono essere proporzionali alla concentrazione di sonde depositate in ciascuno spot ed uguali nei due canali. E' possibile in questo modo monitorare l'aumento lineare dell'intensità del segnale proporzionalmente alla concentrazione delle sonde e catturare, seppure in maniera discontinua, le caratteristiche non lineari delle distorsioni.

5.2 Normalizzazione within-array

In questo caso la normalizzazione viene applicata separatamente ad ogni array. Gli scopi principali sono la correzione dei problemi che si traducono in una differente intensità fra i due canali e di quelli dovuti ad un'eventuale deposizione scorretta delle sonde.

5.2.1 Normalizzazione globale

I metodi globali di normalizzazione assumono che le intensità dei due fluorocromi siano proporzionali, cioè che valga la relazione R =K*G, dove con R si indica il canale rosso, con G il canale verde e K è la costante di proporzionalità. In funzione di questa legge e nell'ipotesi che la maggior parte dei geni non si esprima differenzialmente, la normalizzazione globale sposta il logaritmo del rapporto dei due canali sullo zero, operando quello che viene tipicamente denominato "centraggio" della distribuzione dei dati:

$$\log_2 R/G \xrightarrow{\text{normalizzazione}} \log_2 R/G - c = \log_2 R/(KG)$$

dove $c = \log_2 K$.

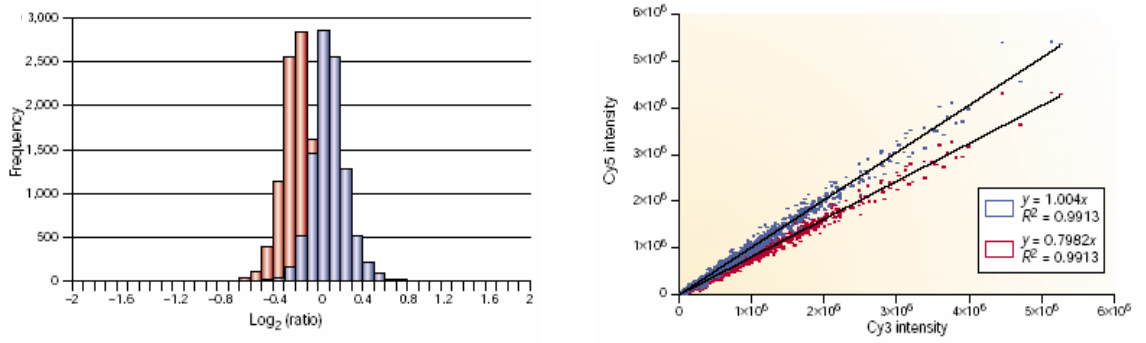


Figura 5.3: sx) Istogramma della distribuzione dei dati prima della normalizzazione (in rosso) e dopo lo spostamento della media (in blu).

dx) Scatterplot dei dati prima della normalizzazione (in rosso) e dopo la normalizzazione (in blu).

Una particolare scelta per il parametro c è la mediana o, in alternativa, la media dei rapporti logaritmici delle intensità.

I metodi di normalizzazione globale non hanno un effetto intensità-dipendente sui dati e, quindi, non riescono a correggere le tendenze non lineari dei dati dovute al diverso comportamento che i fluorocromi presentano in emissione. Questo metodo è disponibile nel pacchetto *LIMMA* con il nome di *median.*

5.2.2 Normalizzazione intensità-dipendente: *LO(W)ESS* e rlowess

In molti casi gli errori sistematici riconducibili alla diversa efficienza di emissione dei fluorocromi sono dipendenti dall'intensità del segnale, come può

essere evidenziato attraverso il grafico MA dei dati. In questi casi si possono correggere le distorsioni attraverso metodi di interpolazione di curve sui dati.

La trasformazione *LO(W)ESS* (LOcally WEighted polynomial regreSSion) [9, 41, 42], così come la sua variante *LOESS*, divide i dati sull'asse delle ascisse in intervalli sovrapposti e interpola su di essi una funzione polinomiale:

$$y = a_0 + a_1x + a_2x^2 + ...$$

I polinomi presentano la caratteristica di poter passare esattamente per tanti punti quanto è il loro grado. Tuttavia, questo approccio presenta il cosiddetto problema dell'"*over-fitting*", ossia si ha un'approssimazione quasi perfetta della funzione bersaglio nei punti conosciuti, ma oscillazioni eccessive al di fuori di essi.

Per ovviare a ciò, l'approccio *LO(W)ESS* utilizza polinomi di grado1, mentre il *LOESS* fa uso di parabole in modo da contenere l'over-fitting e l'eccessiva oscillazione fra i punti delle funzioni interpolanti.

Inoltre, poiché l'approssimazione polinomiale è precisa solo in piccoli intervalli intorno al punto scelto, può essere necessario dividere il dominio dei dati in finestre di dimensioni piccole con l'effetto collaterale di incrementare notevolmente il carico computazionale.

La divisione in piccoli intervalli ha inizio partendo dai geni meno espressi, con una finestra di larghezza data *l,* e i dati che cadono in questi intervalli sono utilizzati per interpolare il polinomio, applicando ad essi dei pesi diversi a seconda della loro posizione nell'intervallo: i dati prossimi al punto di stima hanno un peso maggiore di quelli lontani e ciò può essere realizzato utilizzando una funzione di peso w(x) della forma:

$$w(x) = \begin{cases} \left(1-|x|^3\right)^3, & |x|<1 \\ 0 & ,|x|\geq 1 \end{cases}$$

dove x è la distanza fra i punti di stima. Il procedimento continua facendo scorrere la finestra verso i geni più espressi e interpolando localmente di volta in volta un nuovo polinomio: il risultato è una curva di "smoothing" attraverso cui correggere i dati. L'effetto dell'applicazione di questi metodi ad un insieme di dati può essere osservato in figura 5.4, dove si può notare come la normalizzazione agisca sui dati avvicinandoli all'asse delle ascisse ed eliminando o, comunque, riducendo drasticamente l'andamento non lineare.

Questo metodo è disponibile nel pacchetto *LIMMA* con il nome di *LOESS.*

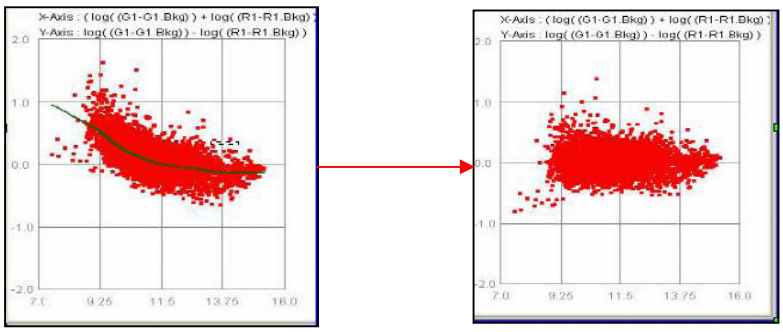


Figura 5.4: Correzione dei dati con l'applicazione di una normalizzazione LO(W)ESS.

Il vantaggio del metodo *LO(W)ESS* è che non ha bisogno di specificare una particolare funzione come modello: i soli parametri necessari sono il grado dei polinomi *d* e il fattore di smoothing *q*, che indica la larghezza della finestra.

Gli svantaggi del metodo *LO(W)ESS* includono il fatto che esso non produce una funzione di regressione o un modello che sia facilmente rappresentabile con una formula matematica. In particolare, il modello di correzione della distorsione del colore trovato su un particolare insieme di dati non può essere direttamente trasferito ad un altro: è necessario riapplicare il metodo ogni volta che si ha un insieme di dati distinto e ciò produce sottili differenze ad ogni applicazione.

Un ulteriore svantaggio è legato al fatto che la procedura è computazionalmente molto pesante, anche se questo è un problema minore nel contesto di tutte le altre problematiche collegate all'analisi dei dati da microarray.

Il più importante svantaggio è la suscettibilità di questo metodo al rumore e agli "outlier" e l'unico modo per contenerla è l'esclusione di spot rumorosi o anomali prima della normalizzazione attraverso l'applicazione di indicatori di qualità.

Il metodo *rlowess* normalizza i dati in modo simile alla *LO(W)ESS*, ma la correzione tiene conto, oltre che dell'intensità dello spot, anche della sua posizione sull'array. E' così possibile mitigare gli artefatti dovuti, per esempio, ad un processo di deposizione delle sonde su vetro malcondotto o ad un'ibridizzazione non uniforme della miscela.

La parte di normalizzazione spaziale viene realizzata interpolando una funzione a due dimensioni che segue l'andamento delle intensità sull'array. Tale funzione, al pari della curva di "smoothing", viene utilizzata per ricalibrare i valori di intensità ed eliminare gli artefatti spaziali.

5.2.3 Trasformazione lineare-logaritmica o *lin-log*

Seppure la trasformazione lineare-logaritmica [26] non sia considerata una vera e propria normalizzazione dei dati, essa produce una correzione dei dati.

Questa trasformazione risulta particolarmente utile per mitigare la destabilizzazione della varianza alle basse intensità prodotta dai metodi di sottrazione del background. Come precedentemente osservato nel capitolo 4, l'intensità del rumore ha un peso differente a seconda dell'intensità del segnale dello spot. In particolare la componente di rumore può essere considerata additiva alle basse intensità e moltiplicativa per le medie ed alte intensità. Per questo motivo un'appropriata trasformazione dei dati alle basse intensità sarà lineare, mentre dovrà essere logaritmica per le altre intensità.

Se si definisce con d_i la soglia di intensità che delimita le due zone, la trasformazione *lin-log* può essere descritta come:

$$Z_{ik} = \begin{cases} \log_2(d_i) - 1/\ln 2 + Y_{ik}/(d_i \times \ln 2) & Y_{ik} < d_i \\ \log_2(Y_{ik}) & Y_{ik} \geq d_i \end{cases}$$

dove Y e Z sono rispettivamente il valore grezzo e trasformato dell'intensità, i e k individuano rispettivamente il fluoroforo e lo spot. E' stato stimato che un buon valore della soglia d_i posiziona il 25-30% dei dati nell'intervallo di intensità che verranno trasformate linearmente [26].

5.2.4 Correzione "paired-slide" o "self-normalization"

La correzione "paired-slide", come la *lin-log*, non è una vera tecnica di normalizzazione, ma produce comunque una correzione dei dati in esperimenti nei quali sia stato adottato un disegno sperimentale in "dye-swap", oppure quando si stanno utilizzando array di piccole dimensioni [41, 42].

Si denoti con $\log_2$ (R/G) - c il rapporto logaritmico normalizzato fra i due canali per il primo vetrino, con $\log_2$ (R'/G') - c' quello per il secondo, mentre c e c' siano le due funzioni di normalizzazione per i due vetrini. Nell'ipotesi che non vi siano comportamenti disuguali dei due fluorocromi sui due vetrini, deve valere che c ≅ c' e che $\log_2$R/G= $\log_2$G'/R',cioè:

$$\frac{1}{2}\left[\log_2 \text{R/G - c - }\left(\log_2 \text{R'/G'-c'}\right)\right] \cong \frac{1}{2}\left[\log_2 \text{R/G} + \log_2 \text{G'/R'}\right] = \frac{1}{2}\log_2 \text{RG'/GR'} = \frac{1}{2}\left(M - M'\right)$$

Con questo metodo è possibile scalare i livelli di espressione relativa per i due microarray senza esplicitare la normalizzazione: questo procedimento prende il nome di "self-normalization".

La validità di questa assunzione può essere verificata utilizzando un insieme di geni con livelli di espressione costante sui due canali. Poiché l'assegnazione dei fluorocromi è invertita sui due microarray, ci si attende che su tali geni il rapporto logaritmico normalizzato sui due vetrini abbia uguale intensità ma segno opposto:

$\log_2$R/G-c ≅ -($\log_2$R'/G'-c')

Quindi, riarrangiando l'equazione e assumendo ancora che c ≅ c' è possibile stimare la funzione di normalizzazione c come:

$$\text{c} \cong \frac{1}{2}\left[\log_2 \text{R/G} + \log_2 \text{R'/G'}\right] = \frac{1}{2}(M + M')$$

Da un punto di vista operativo, la funzione c = c(A) di correzione su tutto il vetrino è stimata attraverso l'interpolazione *LO(W)ESS* di $\frac{1}{2}(M + M')$ vs $\frac{1}{2}(A + A')$.

Se dal disegno dell'esperimento si evidenzia la presenza di un "dye-swap", questa normalizzazione è automaticamente applicata nel pacchetto *LIMMA* e l'effetto "dye" e la successiva correzione dei dati vengono valutati durante l'analisi statistica dei dati.

5.3 Normalizzazione "multiple-slides" o "between arrays"

Dopo la normalizzazione "within array" le distribuzioni dei dati normalizzati di ogni microarray preso singolarmente saranno centrate sulla propria media o mediana. I metodi di normalizzazione "multiple-slides", il cui scopo è quello di consentire confronti fra diversi array, servono per operare una scalatura fra i dati acquisiti con differenti vetrini quando i loro rapporti logaritmici normalizzati presentano una dispersione differente. Non effettuare una normalizzazione tra array quando essa è necessaria può generare un peso non reale sui dati di un particolare vetrino quando si effettua il confronto fra microarray.

E' importante notare, tuttavia, che questo tipo di normalizzazione ha senso esclusivamente fra campioni identici (repliche sperimentali) o, comunque, molto simili (repliche biologiche); per esempio, effettuare la normalizzazione tra array fra dati provenienti da un tessuto sano e uno malato è un errore poiché si sta involontariamente cercando di indurre un'omogeneizzazione delle dispersioni su campioni appartenenti a classi differenti, introducendo un artefatto.

5.3.1 Normalizzazione *scale*

La necessità di normalizzare *between-arrays* è determinata dalla visualizzazione dei boxplot dei dati.

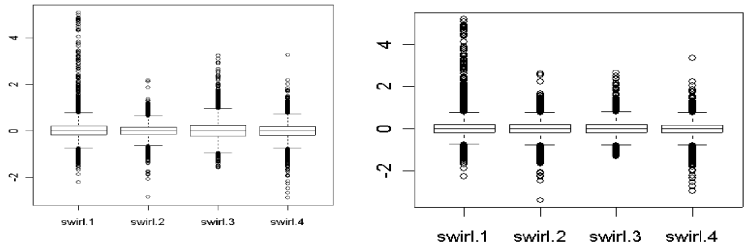


Figura 5.5: Box-plot degli array prima (sx) e dopo (dx) la normalizzazione between-arrays

Nella figura 5.5 (sx) sono visualizzati i dati di un esperimento microarray realizzato con un confronto diretto fra i campioni delle due classi e con inversione della marcatura. I dati sono già stati normalizzati con il metodo *LOESS*, ciò ha prodotto l'eliminazione dell'errore intensità-dipendente e il centraggio della distribuzione sullo zero. Tuttavia, si nota ancora che la dispersione dei dati centrati sullo zero è differente fra i quattro array (l'ampiezza della scatola è differente).

In questo caso è utile applicare una riscalatura dei dati fra array in modo da rendere omogenei i valori di M fra array (figura 5.5 dx).

Un metodo per realizzare la normalizzazione *scale* [41] può essere quello di assumere che tutti i rapporti logaritmici relativi all'i-esimo array si distribuiscano secondo una distribuzione normale con media nulla e varianza $a_i^2\sigma^2$, dove σ^2 è la varianza dei rapporti logaritmici ed a_i^2 è il fattore di scala per l'i-esimo array.

Per ottenere una stima di questo fattore di scala, si procede massimizzando la funzione di verosimiglianza del parametro a_i^2 sui dati ricavando la seguente stima:

$$\hat{a}_i = \frac{MAD_i}{\sqrt[I]{\prod_{i=1}^{I} MAD_i}}$$

dove MAD_i è la Median Absolute Deviation sull'array i ed è definita come: $MAD_i = median_j\{|M_{ij} - median_j(M_{ij})|\}$, con i=1,…, I (I = numero di array) e j = 1,…,N (N = numero totale di spot su ciascun array).

5.3.2 Normalizzazione *quantile* e *Aquantile*

La normalizzazione *quantile* è stata proposta per la prima volta per la normalizzazione degli array sui quali viene utilizzato il protocollo di ibridizzazione one-color [43] ed è stata poi modificata per i microarray two-color [42].

Il metodo di normalizzazione *quantile* modifica i valori delle singole intensità in modo che essi abbiano la stessa distribuzione su entrambi i canali di tutti gli array. Infatti, la normalizzazione *quantile* non distingue fra i due canali, ma cerca di renderli confrontabili fra di loro e su tutti gli array, in modo da eliminare eventuali sbilanciamenti per ogni quantile della distribuzione delle intensità. Essa può essere applicata a entrambi i canali, oppure ad uno solo dei due.

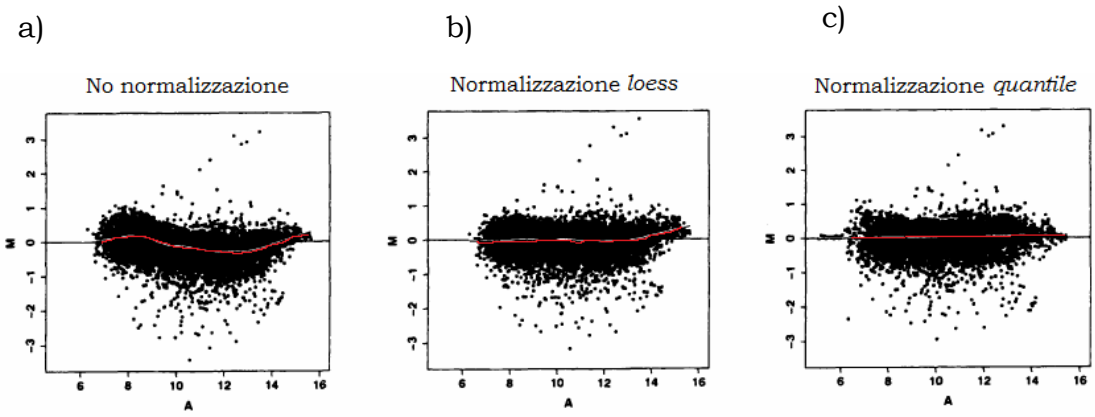


Figura 5.6: MA-plot di dati: a) non normalizzati,, b) normalizzati within-array e c) normalizzati between-arrays.

Questo metodo di normalizzazione è particolarmente potente nell'eliminare eventuali residui di effetti non lineari intensità-dipendenti ancora presenti sui dati. Come esempio si può osservare la figura 5.6 a): è evidente in questo MA plot che alle alte intensità la normalizzazione *LOESS* non è riuscita ad eliminare la non linearità dalla distribuzione dei dati, mentre in figura 5.6 c) anche gli spot ad alta intensità di segnale risultano allineati con l'asse delle ascisse.

I benefici della normalizzazione *quantile* sono realizzati al costo di una riduzione dei valori d'intensità, e di conseguenza dei valori di M, proprio nelle code della distribuzione dei logaritmi dei rapporti, dove è più verosimile trovare una differenza di espressione genica.

Un metodo molto meno efficace, ma anche meno costoso, è la normalizzazione *Aquantile*. In questo caso, infatti, il metodo produce un bilanciamento dei valori di abbondanza di segnale (A), senza modificare i valori di M. Generalmente questo metodo fornisce un risultato che si differenzia molto poco dall'applicazione del metodo di normalizzazione *scale* o addirittura da quello ottenuto con il metodo *LOESS*.

Capitolo 6

Automatizzazione del pre-trattamento dei dati: il software Feature Extraction®

L'avanzamento degli algoritmi di elaborazione delle immagini e di estrazione dei dati ha consentito di rendere sempre più automatizzati e veloci i processi di "gridding", estrazione e pre-trattamento dei dati.

Ne è un esempio il software Feature Extraction® (Agilent Technologies, Santa Clara, US). Feature Extraction® sfrutta cinque algoritmi per il "gridding", il "flagging", la valutazione del rumore dell'immagine e la valutazione degli errori sistematici dei dati. Questi algoritmi sono stati appositamente creati per controllare ed eventualmente correggere le problematiche che si possono avere utilizzando il sistema Agilent dall'inizio alla fine per la realizzazione di un esperimento di espressione genica.

6.1 Algoritmo "FindSpots and SpotAnalysis"

Questo algoritmo posiziona automaticamente la griglia grazie al riconoscimento degli spot posizionati ai quattro angoli del vetrino. La procedura di "gridding" è costituita essenzialmente da due passaggi:

- riconoscimento degli spot ad alta intensità di segnale posizionati agli angoli del vetrino e denominati "Bright Corner"
- riscalatura della posizione dei centroidi degli altri spot sfruttando le informazioni sulla distanza reciproca dei "Bright Corner" e quelle riguardanti lo specifico disegno adottato dal costruttore per la realizzazione del vetrino.

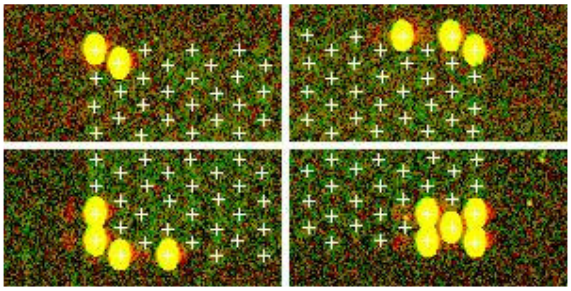


Figura 6.1: "Bright Corner" posizionati agli angoli dei microarray prodotti da Agilent

Se la posizione del centroide di ciascuno spot eccede i limiti di tolleranza impostati sulla sua posizione nominale oppure lo spot ha un segnale troppo debole, ad esso viene associato un indicatore (Flag) che segnala che lo spot non è stato trovato (Not Found).

Successivamente il software determina le aree all'interno delle quali verranno valutate le intensità associate al "foreground" e al "bakground". Per realizzare questo passaggio sono stati appositamente creati due algoritmi denominati "Cookie Cutter" e "Whole Spot".

6.1.1 "Cookie Cutter"

Questo metodo posiziona un cerchietto (in nero nella figura 6.2 sx) intorno al centroide di ciascuno spot e determina tre regioni: il "cookie" (in verde nella figura 6.2 sx), la zona di esclusione e l'anello all'interno del quale verrà determinata l'intensità da associare al "background" (delimitata dalle linee tratteggiate rossa e azzurra nella figura 6.2 sx).

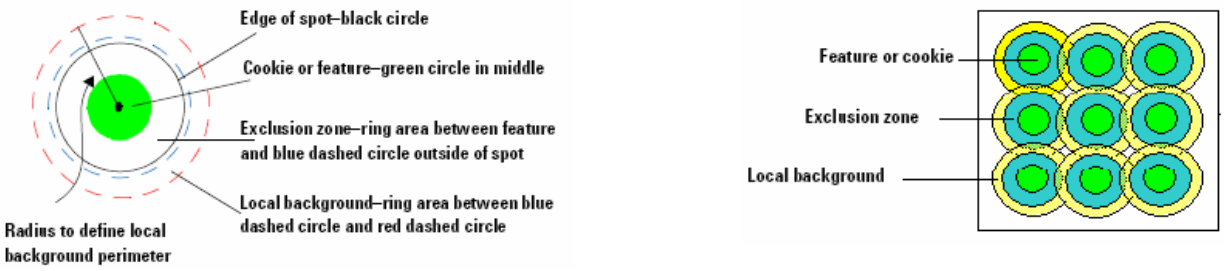


Figura 6.2: Metodo "Cookie Cutter" per la definizione delle aree per il calcolo delle intensità (sx) e sovrapposizione di "cookie" adiacenti (dx)

6.1.2 "Whole Spot"

In questo caso i profili che delimitano le tre zone precedentemente illustrate vengono determinati sulla base di un modello statistico che cerca di modellare la distribuzione del rumore sul vetrino. Poiché tale distribuzione non è la stessa intorno ad ogni spot, i profili delle tre zone saranno frastagliati e differiranno da spot a spot.

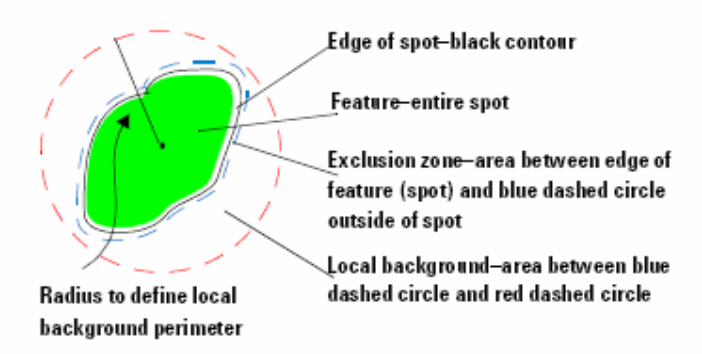


Figura 6.3: Metodo "Whole Spot" per la definizione delle aree per il calcolo delle intensità

Le impostazioni che determinano il raggio della zona dedicata alla valutazione del backgroung possono essere modificate al fine di avere una quantizzazione spot-specifica (figura 6.4 sx) o area-specifica (figura 6.4 dx) del valore di intensità.

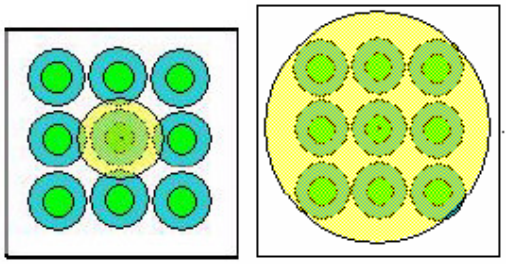


Figura 6.4: Selezione del raggio per il calcolo dell'intensità di "background": spot-specifica (sx) e area-specifica (dx).

6.2 Algoritmo "PolyOutlierFlagger"

Questo algoritmo determina se segnali di intensità associati al "foreground" o al "background" sono anomali. L'eventuale deviazione dalla norma viene valutata in base al confronto fra la deviazione standard rilevata sul segnale e quella teorica che produce il sistema Agilent.

Un segnale può essere definito "outlier" anche in base alla ripetibilità che uno spot evidenzia fra le sue copie presenti sul vetrino. In questo caso viene preliminarmente prodotta la distribuzione statistica delle intensità di tutti gli spot per ciascun canale, in modo da determinare le soglie per definire se un segnale è anomalo (figura 6.5). Queste soglie sono proporzionali al Range InterQuartile (IQR nella figura) secondo l'equazione:

$Cutoff_{PopOutlier}$= 1.42 x IQR

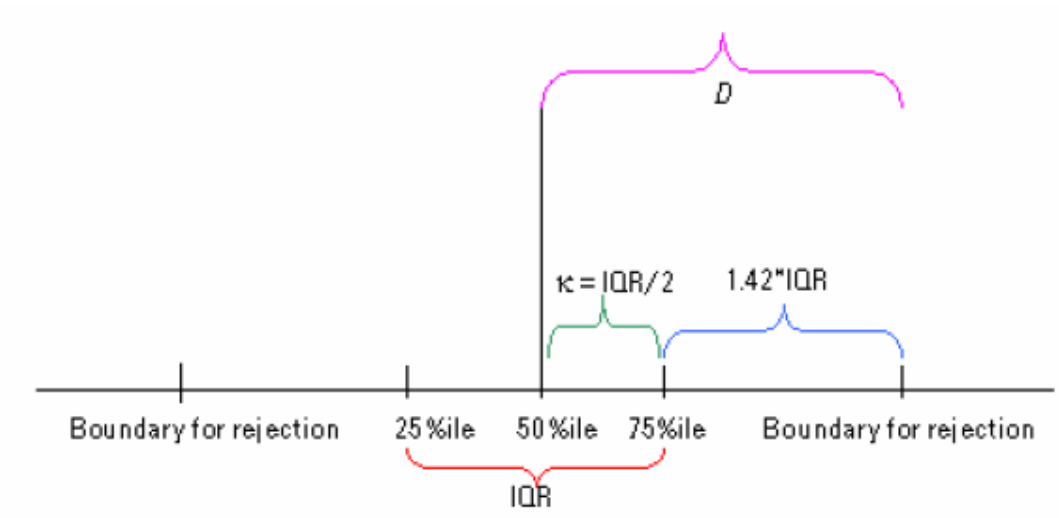


Figura 6.5: Metodo Definizione dei limiti di rigetto per la selezione degli spot "outlier".

Il valore 1.42 che moltiplica il Range InterQuartile è determinato da Agilent in modo da escludere come anomali l'1% dei segnali.

Lo stesso algoritmo viene applicato per determinare se ci sono "outlier" nella distribuzione dei pixel di uno spot. In questo caso i pixel che si posizionano oltre 1.42*IQR vengono esclusi dai dati utilizzati per determinare le statistiche d'intensità per ciascuno spot. sia per il "foreground" che per il "background".

6.3 Algoritmo "BGSubtractor"

Questo algoritmo calcola il segnale d'intensità netto, cioè a "background" sottratto, da associare a ciascuno spot.

Oltre alla valutazione spot-specifica o area-specifica del "background", Feature Extraction® fornisce anche una valutazione del livello di rumore dell'intero vetrino grazie all'algoritmo di "spatial detrend".

Grazie all'uso di un filtro di segnale a media mobile l'algoritmo di "spatial detrend" costruisce una mappa bidimensionale degli spot a bassa intensità su tutto il vetrino.. Il filtro si muove su tutta la superficie del vetrino e cerca di identificare gruppi di spot a bassa intensità per ciascuna finestra di dati. La mappa può essere in seguito utilizzata per correggere i dati di intensità del "foreground" ed ottenere un valore di intensità che sia mediamente descrittivo della superficie e, quindi, del "background" su tutto l'array. Se il valore di intensità rilevato prima dell'interpolazione della superficie risulta molto diverso, in termini di distanza IQR, dal valore ottenuto dopo l'interpolazione, l'algoritmo di "spatial detrend" associa allo spot anche un indice che lo identifica come "outlier".

L'algoritmo "BGSubtractor" calcola anche un insieme di indicatori di qualità che possono essere utilizzati per determinare la qualità di ciascuno spot per quel che riguarda il rapporto segnale/rumore.

Il primo di questi indicatori è denominato "IsPosAndSignif" ed associa un valore pari ad 1 allo spot se il t-test che confronta le intensità del "background" e quella del "foreground" produce un risultato statisticamente significativo.

Un altro indicatore di qualità è "IsWellAboveBG", che riprende la definizione classica di SNR, cioè determina se il segnale di intensità, al quale è stato sottratto il "background", supera la deviazione standard del "background" di un fattore stabilito da Agilent e pari a 2.6. In questo caso, per associare un indicatore pari ad 1 allo spot, quest'ultimo deve anche aver ottenuto 1 nella valutazione dell'indicatore "IsPosAndSignif".

6.4 Algoritmo "Dye Normalization"

Questo algoritmo si occupa di valutare ed eliminare la variabilità sistematica dei dati associata al sistema di rivelazione del segnale (sistema di marcatura, differenze fisico/chimiche fra i due fluorocromi, scanner).

Feature Extraction® consente di selezionare differenti insiemi di spot da utilizzare per la normalizzazione: tutti i geni sul vetrino che superano le soglie di qualità impostate sugli indicatori, oppure un ristretto numero di spot che vengono identificati dall'utente come "housekeeping" e, infine, tutti quei geni che sono vicini alla tendenza centrale dei dati (media o mediana).

Per l'automatizzazione del processo di selezione di quest'ultimo insieme di spot, definiti anche "housekeeping" virtuali, Agilent ha appositamente messo a punto il "Rank consistency filter".

Questi spot sono selezionati in base ai seguenti criteri:

- il loro indicatore "IsPosAndSignif" è pari a 1;
- non sono "outlier";
- non sono spot di controllo;
- non sono spot saturati;
- superano la soglia impostata per il "Rank consistency filter".

Il filtro valuta quanto i ranghi associato all'intensità di un canale siano correlati a quelli dell'altro canale per tutti gli spot selezionati dai primi quattro punti dell'elenco sopra. Il segnale netto per ciascun canale e per ciascuno spot viene trasformato in rango e viene valutata la correlazione fra i due ranghi (Correlation Strength) utilizzando la seguente formula:

$$CS = \frac{\left|\rho_R - \rho_G\right|}{N}$$

dove ρ indica il rango per ciascun canale e N è il numero di spot che superano la selezione generata dai primi quattro punti dell'elenco sopra.

Lo spot è definito "consistente in rango" se il valore di CS valutato per i suoi segnali di intensità supera una soglia τ imposta da Agilent sulla distribuzione dei ranghi e non nota o modificabile da parte dell'operatore.

Una volta che l'utente ha scelto quale insieme di spot utilizzare per la normalizzazione, la successiva fase di correzione dei dati avviene a seconda del metodo di normalizzazione selezionato.

Agilent mette a disposizione un metodo di normalizzazione globale, un metodo non lineare, che fa uso dell'interpolazione *LO(W)ESS* già illustrato e un metodo misto fra i due precedenti. Non sono invece previsti metodi di normalizzazione "between arrays".

Capitolo 7

Metodi di analisi statistica dei dati

I metodi di analisi statistica dei dati di espressione genica interessano un ampio settore della bioinformatica. I primi metodi per identificare la lista dei geni differenzialmente espressi applicavano una soglia empirica alla distribuzione dei log-"fold-change" senza produrre una valutazione statistica degli errori commessi nel dichiarare che un gene fosse differenzialmente espresso. Un miglioramento in questo senso è stato realizzato da Tusher e colleghi [20]con l'Analisi della Significatività statistica dei Microarray, che assegna l'espressione differenziale ad un gene sulla base di una permutazione dei dati disponibili e propone una misura di False Discovery Rate (FDR) come parametro statistico di confidenza nella risposta ottenuta.

Di tutt'altro genere sono i metodi empirici bayesiani che, insieme allo strumento messo a punto da Tusher, si dimostrano particolarmente efficienti quando si vuole realizzare un test multiplo simultaneo di ipotesi su un insieme numeroso di soggetti, per ognuno dei quali sono disponibili poche osservazioni.

Questo è proprio il caso degli esperimenti microarray, in cui il numero di osservazioni per lo stesso gene è generalmente molto basso, rispetto al numero totale di geni analizzati.

I metodi bayesiani fanno inferenza su ogni singolo gene, ossia generano stime dei parametri statistici che lo possono descrivere, traendole da tutto l'insieme di dati: per questo motivo tali metodi vengono detti empirici. Un contesto operativo o "framework" bayesiano genera queste stime avvalendosi del teorema di Bayes, di ipotesi sulle distribuzioni a priori dei parametri, formulate dall'analista, e dell'insieme dei dati: sono queste le componenti di un processo capace di produrre in maniera automatica un aggiornamento dei parametri delle distribuzioni coinvolte, al fine di generare gli elementi confrontati nella statistica caratteristica del "framework".

Ancora differenti sono i metodi che si propongono di misurare le fonti di variabilità che si abbattono sui dati microarray e cercano di quantificare la varianza "spiegata" da tali sorgenti in relazione alla varianza totale dell'insieme di dati, nel tentativo di eliminarla.

Questo obiettivo può essere ottenuto disegnando in maniera opportuna l'esperimento, in modo da raccogliere un'adeguata, ma spesso proibitiva, quantità di misure e contribuire a monitorare gli effetti di alcune delle fonti di variabilità. Ad un disegno sperimentale idoneo si deve aggiungere la capacità di realizzare un modello descrittivo dei dati, che renda possibile la diversificazione dell'effetto obiettivo dello studio da quelli legati a fonti di variabilità indesiderate.

Kerr e Churchill [24, 44, 45] sono stati i primi a studiare le potenziali sorgenti di variabilità in esperimenti microarray e ad incorporarle in un modello additivo attraverso il metodo statistico di analisi della varianza a più fattori ANOVA (ANalysis Of VAriance).

A ciascuno dei tre metodi brevemente illustrati corrisponde in Bioconductor un pacchetto che raccoglie tutte le funzioni necessarie a realizzare l'analisi dei

dati di espressione genica seguendo i fondamenti statistici da essi proposti. Questi pacchetti sono denominati *SAM* (Significance Analysis of Microarrays) [20], *LIMMA* (LInear Model of MicroArray data) [21] e *MAANOVA* (MicroArray ANalysis Of VAriance) [22].

7.1 Analisi della significatività sui microarray

La presenza di rumore e di numerosi fattori di variabilità dei dati, non sempre ben quantificabili ed eliminabili, rende necessaria l'adozione di approcci statistici per la selezione dei geni differenzialmente espressi.

In questo contesto, il metodo che va sotto il nome di analisi della significatività statistica conferisce una solida base statistica ad un criterio di selezione a soglia.

Nell'Analisi della Significatività statistica per Microarray (*SAM*) [20] viene assegnato un punteggio ad ogni gene sulla base di un procedimento iterativo.

Capita spesso che un metodo di analisi statistica utilizzi tecniche di "bootstrap" o permutazioni, cioè un campionamento dei dati con o senza sostituzione delle osservazioni, per creare degli insiemi di dati surrogati a partire dai quali effettuare le necessarie speculazioni statistiche; questi approcci hanno tanto più valore quanto minore è il numero delle informazioni o dei dati disponibili.

SAM si basa su una variante del t-test, in cui vengono messe a confronto le medie delle due distribuzioni di dati che si vogliono paragonare, attraverso la verifica di due ipotesi: l'ipotesi nulla, secondo cui i due campioni di dati provengono dalla stessa popolazione, e l'ipotesi alternativa, che afferma che i dati appartengono a due popolazioni distinte.

Dal punto di vista dell'espressione differenziale dei geni queste ipotesi si possono definire come:

- *Ipotesi nulla:* i due valori di espressione che si stanno confrontando indicano che il gene non è differenzialmente espresso;
- *Ipotesi alternativa:* i due valori di espressione indicano che il gene è differenzialmente espresso.

Di seguito è riportato il diagramma delle operazioni eseguite per visualizzare meglio il procedimento di calcolo della statistica utilizzata da *SAM*:

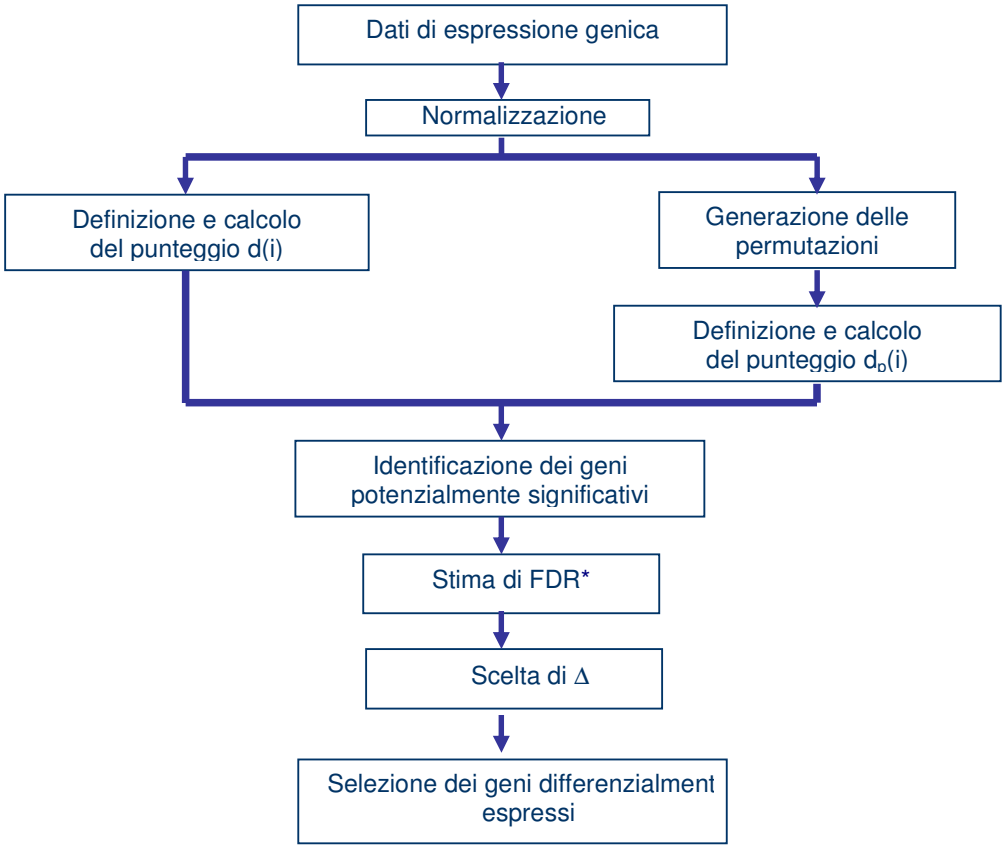


Figura 7.1: Diagramma delle operazioni effettuate nell'analisi della significatività statistica.

**FDR (False Discovery Rate)*

SAM non mette a disposizione metodi di sottrazione del background o metodi di normalizzazione dei dati, per cui è necessario effettuare queste operazioni con un altro pacchetto prima di iniziare il calcolo dei punteggi e fornire a *SAM* un insieme di dati opportunamente ripulito.

Il primo passo calcola il punteggio *d(i)* per ogni gene *i* nelle due condizioni che si vogliono verificare

dove:

$$d(i) = \frac{\bar{x}_{C1}(i) - \bar{x}_{C2}(i)}{s(i) + s_0}$$

- al numeratore vi è la differenza fra le medie delle misure relative alle due condizioni C_1 e C_2 per il gene i-esimo (possono essere, per esempio, trattato e controllo);
- al denominatore vi è la somma fra la stima della deviazione standard del numeratore e un valore additivo s_0, detto "*fudge factor*".

La deviazione standard del numeratore può essere calcolata in base alla seguente formula di stima:

$$s(i) = \sqrt{\left(\frac{\frac{1}{n_1} + \frac{1}{n_2}}{n_1 + n_2 - 2} \right) \left\{ \sum_{h=1}^{n_1} [x_h(i) - \bar{x}_{C_1}(i)]^2 + \sum_{k=1}^{n_2} [x_k(i) - \bar{x}_{C_2}(i)]^2 \right\}}$$

dove:

- le due sommatorie sono estese al numero di misure effettuate nei due stati;
- n_1 è il numero di misure nello stato C_1;
- n_2 è il numero di misure nello stato C_2.

A bassi livelli di espressione la varianza di *d(i)* può essere alta a causa di piccoli valori di *s(i)*. Per assicurare l'indipendenza della distribuzione dei *d(i)* dal livello di espressione del gene è necessario aggiungere un fattore additivo s_o al denominatore del punteggio. Il valore di s_0 viene scelto in modo da minimizzare il coefficiente di variazione di *d(i)* in funzione di *s(i)* attraverso un procedimento a finestre mobili sui dati. In generale si sceglie s_0 in maniera che il coefficiente di variazione di *d(i)* sia approssimativamente costante al variare di *s(i).*

L'acquisizione di una stima di confidenza, espressa sottoforma di FDR, sui dati richiede la realizzazione di numerosi esperimenti al fine di ottenere un'informazione il più possibile completa sui livelli di espressione di tutti i geni. Poiché eseguire molti esperimenti è dispendioso sia in termini di tempo che economici, vengono effettuate una serie di permutazioni dei dati, ognuna delle quali produce un nuovo valore del punteggio *d(i);* tali permutazioni devono essere bilanciate. Una permutazione è bilanciata se per ogni gruppo di *g* esperimenti, con *g* pari al numero di campioni che sono stati ibridizzati, vi sono $g/2$ esperimenti per campione.

Per stimare l'ordine delle statistiche *d(i)*, vengono calcolati per ogni permutazione *p* i punteggi $d_p(i)$ da attribuire al gene *i* di ogni coppia di esperimenti, secondo la definizione:

$$d_p(i) = \frac{\bar{x}_{G1}(i) - \bar{x}_{G2}(i)}{s(i) + s_0}$$

dove con G_i si indicano i due gruppi della permutazione, ossia le due condizioni sperimentali.

I punteggi così ottenuti sono ordinati in senso ascendente:

$$d_p(1) \geq d_p(2) \geq d_p(3) \geq \ldots \geq d_p(k)$$

dove k indica la posizione del punteggio all'interno dell'insieme ordinato dei $d_p(i)$.

Si definisce la differenza relativa attesa sul numero di permutazioni come:

$$d_E(k) = \sum_{p=1}^{n_p} \frac{d_p(k)}{n_p}$$

come nell'esempio indicato in figura 7.2.

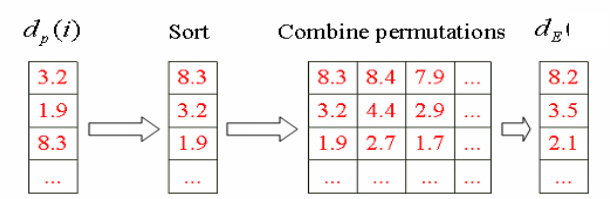


Figura 7.2: Punteggi di permutazione e punteggio atteso su tutte le permutazioni

Per identificare i geni significativamente espressi si ordinano i punteggi dei dati originali in senso ascendente e si indica con $d^i(k)$ il punteggio $d(i)$ del gene che era in posizione i-esima e dopo l'ordinamento si trova in posizione k-esima.

Per mettere in relazione i punteggi $d^i(k)$ con le differenze relative attese $d_E(k)$ si fa uno scatterplot che prende il nome di "*SAM* plot".

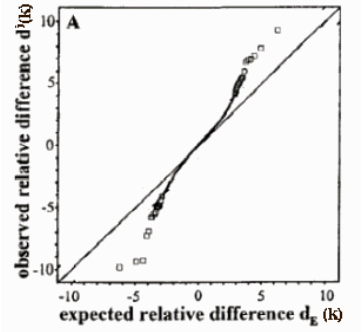


Figura 7.3: "SAM plot"

Dalla figura 7.3 si può osservare che per diversi geni si ha che $d^i(k) \cong d_E(k)$.

Una volta stabilita una soglia Δ si individuano il più piccolo $d(i)$ positivo (t_1) e il più grande $d(i)$ negativo (t_2) tali che:

$$\left| d^i(k) - d_E(k) \right| \geq \Delta$$

e il gene i-esimo viene definito potenzialmente differenzialmente espresso se vale che $d^i(k) \geq t_1$ o $d^i(k) \leq t_2$

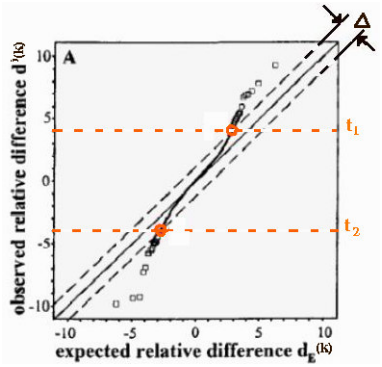


Figura 7.4: SAM plot con apposizione della soglia superiore t_1 e della soglia inferiore t_2

Per dare una valutazione statistica dell'affidabilità con cui si è individuato l'insieme di geni differenzialmente espressi si stima il False Discovery Rate (FDR):

$$FDR \approx \frac{\frac{1}{n_p}\sum_{p=1}^{n_p} card\{i \mid d_p(i) \geq t_1 \vee d_p(i) \leq t_2\}}{card\{i \mid d(i) \geq t_1 \vee d(i) \leq t_2\}}$$

dove, fissati i valori di soglia t_1 e t_2 , al numeratore si ha la media del numero di geni individuati come differenzialmente espressi attraverso le permutazioni e al denominatore il numero di geni differenzialmente espressi ottenuti dall'analisi dei dati reali.

I geni differenzialmente espressi prodotti da una permutazione p sono detti "falsamente significativamente espressi" e sono individuati con la stessa procedura con cui si selezionano i geni significativamente espressi, ma sostituendo $d_i(k)$ con $d_p(k)$, ossia:

$$\left| d_p\ (k) - d_E(k) \right| < \Delta$$

Ovviamente questa stima è differente a seconda della soglia impostata, per cui è possibile determinare il valore di Δ a seconda del FDR che si desidera avere sui dati. Come si può osservare nella tabella 7.1, ad un minore FDR corrispondono Δ maggiori e, come immaginabile, diminuisce il numero di geni differenzialmente espressi individuati.

Parameter	Number falsely significant	Number called significant	FDR
SAM			
Δ = 0.4	134.9	288	47%
Δ = 0.5	78.1	192	41%
Δ = 0.6	56.1	162	35%
Δ = 0.9	19.1	80	24%
Δ = 1.2	8.4	46	18%

Tabella 7.1: Lista dei geni differenzialmente espressi in funzione di FDR e Δ

7.2 Inferenza statistica classica e approccio bayesiano empirico

Un processo di inferenza statistica su un insieme di dati propone un modello delle osservazioni collezionate e lo verifica attraverso l'analisi dei dati stessi, producendo delle stime dei parametri descrittivi della distribuzione dei dati ipotizzata.

Nell'approccio bayesiano la probabilità associata a tali stime può essere interpretata come l'aspettativa che ciascuno esprime sulla possibilità che "verosimilmente" si possa ottenere una determinata realizzazione statistica di un qualche processo. Ciò significa che individui differenti possono attribuire una diversa probabilità ad uno stesso evento.

Il concetto di probabilità assume, dunque, il significato che gli viene attribuito nel linguaggio comune, ossia è una misura del "grado di fiducia" nel verificarsi dell'evento; Conseguenza diretta di tale interpretazione è che si respinge il fondamento che vi sia un processo di generazione dei dati verificabile attraverso un procedimento dicotomico quale il test delle ipotesi.

Il concetto intuitivo sfruttato nell'approccio bayesiano è che la probabilità dipende dallo stato di conoscenza (o di ignoranza) del fenomeno in esame; questa conoscenza è, in genere, differente da persona a persona.

Tali concetti sono in conflitto con l'impostazione classica della statistica, secondo cui le proprietà probabilistiche sono definite come proprietà asintotiche, ossia legate ad un numero infinito di dati ottenibili solo attraverso esperienze replicabili, e l'inferenza sui parametri è effettuata escludendo l'eventualità di utilizzare informazioni pregresse sul fenomeno che si sta analizzando.

Lo schema operativo dell'approccio Bayesiano alla modellazione dei dati ha la seguente struttura:

Passo 1: Fare inferenza basata su tutte le informazioni a disposizione e generare un'ipotesi di distribuzione a priori per il parametro che si sta considerando.

Passo 2: Aggiornare le stime avvalendosi del teorema di Bayes [46] e delle distribuzioni a priori, al fine di generare una distribuzione a posteriori del parametro.

Passo 3: Verificare che i nuovi dati confermino le ipotesi a priori.

Il metodo bayesiano è iterativo, cioè le stime dei parametri generate al passo precedente sono gli ingressi del passo successivo e il processo si interrompe quando non si osservano apprezzabili variazioni sui parametri stimati. La differenza operativa più evidente fra un approccio classico e uno bayesiano è che mentre la statistica classica considera i dati D come realizzazioni di variabili aleatorie ed i parametri ignoti θ come deterministici, la statistica bayesiana considera i dati come costanti ed i parametri ignoti sono variabili aleatorie caratterizzate da una funzione densità di probabilità *a priori* $P(\theta)$.

7.2.1 Scelta della distribuzione *a priori* e stimatori della media e della varianza

La scelta della distribuzione *a priori* per i parametri che devono essere stimati può essere effettuata sulla base del significato che si vuole attribuire ad essa e delle diverse informazioni a disposizione dell'analista.

In generale esistono tre criteri per procedere a tale scelta e ognuno di essi esprime un differente modo di intendere questa distribuzione:

Metodo bayesiano classico: assume che la *a priori* non deve esprimere l'influenza del ricercatore, per cui si scelgono *a priori* che siano il meno informative possibile sull'insieme di dati.

Metodo bayesiano parametrico moderno: assume che la scelta della *a priori* deve essere funzionale ad avere un processo computazionale più snello, per cui si scelgono *a priori* con proprietà convenienti.

Metodo bayesiano soggettivo: assume che la *a priori* è un riassunto delle assunzioni del ricercatore, per cui si sceglie una *a priori* basata su conoscenze precedenti (risultati di precedenti studi, opinioni di altri gruppi di studio).

Nel seguito verrà illustrato soltanto il metodo bayesiano parametrico moderno, che è alla base dell'analisi statistica realizzata utilizzando il pacchetto *LIMMA*.

7.2.2 Metodo bayesiano parametrico moderno per la scelta delle distribuzioni *a priori*

L'ipotesi sui dati afferma che essi si distribuiscono seguendo una variabile aleatoria normale con media μ e varianza σ^2, entrambe variabili casuali sconosciute.

In questo approccio, la scelta delle distribuzioni *a priori* per i parametri viene effettuata in base a considerazioni di ordine pratico, ossia queste distribuzioni devono contribuire a semplificare il calcolo che conduce, attraverso l'applicazione del teorema di Bayes, all'aggiornamento del valore dei parametri considerati.

Si supponga di conoscere la varianza della popolazione $\sigma^2 = \sigma_0^2$, che, per esempio, in prima istanza può essere ritenuta uguale alla varianza campione. In quest'ottica una scelta conveniente per la distribuzione *a priori* della media è la distribuzione normale $N(m, d^2)$, dove m è la media campione e d^2 è la varianza campione.

Questa scelta può essere operata sia in base alla congruenza con le ipotesi sulla media espresse dall'analista che, soprattutto, osservando che la distribuzione normale è una coniugata.

Una distribuzione *a priori* si definisce coniugata se, dopo aver applicato il teorema di Bayes ad essa, la distribuzione *a posteriori* risultante appartiene sempre alla stessa famiglia di distribuzioni, ad esempio se la *a priori* è una distribuzione normale anche la *a posteriori* sarà una distribuzione normale. Quello che cambia fra le due distribuzioni sono i parametri, che vengono modificati dall'applicazione del teorema di Bayes in virtù dei dati campionari disponibili.

$p(\mu) \sim N(m, d^2)$ $\qquad$ $p(\mu|y) \sim p(\mu)p(y|\mu) \sim N(\hat{\mu},\hat{\sigma}^2_{\mu})$

dove:

$$\hat{\mu} = \frac{\frac{m}{d^2} + \frac{\bar{y}}{\sigma_0^2}}{\frac{1}{d^2} + \frac{1}{\sigma_0^2}} \quad \text{e} \quad \hat{\sigma}_\mu^2 = \left(\frac{1}{d^2} + \frac{n}{\sigma_0^2} \right)^{-1}$$

sono i parametri della distribuzione *a priori* automaticamente aggiornati dal teorema.

Un discorso simile può essere fatto con la distribuzione *a priori* della varianza, per la quale una scelta opportuna è la distribuzione gamma inversa, che è anch'essa una coniugata.

Effettuando queste scelte la distribuzione congiunta *a priori* dei due parametri, nell'ipotesi che essi siano indipendenti, sarà una Γ normale-inversa:

$p(\mu, \sigma^2) \sim \text{N-Inv-}\Gamma(\mu_0, \sigma_0^2/k_0; \nu_0, \sigma_0^2)$

dove μ_0, σ_0^2/k_0, ν_0 e σ_0^2 sono i parametri descrittivi della distribuzione.

Poiché il prodotto di due coniugate dà sempre una coniugata, la distribuzione *a posteriori* dei due parametri, condizionata all'insieme dei dati che si ottiene dall'applicazione del teorema di Bayes alla distribuzione congiunta, appartiene alla stessa famiglia della distribuzione *a priori*, ma con i parametri aggiornati:

$p(\mu,\sigma^2 \mid y) \sim \text{N-Inv-}\Gamma(\mu_i, \sigma_i^2/k_i; \nu_i, \sigma_i^2)$

dove:

$$\mu_i = \frac{k_0}{k_0 + k_n}\mu_0 + \frac{n}{k_0 + k_i}\bar{y}$$

$$k_i = k_0 + n$$

$$\nu_i = \nu_0 + n$$

$$\nu_i \sigma_i^2 = \nu_0 \sigma_0^2 + (n-1)s^2 + \frac{k_0 n}{k_0 + n}(\bar{y} - \mu_0)^2$$

sono i parametri aggiornati ed *i* è il minimo indice di iterazione per il quale le stime non variano significativamente.

7.2.3 Statistica "B" e modello gerarchico per i dati di espressione genica

Sulla base degli argomenti appena trattati è ora possibile introdurre la statistica *B*, o fattore di Bayes, per la discriminazione dei geni differenzialmente espressi, inserendola in un "framework" bayesiano basato su di un modello gerarchico dei dati di espressione genica.

Si indichi con:

$$M_{ij} = \log_2 \frac{R_{ij}}{G_{ij}}$$

il logaritmo del rapporto dei due canali per il gene *i* sull'array *j*, con i=1,…,N e j=1,…,n.

Si supponga che i dati M_{ij}, relativi ad ogni gene i, si distribuiscano seguendo una variabile aleatoria normale con media μ_i e varianza σ_i^2:

$$M_{ij} \mid \mu_i, \sigma_i^2 \sim N(\mu_i, \sigma_i^2) \qquad \forall\, i$$

Per indicare se un gene *g* è differenzialmente espresso oppure non ha mutato la sua espressione in seguito al trattamento effettuato, si può utilizzare un insieme di indicatori o indici definito nel modo seguente:

$$I_g = \begin{cases} 0 & \text{se il gene non è differenzialmente espresso} \\ 1 & \text{se il gene è differenzialmente espresso} \end{cases}$$

Per ogni gene *g* si può calcolare la probabilità *a posteriori* che esso sia differenzialmente espresso (I_g=1), dato l'insieme di osservazioni M_{ij}, e metterla in rapporto con la probabilità *a posteriori* che la sua espressione sia rimasta immutata (I_g=0) dato l'insieme di osservazioni M_{ij}; ciò corrisponde a definire il rapporto degli "odds" a posteriori per il gene *g*.

La statistica B per ogni singolo gene viene definita come il logaritmo del rapporto dei suoi "odds" a posteriori:

$$B_g = \log \frac{\Pr(I_g = 1 \mid M_{ij})}{\Pr(I_g = 0 \mid M_{ij})}$$

per cui $\Pr(I_g=1 \mid M_{ij}) > \Pr(I_g=0 \mid M_{ij})$ se e solo se $B_g > 0$.

Applicando il teorema di Bayes all'espressione di B_g e nell'ipotesi che gli M_{ij} siano indipendenti al variare di i, si può scrivere:

$$\begin{aligned} B_g &= \log \frac{p}{1-p} \frac{\Pr(M_{ij} \mid I_g = 1)}{\Pr(M_{ij} \mid I_g = 0)} \\ &= \log \frac{p}{1-p} \frac{\Pr(M_{i=g} \mid I_g = 1)}{\Pr(M_{i=g} \mid I_g = 0)} \frac{\prod_{i \neq g} \Pr(M_i \mid I_g = 1)}{\prod_{i \neq g} \Pr(M_i \mid I_g = 0)} \\ &= \log \frac{p}{1-p} \frac{\Pr(M_{i=g} \mid I_g = 1)}{\Pr(M_{i=g} \mid I_g = 0)} \end{aligned}$$

dove:

M_g è il vettore delle n osservazioni per il gene g;

p è la proporzione di geni differenzialmente espressi nell'esperimento, definita come $p = \Pr(I_i=1)$ per ogni i=1,…,N.

Per calcolare B_g è necessario quantificare l'espressione di $\Pr(M_g \mid I_g=1)$ e $\Pr(M_g \mid I_g=0)$, che sono le distribuzioni marginali di M_g rispetto agli indici e che possono essere indicate seguendo la notazione statistica con:

$$\Pr(M_i \mid I_g = 1) \equiv f_{I_i=1}(M_i)$$

$$\Pr(M_i \mid I_g = 0) \equiv f_{I_i=0}(M_i)$$

Ciò può essere realizzato attraverso la definizione di un modello gerarchico dei dati, ossia un modello nel quale la distribuzione di un parametro è condizionata a quella dell'altro. Definendo un modello gerarchico per le osservazioni è possibile utilizzare le informazioni relative all'intero insieme di dati per descrivere le distribuzioni marginali della media μ_i e della varianza σ_i^2, cioè dei parametri della distribuzione $N(\mu_i, \sigma_i^2)$ relativa alle osservazioni del gene i-esimo.

Utilizzando queste informazioni, la statistica B per ogni gene *g* assume la forma:

$$B_g = \log \frac{p}{1-p} \frac{1}{\sqrt{1+nc}} \left[\frac{a + s_g^2 + M_{g.}^2}{a + s_g^2 + \frac{M_{g.}^2}{1+nc}} \right]$$

La sola parte gene-specifica della statistica *B* è il rapporto fra parentesi quadre, che è sempre un numero ≥1 perché il denominatore è sempre minore o uguale al numeratore; ciò significa che un incremento nell'espressione differenziale, cioè un incremento della media $M_{g.}$ fa aumentare il valore della statistica anche quando la varianza è piccola e la presenza della costante di scala *a* garantisce che il rapporto non assuma valori troppo grandi a causa di medie $M_{g.}$ troppo piccole.

E' necessario, inoltre, porre l'attenzione sul fatto che, diversamente da altre statistiche, non esiste un valore di soglia statistico di *B* rispetto al quale dichiarare che un gene è differenzialmente espresso. Al crescere del valore della statistica si incrementa la probabilità che il gene possa essere considerato a ragione differenzialmente espresso e, analogamente, per valori negativi di *B* è molto più verosimile supporre che l'espressione differenziale sia assente.

7.3 Fonti di variabilità sui dati di espressione genica e modellazione della varianza dei dati

Le sorgenti di variabilità che si hanno per i dati di espressione genica possono includere sia fattori sperimentali sia rumore casuale o "*random*"; il metodo dell'analisi della varianza cerca di quantificare tale variabilità e di esaminare se sia statisticamente comparabile con quella attribuita alle sorgenti "random".

Si supponga, per esempio, di trattare con un farmaco un gruppo di cavie e di confrontare mediante microarray i campioni ottenuti dopo il trattamento con quelli di un gruppo di controllo non trattato: l'analisi della varianza consente di esaminare le differenze rilevate fra i gruppi, dividendole in effetto del trattamento ed effetto dovuto ai fattori sperimentali che si abbattono sull'espressione differenziale.

Il processo è concettualmente simile alla normalizzazione, poiché si tratta di eliminare, anche in questo caso, gli errori sistematici che contribuiscono a corrompere il dato di espressione, ma, in più, l'analisi della varianza permette di rilevare direttamente l'espressione differenziale sui dati ripuliti.

Il tipo più semplice di esperimento microarray consiste nel cercare di misurare i cambiamenti nell'espressione genica in campioni che differiscono per un unico fattore, ad esempio la somministrazione di un farmaco.

Si indicano con il termine *varietà* tutte le categorie del fattore di interesse: nel caso della somministrazione del farmaco le due categorie saranno *trattato* e *non-trattato* (*controllo*).

Nel loro lavoro Kerr e Churchill [24] hanno messo in evidenza che la variabilità può essere dovuta essenzialmente a quattro sorgenti principali:

- Effetto Array (A);
- Effetto Fluorocromo (D);
- Effetto Varietà o Trattamento (V o T);
- Effetto Gene (G).

Sotto il nome di effetto "Array" vengono classificate le variazioni di segnale fra array, mediate su tutti i geni, i fluorocromi e i trattamenti. Una problematica frequente che può condurre alla rilevazione di questo effetto può essere la non uniformità del processo di ibridazione del campione marcato.

L'effetto "Fluorocromo" o "Dye" misura le differenze intrinseche fra dei due fluorocromi. Nel caso di microarray "dual-color si può facilmente rilevare sin dalla fase di acquisizione dell'immagine che il fluorocromo Cy5 (rosso) ha un'efficienza di emissione più bassa rispetto al fluorocromo Cy3 (verde). Questo comportamento è dovuto ad una differente capacità dei due fluorocromi di incamerare e riemettere l'energia prodotta dal laser e si ripercuote sul bilanciamento del segnale nei due canali.

L'effetto "Varietà" si riscontra quando le categorie del fattore di interesse presentano livelli di espressione diversi, dovuti a fattori non riconducibili al trattamento. Questo potrebbe verificarsi, nel caso della somministrazione del farmaco, se venisse preso come controllo un tessuto diverso da quello trattato: l'espressione differenziale sarebbe riconducibile anche alle differenze fra i due tessuti.

L'effetto "Gene" si può verificare quando alcuni geni mostrano una diversa risposta all'ibridazione; ciò si manifesta con la generazione di una variazione del segnale, di intensità indipendente dalla quantità di campione ibridizzato.

Gli effetti descritti sono soltanto i fattori principali. Con quattro fattori principali è possibile considerare $2^4=16$ effetti complessivi ripartiti in:

- quattro effetti principali,
- sei interazioni a due fattori,
- quattro interazioni a tre fattori,
- una interazione a quattro fattori.

Anche nel caso dei fattori di interazione è possibile identificare alcune cause alla base dell'insorgenza degli effetti combinati.

L'effetto combinato del fluorocromo e del trattamento (DV) si può ricondurre ad una differente efficienza di incorporazione del marcatore nei campioni di cDNA da analizzare. Si supponga, per esempio, che il fluorocromo verde presenti una differente efficienza di incorporazione rispetto a due diverse varietà, mentre il fluorocromo rosso si comporti in maniera equivalente con entrambe. Questa situazione è schematizzata nella figura 7.5, dove la linea orizzontale è indicativa del comportamento costante del fluorocromo rosso, mentre quella obliqua evidenzia la differenza di incorporazione del fluorocromo verde sui due campioni T_1 e T_2.

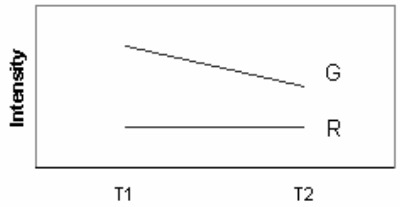


Figura 7.5: Schematizzazione dell'effetto combinato DV

In un esperimento in cui venissero realizzate due ibridizzazioni su due microarray con marcatura invertita dei campioni, sarebbero rilevate delle differenze in espressione non imputabili ad un effetto del trattamento, ma attribuibili al comportamento non omogeneo del fluorocromo verde.

L' effetto di interazione fra l'array e il gene (AG) si può verificare se lo stesso gene su diversi array è presente con una concentrazione diversa di sonde di cDNA disponibili per l'ibridazione. Questo effetto viene spesso denominato "Spot-effect" perché dipende fortemente dal processo di deposizione delle sonde sul microarray e, per eliminarlo, si possono seguire due strategie:

- considerare ogni spot come un'unità a sé, anche se così facendo si perdono le informazioni globali sul gene (per esempio le repliche sperimentali);
- cercare di ricostruire un modello statistico della densità dello spot o delle proprietà della punta di deposizione.

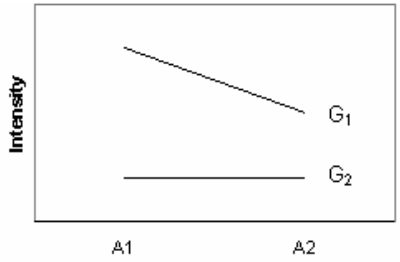


Figura 7.6: Schematizzazione dell'effetto combinato AG

Generalmente è difficile riuscire a modellare lo "Spot-effect", per cui si tende a migliorare la misura relativa ad ogni spot aumentando il numero delle repliche sperimentali, in modo da avere più dati a disposizione per l'interpolazione del loro modello.

L'effetto Dye-Gene (DG) si realizza se ci sono interazioni gene-specifiche fra il gene e il fluorocromo ed è detto anche "effetto dye gene-specifico".

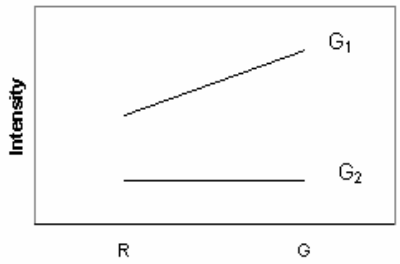


Figura 7.7: Schematizzazione dell'effetto combinato DG

L'interazione fra il trattamento e il gene (VG) si realizza quando un gene mostra espressione differenziale nelle diverse varietà ibridizzate sul microarray e questa differenza è riconducibile proprio al trattamento. La quantificazione di questo effetto è l'obiettivo principale dell'esperimento e la sua schematizzazione è mostrata in figura 7.8.

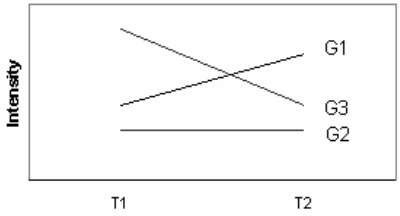


Figura 7.8: Schematizzazione dell'effetto combinato VG

Le interazioni AD, AT e ADT non sono gene-specifiche ed è difficile connettere ognuno di questi effetti combinati ai processi che si realizzano sui microarray.

Le interazioni ADG, ATG, DTG e ADTG sono invece gene-specifiche. La presenza di tali interazioni dimostrerebbe che ci sono variazioni attribuibili a particolari coppie array-fluorocromo, array-trattamento, fluorocromo-trattamento o combinazioni di array-fluorocromo-trattamento in relazione ad un particolare gene. Queste interazioni di ordine superiore al secondo sono difficili da collegare a processi fisici o chimici che si realizzano nei microarray e generalmente si assume che non si verifichino.

Esistono diversi modelli per quantificare le fonti di variabilità illustrate; la possibilità di valutare tutti gli effetti che compaiono in essi è consentita, oltre che da un adeguato numero di dati sperimentali, anche da un'opportuna pianificazione del modello statistico e del disegno dell'esperimento.

7.3.1 Modelli additivi ANOVA per l'analisi dell'espressione

La formulazione di un modello dei dati di intensità è legata non soltanto all'identificazione delle sorgenti di variabilità, ma anche ad un'adeguata caratterizzazione statistica degli effetti.

E' possibile distinguere fra effetti fissi, statisticamente modellabili con variabili aleatorie indipendenti ed identicamente distribuite, ed effetti "random", che presentano le caratteristiche di variabili aleatorie generate da processi tipicamente utilizzati per descrivere l'errore non sistematico di misura o errore "random".

Sulla base della descrizione statistica degli effetti, vengono definiti modelli "random", in cui tutti gli effetti coinvolti vengono considerati casuali, modelli misti nei quali viene individuata una parziale componente sistematica, e modelli fissi in cui tutti gli effetti sono sistematici a meno dell'errore di misura.

Per poter utilizzare questi modelli è necessario operare delle trasformazioni sui dati in modo da renderli idonei per la successiva elaborazione.

La trasformazione più frequentemente utilizzata sui dati grezzi di intensità è quella logaritmica, al fine di generare un modello additivo piuttosto che moltiplicativo. Utilizzando questa scala si indica con y_{ijkg} il logaritmo dell'intensità della fluorescenza misurata per l'array *i*, il fluorocromo *j*, la varietà *k* e il gene *g*.

Assumendo che lo stesso insieme di geni sia depositato su ogni array dell'esperimento, si ha a disposizione un insieme completo di osservazioni per ogni combinazione di array, fluorocromo e varietà: in conseguenza di ciò l'effetto gene e le sue combinazioni sono ortogonali, ossia indipendenti, a tutti gli altri effetti e l'esperimento si dice *bilanciato*.

Questo porta a suddividere gli effetti in due gruppi: effetti globali, che coinvolgono solo gli effetti principali A, D e V, ed effetti gene-specifici, che coinvolgono G. L'effetto di interesse VG è, quindi, gene-specifico.

Se gli effetti non sono ortogonali, ossia la quantificazione di uno fornisce informazioni ridotte o complete anche sull'altro, si parla di confusione dell'informazione, ossia di mascheramento parziale o totale degli effetti.

Modelli additivi misti

La scelta più generale operabile quando non vi sono informazioni per caratterizzare gli effetti come variabili indipendenti ed identicamente distribuite è considerarli tutti come effetti casuali e generare un modello "random". Malgrado questa scelta garantisca la completa generalizzabilità del modello, essa ha lo svantaggio di essere estremamente onerosa dal punto di vista della quantificazione di questi effetti.

Una valutazione più approfondita degli effetti e dei processi fisici che essi cercano di modellare può portare alla definizione di un modello misto, nel quale non tutte le sorgenti di variabilità vengono considerate casuali.

Questo tipo di modello per i dati di espressione genica è stato introdotto per la prima volta nel lavoro di Wolfinger e collaboratori [47] e si sviluppa in due stadi: un primo stadio di normalizzazione, cui segue un secondo di calcolo degli effetti gene-specifici.

La normalizzazione serve ad eliminare il contributo degli effetti principali globali ed essa viene realizzata attraverso la definizione di un sotto-modello per la loro stima, diverso a seconda degli effetti globali da stimare.

Un modello completo, in cui vengono presi in considerazione tutti gli effetti globali e gene-specifici, può avere la seguente forma:

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \varepsilon_{ijkg}$$

dove:

- il termine μ si riferisce all'intensità media totale calcolata su tutti i geni di tutti gli array;
- il termine ε rappresenta l'errore "random"; questo è una quantità aleatoria che si distribuisce secondo una variabile di Fisher con media nulla e varianza σ^2 e rappresenta tutta l'informazione che non si riesce a modellare.

Supponendo di aver utilizzato un disegno dell'esperimento che mascheri l'effetto della varietà con quello combinato dell'array e del fluorocromo (AD), come accade quando si inverte la marcatura dei due campioni confrontati nell'ibridizzazione su due vetrini, il modello parziale di normalizzazione ha la seguente forma:

$$y_{ijkg} = \mu + A_i + D_j + AD_k + x_{ijkg}$$

dove x_{ijkg} rappresenta il termine dei residui del modello, che, per ipotesi, hanno distribuzione normale.

La stima degli effetti globali A, D e AD e della media totale μ serve a "centrare" la distribuzione dei dati rispetto a questi effetti e assolve, quindi, lo stesso compito della normalizzazione globale illustrata nel capitolo 5.

In questo modo i dati grezzi di intensità vengono "ripuliti" senza ricorrere a tecniche di normalizzazione, ma solo attraverso la quantificazione di queste sorgenti di variabilità.

Il modello gene-specifico corregge invece gli effetti non-lineari di distorsione. La correzione delle fonti di variabilità attraverso una loro modellazione statistica *a posteriori* sui dati osservati e la successiva eliminazione implicano la progettazione dell'esperimento in funzione dell'acquisizione di un numero sufficiente di osservazioni per la loro quantificazione. Ai fini pratici questo si traduce in un ampliamento dell'esperimento che non sempre può essere realizzato, sia per l'eccessiva complicazione della procedura sperimentale che per il costo aggiuntivo. Per questo motivo, il pacchetto *MAANOVA*, nel quale vengono realizzati i modelli appena illustrati, mette a disposizione anche alcune tecniche di normalizzazione *within-arrays* al fine di consentire l'analisi della varianza anche per quegli esperimenti che non sono stati progettati per la valutazione sui dati delle fonti di variabilità. Non è invece fornito alcun metodo di sottrazione del "background", quantificato nel modello attraverso l'effetto "Array".

I residui x_{ijkg} del modello del primo stadio diventano i dati del modello del secondo stadio, che è, invece, un modello gene-specifico, come deducibile dal pedice *g* di ogni effetto, e serve a generare la stima dell'effetto di interesse VG e degli altri effetti combinati gene-specifici. Questo avviene grazie all'interpolazione ai minimi quadrati della formula che schematizza il secondo stadio del modello:

$$x_{ijkg} = \mu_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \varepsilon_{ijkg}$$

dove il termine μ_g si riferisce all'intensità media totale calcolata sul gene che si sta considerando.

Alcuni effetti possono essere considerati casuali in base alla considerazione che non vi è un'effettiva certezza che essi si abbattano in maniera sistematica su tutti i dati.

E' questo il caso dell'effetto AG che, verosimilmente, potrebbe avere un'entità diversa sullo stesso gene in array differenti. Dal punto di vista statistico l'effetto AG viene, quindi, trattato come se fosse una variabile aleatoria con distribuzione normale a media nulla e il modello così definito viene detto *misto*.

Modelli additivi fissi

I modelli additivi fissi proposti da Kerr e Churchill ereditano, dal modello misto appena illustrato, la formulazione a due stadi; anche in questo caso, infatti, lo stadio gene-specifico di valutazione dell'effetto VG viene preceduto da quello di normalizzazione.

Un modello ANOVA semplice include solo i fattori principali e l'effetto VG e può essere schematizzato con la formula seguente:

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + \varepsilon_{ijkg}$$

Un modello più plausibile aggiunge le variazioni spot a spot, includendo l'effetto combinato AG, e la sua struttura è descritta dalla formula seguente:

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + \varepsilon_{ijkg}$$

Un'altra possibilità è quella di aggiungere l'interazione fluorocromo-gene (DG), generando così il modello completo già illustrato nel paragrafo precedente.

Questi modelli sono relativi a situazioni in cui ogni gene è presente in una sola copia per array. Se i geni sono depositati in *r* copie per ogni array, la varianza del termine di interesse VG decrementa di un fattore 1/r ed è possibile inserire un "effetto replica S" nel modello, per catturare le differenze fra gli spot duplicati all'interno dell'array, così come indicato nella formula:

$$y_{ijkgr} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + S_{r(ig)} + \varepsilon_{ijkgr}$$

Poiché si assume che tutti gli effetti presenti nei tre modelli siano variabili aleatorie indipendenti e identicamente distribuite con media nulla, a meno del termine di errore "random", si parla di *modello fisso*.

Le stime degli effetti del modello sono realizzate attraverso un'interpolazione ai minimi quadrati, minimizzando la quantità:

$$\sum_{ijkgr} [y_{ijkgr} - \mu - A_i - D_j - V_k - G_g - (VG)_{kg} - (AG)_{ig} - (DG)_{jg} - S_{r(ig)}]^2$$

con i vincoli che:

$$\sum A_i = \sum D_j = \sum V_k = \sum G_g = \sum_g (AG)_{ig} = \sum_i (AG)_{ig} = \sum_g (VG)_{kg} = \sum_k (VG)_{kg} = \sum_g (DG)_{jg} = \sum_j (DG)_{jg} = \sum_r S_{r(ig)} = 0$$

L'effetto di interesse VG_{kg} per ogni gene g e trattamento k è ottenuto attraverso la stima ai minimi quadrati :

$$VG_{kg} = t_{..kg.} - t_{..k..} - t_{...g.} + t_{.....}$$

dove t rappresenta il logaritmo delle intensità e ogni punto dei pedici identifica il termine sul quale è stata eseguita la media.

7.3.2 "Nested" F-test e determinazione dei geni differenzialmente espressi

Una volta effettuata l'interpolazione ai minimi quadrati dei parametri del modello si può passare alla determinazione dei geni differenzialmente espressi.

Con questa tecnica di analisi dei dati di intensità, si decide se un gene è differenzialmente espresso realizzando un F-test delle ipotesi sul modello che è stato interpolato.

Seguendo lo schema classico del test delle ipotesi si definiscono:

- *ipotesi nulla o modello nullo:* il trattamento non ha effetto sul gene e $(VG)_{1g}=\ldots=(VG)_{kg}=0$ nel modello;
- *ipotesi alternativa o modello alternativo:* il gene è differenzialmente espresso e vi è almeno un *k* per il quale il termine $(VG)_{kg}\neq 0$ nel modello.

L'adeguatezza dei due modelli viene verificata attraverso un "nested" F-test. Due modelli vengono dichiarati "nested" o "annidati" se il modello definito completo o alternativo contiene tutti i termini del modello definito parziale o nullo e almeno un termine addizionale diverso da zero.

Se si definisce il modello nullo secondo la classica formulazione statistica come:

$$E(y)=\beta_0+\beta_1 x_1+\beta_2 x_2\ldots+\beta_g x_g$$

dove *E(y)* rappresenta l'aspettazione dei dati y, allora il modello alternativo che lo contiene avrà la forma:

$$E(y)=\beta_0+\beta_1 x_1+\beta_2 x_2\ldots+\beta_g x_g+\beta_{g+1}x_{g+1}+\ldots+\beta_k x_k$$

e dal punto di vista del test delle ipotesi, esse verranno definite come segue:

$$H_0:\beta_{g+1}=\beta_{g+2}=\ldots=\beta_k=0$$

$$H_a:\text{almeno un parametro }\beta_a\text{ con a}=\text{g}+1,\ldots\text{k è diverso da }0$$

Per testare queste ipotesi è possibile utilizzare un F-test in cui la classica statistica F viene sostituita con una che realizza il confronto fra i residui dei due modelli piuttosto che fra le varianze dei dati e che è definita come segue:

$$F=\frac{\left(SSE_{reduced}-SSE_{full}\right)/(k-g)}{SSE_{full}/\left[n-(k+1)\right]}$$

dove SSE indica la somma degli errori quadratici dei residui per i due modelli secondo la definizione classica, k sono i gradi di libertà per il modello nullo, g quelli per il modello alternativo e n è il numero delle osservazioni.

Questa statistica si distribuisce ancora come una variabile F di Fisher con $k-g$ gradi di libertà per il numeratore e $n-(k+1)$ gradi di libertà per il denominatore. La regola di rigetto dell'ipotesi nulla stabilisce che il modello nullo viene rifiutato se $F > F_{k-g,n-(k+1)}$., dove $F_{k\text{-}g,n\text{-}(k+1)}$ è il valore critico della statistica F tabulata.

Questo test è anche conosciuto con il nome di "F-test parziale" e per i dati di intensità ricavati dai geni si traduce in:

$$F = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{rss_1 / df_1}$$

dove *rss* è l'equivalente di SSE e *df* sono i gradi di libertà del modello nullo (pedice 0) e alternativo (pedice 1). Questa statistica è gene-specifica, poiché vengono utilizzati i dati dell'interpolazione del modello gene-specifico.

In un F-test è possibile utilizzare altre statistiche per la discriminazione dei geni differenzialmente espressi.

Se, per esempio, si vuole considerare una varianza dell'errore σ^2_{pool} comune su tutti i geni di tutti gli array, la statistica F può essere definita come:

$$F = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{\sigma^2_{pool}}$$

utilizzando un'informazione globale in supporto di quella gene-specifica espressa dal numeratore.

Una via di mezzo fra le due statistiche appena definite può venire dal considerare una combinazione di varianza globale e gene-specifica al denominatore della statistica da computare:

$$F = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{(rss_1 / df_1 + \sigma^2_{pool})/2}$$

Le tre statistiche sono praticamente equivalenti e l'adozione di una di esse può dipendere dalle informazioni che si hanno sui dati e dalle ipotesi fatte su di essi.

L'analisi dei residui del modello è utile non solo per determinare l'espressione differenziale dei geni, ma ha primariamente lo scopo di verificare l'adeguatezza del modello.

Infatti, dallo *scatterplot* dei residui è possibile rilevare la presenza di andamenti non casuali (o tendenze) sui dati dei residui; ciò indica l'inclusione di elementi di informazione in un elemento del modello che viene considerato "random" e, quindi, per definizione non informativo. Riscontrare una situazione del genere deve portare ad un'analisi più approfondita degli effetti da

considerare nel modello che viene interpolato, per non rischiare di mantenere errori sistematici che corrompono i dati o di non quantificare effetti di interesse.

Capitolo 8

Metodi di estrazione dell'informazione biologica

Una volta che sono stati identificati i geni la cui espressione è significativamente differente fra i campioni analizzati, il passo successivo è cercare di comprendere se essi possono fornire una spiegazione molecolare del fenomeno che è stato osservato.

A tal fine possono essere consultate molte banche dati, la maggior parte delle quali fa parte dell'NCBI (National Center for Biotechnology Information), che forniscono informazioni d diversa natura, più o meno accessibili gratuitamente: con questa operazione si procede alla annotazione dei risultati.

Le informazioni dai database di annotazione possono essere utilizzate per identificare nuovi "target" molecolari per studi successivi, per stabilire una scala di priorità per significatività biologica dei risultati dell'esperimento, o per identificare i percorsi di co-regolazione genica, comunemente detti "pathway", presenti nei risultati.

L'interrogazione delle banche dati può avvenire un gene per volta oppure sfruttando le potenzialità di software appositamente creati per abilitare la sottomissione delle liste complete dei geni differenzialmente espressi. Nel primo caso si parla di "single-gene analysis", mentre nel secondo di "pathway analysis" o "enrichment analysis".

Quest'ultimo approccio all'interpretazione dei dati è sicuramente più potente del primo, dal momento che l'ultima fase dell'analisi dei dati di espressione cerca di ricostruire il fitto scambio di informazioni fra geni che viene visualizzato con l'uso di una tecnologia di alto livello come i microarray.

8.1 Banche dati di annotazioni geniche

GenBank

Indirizzo: http://www.ncbi.nlm.nih.gov/Genbank/index.html [48]

Contiene la collezione annotata di tutte le sequenze pubbliche di DNA. Approssimativamente sono in Genbank 82853685 sequenze.

UniGene

Indirizzo: (http://www.ncbi.nlm.nih.gov/entrez/query.fgci?db=unigene) [49]

Organizza le sequenze contenute in GenBank in set non ridondanti di "cluster". Ciascun "cluster" rappresenta un gene e ad esso afferiscono tutte le sequenze collegate contenute in GenBank. Ogni "cluster" è identificato con un codice che spesso viene utilizzato sui file descrittivi dei microarray.

Entrez Gene (LocusLink)

Indirizzo: http://www.ncbi.nlm.nih.gov/entrez/query.fgci?db=gene/ [50]

Fornisce un'interfaccia per reperire informazioni relative a sequenze geniche controllate da gruppi di esperti e cataloga le sequenze in base al codice RefSeq.

Ensembl Genome Browser

Indirizzo: http://ensembl.org/ [51]

Fornisce e mantiene continuamente aggiornate le informazioni complete sul genoma di alcuni organismi eucariotici, in particolare sugli organismi vertebrati. Procura la corrispondenza fra diversi identificativi genici, compresi i codici delle sonde presenti sui microarray.

KEGG Pathway

Indirizzo: http://www.genome.jp/kegg/pathway.html/ [52]

La Kyoto Encyclopedia of Genes and Genomes contiene informazioni sui percorsi di co-regolazione che sono alla base della trasmissione del segnale genico. In essa sono presenti le mappe o "pathway" che visualizzano in maniera grafica i differenti livelli di interazione fra geni.

OMIM

Indirizzo: http://www.ncbi.nlm.nih.gov/entrez/query.fgci?db=OMIM/ [53]

La banca dati Online Mendelian Inheritance in Man collega circa 12000 geni con le malattie genetiche mendeliane conosciute.

HomoloGene

Indirizzo: http://www.ncbi.nlm.nih.gov/entrez/query.fgci?db=homologene/ [54]

Strumento che compara sequenze nucleotidiche fra coppie di organismi, fornendo il loro grado di omologia. Utilizzando HomoloGene le annotazioni possono essere combinate fra organismi.

GeneOntology

Indirizzo: http://geneontology.org/ [55]

GeneOntology è un consorzio che cerca di costituire un vocabolario unico per la descrizione dei prodotti genici in tre categorie. Esse sono la funzione molecolare, il processo biologico e la componente cellulare.

8.2 Strumenti per "single-gene analysis"

Il primo obiettivo dell'interpretazione dei dati è riuscire a capire se esiste una spiegazione del fenomeno molecolare osservato in relazione alla letteratura disponibile. In questa fase è quindi fondamentale riuscire a reperire il maggior numero di pubblicazioni inerenti l'argomento.

Il database che maggiormente viene consultato è Pubmed (http://www.ncbi.nlm.nih.gov/sites/entrez/), che contiene la maggior parte delle pubblicazioni di carattere medico, biologico, ma anche bioinformatico e biostatistico.

Le informazioni contenute in Pubmed sono solo una minima parte di quelle che devono essere utilizzate per l'interpretazione dei risultati. Infatti, ad esse bisogna aggiungere dati di carattere più puramente biologico, come per esempio, informazioni sulla sequenza del gene che si sta analizzando, il suo posizionamento cromosomico, il numero e la sequenza dei trascritti da esso generati, i suoi omologhi in altri organismi, etc.

Tutte queste notizie possono essere tratte dai database illustrati nel precedente paragrafo, ma la loro visualizzazione simultanea procura al ricercatore un quadro molto più informativo e pratico da consultare.

Per questo scopo sono stati realizzati molti strumenti che attingono dalle banche dati e presentano le informazioni in maniera più fruibile e diretta. Uno di questi è GeneCards® [56].

8.2.1 GeneCards®

Indirizzo: http://www.genecards.org/

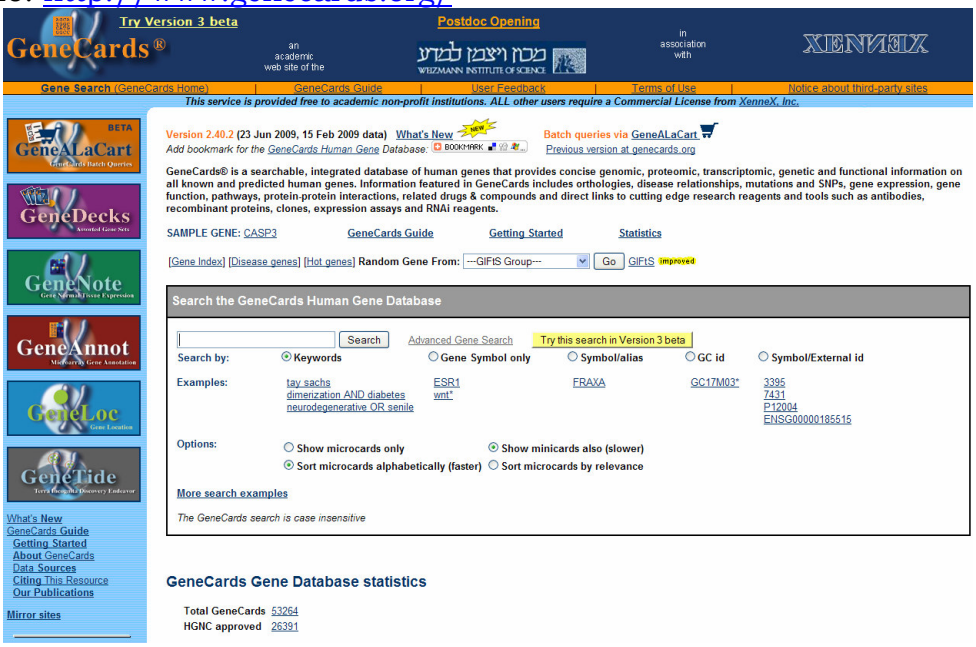


Figura 8.1: Interfaccia iniziale di GeneCards

GeneCards®, creato e mantenuto dal Weizmann Institute of Science di Rehovot (Israele) con la sponsorizzazione di diverse aziende produttrici di materiale utilizzato nella realizzazione di esperimenti di biologia, può essere interrogato sottoponendo uno dei codici identificativi del gene di interesse cui segue la visualizzazione di numerose informazioni in una Card, cioè in una scheda riassuntiva dalla quale si può partire per ricollegarsi ai database di origine. Le informazioni comprendono la lista completa di tutti gli Alias, cioè dei nomi alternativi del gene, con i rispettivi collegamenti alle banche dati HGNC, Entrez Gene, UniProtKB/Swiss-Prot (http://www.genenames.org/, http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene, http://www.uniprot.org/), [57, 58] informazioni sulla sequenza, corrispondenze con altri organismi, dati di espressione genica in diversi tessuti umani, posizionamento nei "pathway" e caratterizzazione in GeneOntology e molto altro. Le informazioni visualizzate nelle Card sono relative all'uomo, ma GeneCards può essere interrogato anche utilizzando codici di geni di altri organismi: di essi GeneCards ricava l'omologo umano passando attraverso HomoloGene. Inoltre, Genecards scarica direttamente da PubMed una lista di collegamenti a pubblicazioni che si riferiscono al gene per il quale si è fatta la ricerca e una lista più specifica che collega il gene a diverse patologie. Vi è anche una tabella che illustra le possibili interazioni del gene con alcune molecole chimiche la cui caratterizzazione è contenuta in diversi database, fra cui PharmGKB (http://www.pharmgkb.org/) [59].

Queste informazioni insieme a molte altre contenute nella Card relativa al gene, costituiscono un punto di partenza dettagliato per l'identificazione della direzione da prendere per l'approfondimento successivo.

8.3 Strumenti per l'analisi "pathway-level"

L'interrogazione in GeneCards avviene gene per gene perdendo in questo modo la complessa, ma estremamente ricca di significato, visione d'insieme procurata dai microarray.

A tale scopo sono stati prodotti diversi software che realizzano la cosiddetta "pathway analysis" o "enrichment analysis" attraverso l'uso delle informazioni contenute in KEGG. Con questo termine si intende un tipo di analisi mirata alla rivelazione di particolari "temi biologici" dei quali è "arricchita" la lista dei geni differenzialmente espressi, che potrebbero suggerire la formulazione di ipotesi biologiche a giustificazione della variazione di espressione genica osservata fra i campioni.

Gli strumenti che realizzano l'analisi di "pathway" si dividono essenzialmente in due classi: quelli che sovrappongono semplicemente la lista di geni differenzialmente espressi alle mappe di co-regolazione contenute in KEGG, e realizzano così una Over Representation Analysis (ORA), e quelli che cercano di associare un parametro statistico ad un "pathway" che sia esplicativo dell'importanza che esso assume dati i geni differenzialmente espressi che vengono mappati al suo interno.

Al primo gruppo di strumenti appartiene PathwayExplorer [60], mentre al secondo PathwayExpress [61].

8.3.1 Pathway Explorer

Indirizzo: https://pathwayexplorer.genome.tugraz.at/

L'accesso a PathwayExplorer può avvenire o via web, solo per l'organismo *homo sapiens*, oppure il software può essere installato sul proprio pc, scaricando in locale da KEGG i file che contengono la schematizzazione delle mappe per tutti gli organismi di cui è noto il genoma.

L'interfaccia di PathwayExplorer è estremamente semplice ed intuitiva:

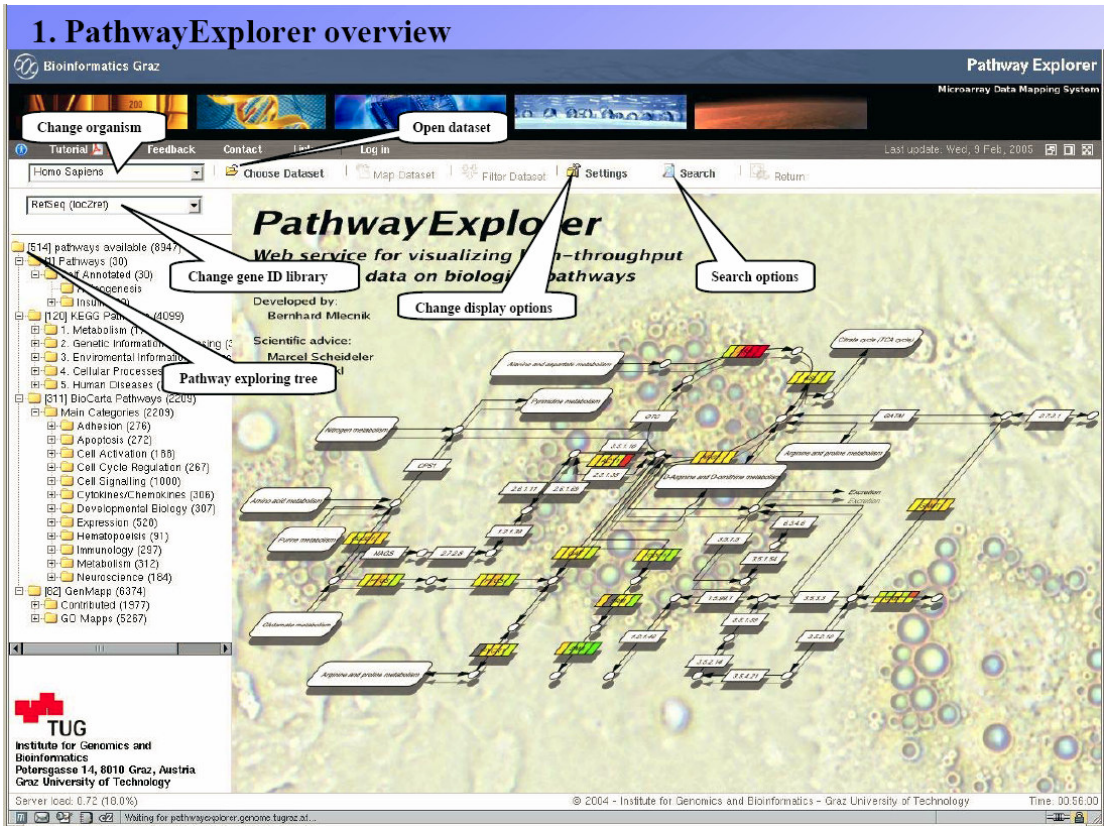


Figura 8.2: Interfaccia iniziale di PathwayExplorer

Il software a seguito della sottomissione della lista dei geni differenzialmente espressi, costituita almeno da una colonna di identificativi del tipo suggerito da PathwayExplorer e da quella dei "fold-change", posiziona all'interno delle mappe quei geni per i quali trova la corrispondenza con la banca dati di codici, specificata in fase di sottomissione.

Per ciascuna mappa viene visualizzata la posizione dei geni appartenenti alla lista sottomessa, mentre i differenti colori rappresentano i diversi livelli di espressione eventualmente caricati nel file.

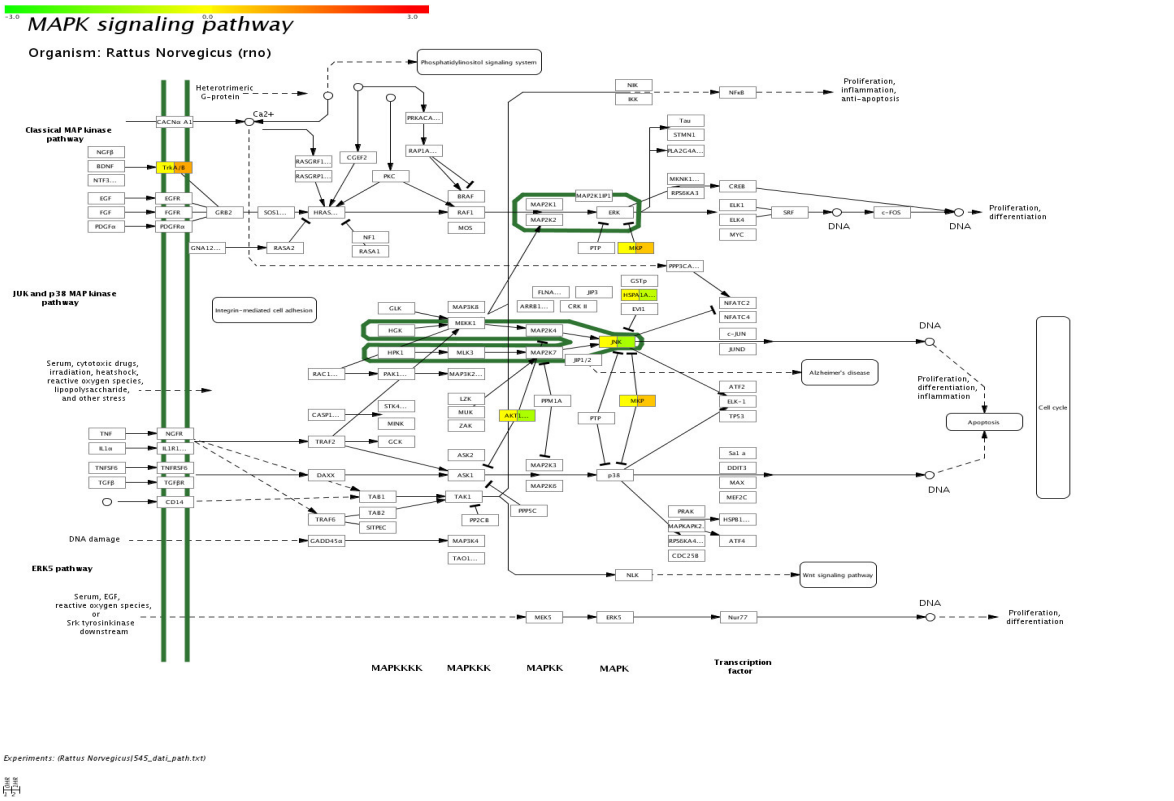


Figura 8.3: Visualizzazione della posizione di sei geni differenzialmente espressi utilizzando PathwayExplorer

La rappresentazione del risultato di un esperimento sulle mappe di co-regolazione genica semplifica l'interpretazione del risultato, dal momento che fornisce automaticamente la ricostruzione delle possibili interazioni biologiche fra geni.

8.3.2 PathwayExpress

Indirizzo:http://vortex.cs.wayne.edu/projects.htm#Pathway-Express

PathwayExpress fornisce un parametro, denominato "Impact Factor", che è esplicativo dell'importanza che una mappa di co-regolazione assume rispetto alle altre in relazione ai geni differenzialmente espressi che vengono mappati in esso.

All'interno della topologia di un "pathway" ciascun gene occupa una posizione che può essere utilizzata per organizzare gerarchicamente la lista dei geni coinvolti nella mappa. Schematizzando ciascun "pathway"come una rete costituita da nodi, rami di collegamento e foglie, è intuibile come un gene situato su di un nodo sia più importante di uno posizionato su una foglia collocata alla terminazione di un percorso. L'impatto che un gene differenzialmente espresso può avere sulla trasmissione del segnale genico è assai maggiore se esso si trova a monte del "pathway" o di una sua sottorete, piuttosto che in posizione isolata o alla sua terminazione.

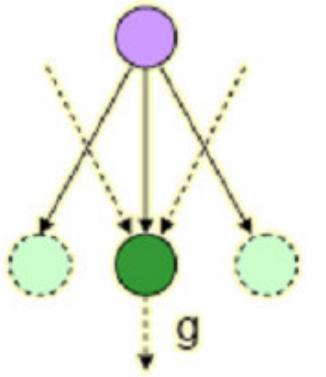


Figura 8.4: Schematizzazione della propagazione del segnale genico

La mappatura all'interno di un "pathway" di geni differenzialmente espressi con queste caratteristiche di importanza aumenta l'impatto biologico del "pathway".

PathwayExpress traduce questo concetto biologico in una misura statistica di rilevanza biologica del "pathway". Essa combina in un unico parametro la topologia della rete nei termini appena illustrati e la quota di geni differenzialmente espressi mappati all'interno di un "pathway" rispetto all'insieme dei geni presenti in esso. L'"Impact Factor" è accompagnato da una misura della sua significatività statistica, ricavata in termini di p-value valutando la probabilità che prendendo casualmente un insieme di geni appartenenti al "pathway" della stessa numerosità di quelli differenzialmente espressi mappati si possa misurare lo stesso effetto di perturbazione della mappa. Il p-value viene corretto per i test multipli utilizzando il False Discovery Rate.

PathwayExpress : Results - Demo Run - hepatic cell line treated with palmitate

Bar Graph

Input Genes / Pathway Genes	Hypergeometric Raw	Flag	IF Analysis p-value Raw	Corrected	Impact Factor	
3 / 89	7.236E-3		0.000E0	0.000E0	55.97	Antigen processing and presentation - KEGG
7 / 69	2.370E-8		1.632E-8	1.632E-8	21.023	Complement and coagulation cascades - KEGG
8 / 87	5.018E-9		3.077E-8	3.077E-8	20.358	Small cell lung cancer - KEGG
9 / 200	2.756E-7		9.688E-7	9.688E-7	16.722	Focal adhesion - KEGG
5 / 42	1.154E-6		5.492E-6	5.492E-6	14.877	Bladder cancer - KEGG
8 / 262	2.233E-5		2.027E-5	2.027E-5	13.479	MAPK signaling pathway - KEGG
5 / 68	1.290E-5		4.065E-5	4.065E-5	12.73	p53 signaling pathway - KEGG
5 / 87	4.278E-5	?	6.961E-5	6.961E-5	12.149	ECM-receptor interaction - KEGG
4 / 64	1.873E-4		9.957E-5	9.957E-5	11.761	Glioma - KEGG
6 / 138	3.502E-5		1.443E-4	1.443E-4	11.358	Cell Communication - KEGG
4 / 57	1.193E-4		4.445E-4	4.445E-4	10.128	Acute myeloid leukemia - KEGG
4 / 76	3.631E-4	?	9.434E-4	9.434E-4	9.298	Chronic myeloid leukemia - KEGG
3 / 136	2.251E-2	?	1.444E-3	1.444E-3	8.825	Tight junction - KEGG
3 / 90	7.461E-3	?	1.929E-3	1.929E-3	8.502	TGF-beta signaling pathway - KEGG
4 / 88	6.337E-4		1.961E-3	1.961E-3	8.484	Hematopoietic cell lineage - KEGG

Figura 8.5: Risultati di un'analisi realizzata con PathwayExpress

L'uso di uno strumento come quello appena illustrato modifica in maniera sostanziale la strutturazione stessa dell'interpretazione biologica dei risultati dal momento che l'informazione da esso fornita sposta il piano di osservazione dei risultati dai singoli geni all'intero "pathway" che li contiene. In questo senso l'interpretazione dovrebbe tentare di mettere in relazione i risultati ottenuti dall'esperimento e il fenomeno molecolare studiato attraverso l'indicazione fornita dai "pathway" statisticamente rilevanti, prima ancora che dai geni differenzialmente espressi in esso mappati.

Il successivo e più complesso passaggio è capire se la mappatura dei geni in un "pathway" è coerente con le informazioni biologiche disponibili su di essi e se è possibile produrre un'ipotesi biologica che spieghi il loro simultaneo coinvolgimento.

Quest'ultimo passaggio è ancora ben lontano dal poter essere automatizzato poiché è strettamente dipendente dalla capacità di schematizzare matematicamente le interazioni fra geni visualizzate nei "pathway". Alla loro base vi sono molteplici processi biologici, messi in atto dalla cellula per la regolazione dell'espressione e che possono modulare la trasmissione del segnale in maniera differente a seconda di quale di essi sceglie la cellula per rispondere allo stimolo. La conoscenza di questi processi costituisce la parte preponderante dell'interpretazione biologica dei risultati e, al momento, non vi è ancora uno strumento bioinformatico che riesca a sostituire il biologo in questo compito.

8.4 Rendere i dati pubblici: standard MIAME

Una volta che l'esperimento di espressione genica è terminato i dati possono essere resi pubblici a beneficio della comunità scientifica e dello scambio fra gruppi di ricerca.

Nell'ambito degli esprimenti di espressione genica è stato sviluppato uno standard per la presentazione delle informazioni contenute nell'esperimento, dai protocolli sperimentali fino a quelli di analisi dei dati. Questo standard è denominato MIAME (Minimum Information About a Microarray Esperiment) [62] e il protocollo relativo fornisce le linee guida per la collezione e l'impacchettamento delle informazioni.

Lo standard MIAME descrive sottoforma di "checklist", disponibile sul sito Microarray Gene Expression Data (MGED) (http://www.mged.org/), le informazioni minime che devono essere fornite quando si descrive un esperimento microarray.

Alcune riviste specializzate come condizione per la pubblicazione del lavoro relativo ad un esperimento microarray richiedono sempre più spesso che i dati vengano presentati seguendo le linee guida MIAME e che vengano resi pubblici sottomettendoli a un database che li rende disponibili.

I due più importanti database ai quali è possibile sottomettere i propri dati sono GEO [63] e ArrayExpress [64].

8.4.1 GEO Omnibus

Indirizzo: http://www.ncbi.nlm.nih.gov/geo/

Il Gene Expression Omnibus (GEO), mantenuto da NCBI, è uno dei "contenitori informatici" di dati di espressione genica maggiormente utilizzato. GEO offre una piattaforma pre-costituita per immagazzinare i dati di espressione genica e, allo stato attuale, ospita un totale di 322424 sottomissioni di dati di espressione da esperimenti microarray. Questa ampia collezione fornisce ai ricercatori materiale da utilizzare per confrontare le procedure sperimentali adottate o i risultati dei propri esperimenti, oppure per verificare ipotesi prima della pianificazione di un nuovo esperimento. Un altro valido utilizzo di questi dati è nella realizzazione di meta-analisi statistiche.

8.4.2 ArrayExpress

Indirizzo: http://www.ebi.ac.uk/microarray-as/ae/

ArrayExpress è un altro database pubblico di dati di espressione genica realizzato e mantenuto dall'EBI (European Bioinformatics Institute). In esso è possibile reperire i dati relativi agli esperimenti di espressione genica sottomessi dai gruppi di ricerca seguendo lo standard MIAME. L'infrastruttura informatica di ArrayExpress è costituita dalla base di dati, dai moduli di sottomissione in formato MAGE-ML (MicroArray Gene Expression-Markup Language), da uno strumento per fare la sottomissione diretta senza passare dalla formattazione manuale in MAGE-ML e da un'interfaccia per l'interrogazione del database e il recupero dei dati.

Mentre la sottomissione via MAGE-ML si presenta poco intuitiva per chi non è esperto di strumenti informatici per la manipolazione dei file di testo, lo strumento di sottomissione on-line è facilmente utilizzabile ed è corredato da una chiara documentazione sulle modalità e sui file da sottomettere.

La sottomissione avviene per passi e a ciascuno di essi corrisponde un differente protocollo sperimentale o di analisi dei dati che segue la scansione temporale dell'esperimento. Ad ogni protocollo viene associato un codice che potrà essere utilizzato per effettuare una ricerca selettiva, mentre l'EBI collega all'intero esperimento un codice che può essere utilizzato all'interno di una pubblicazione per comunicare la disponibilità dei dati.

L'EBI rilascia anche un punteggio che viene associato al dataset e che è esplicativo del grado di compatibilità dei dati immessi con lo standard MIAME.

Capitolo 9

Applicazione dei metodi in esperimenti di espressione genica realizzati mediante microarray: risultati e discussione

In questo capitolo verrà presentato il confronto critico fra i metodi descritti, applicati a quattro esperimenti di espressione genica differenziale realizzati presso il Laboratorio Microarray del Dipartimento di Patologia Sperimentale, Biotecnologie Mediche, Infettivologia ed Epidemiologia dell'Università di Pisa, durante lo svolgimento di questo dottorato.

I pacchetti utilizzati per realizzare l'analisi e la visualizzazione dei risultati sono *LIMMA*, *MAANOVA* e *SAM* e sono sviluppati in linguaggio R. Per l'analisi delle componenti principali è stato utilizzato *Genesis* [65]. L'analisi di "pathway" è stata condotta utilizzando *Pathway Express* e *Pathway Explorer*.

9.1 Esperimento E1: analisi dell'espressione genica in tessuto cerebrale di ratti trattati con fenitoina [66].

Questo primo esperimento ha avuto l'obiettivo di indagare il meccanismo molecolare di azione della fenitoina, un farmaco comunemente usato per il trattamento dell'epilessia. L'obiettivo principale del progetto era identificare profili di espressione genica nel tessuto cerebrale dopo trattamento con fenitoina, che potessero supportare un ruolo potenziale di questo farmaco come stabilizzante dell'umore, ruolo per cui recentemente sono emerse alcune evidenze cliniche [67-70].

A tal scopo è stato realizzato un esperimento microarray nel quale sono stati messi a confronto i profili di espressione dell'ippocampo e della corteccia frontale di tre ratti trattati cronicamente con fenitoina rispetto a quelli di tre ratti non trattati.

L'esperimento è stato realizzato in collaborazione con il Dipartimento di Biochimica Clinica dell'Università del Negev "Ben-Gurion" a Beer-Sheva in Israele. Il gruppo israeliano si è occupato del trattamento degli animali e del prelievo dei tessuti cerebrali.

L'RNA totale è stato estratto dai tessuti prelevati da ciascun ratto.

La concentrazione e la purezza di ciascun campione di RNA sono state misurate utilizzando lo spettrofotometro NanoDrop ND-1000 (NanoDrop Technologies, Wilmington, Del, USA), valutando l'assorbanza a 260 e a 280 nm e il rapporto di questi due valori, che è risultato superiore a 1.9 per tutti i campioni.

L'integrità dell'RNA è stata valutata utilizzando lo strumento Bioanalyzer 2100 di Agilent (Agilent Technologies, Palo Alto, CA, USA) e per tutti i campioni l'RNA Integrity Number è risultato superiore a 7, che rappresenta un buon valore per il tessuto cerebrale.

Gli RNA sono stati quindi amplificati, marcati con i fluorofori Alexa 555 e Alexa 647 e ibridizzati su vetrini Whole Rat Genome Oligo Microarray G4131A (Agilent Technologies, Palo Alto, CA USA), che contengono 44000 sonde rappresentative di circa 41000 trascritti di ratto.

Le immagini dei microarray sono state acquisite a 5 μm e PMT variabile utilizzando lo scanner Axon 4000B (Axon Instruments, USA) e l'estrazione dei dati grezzi è stata effettuata utilizzando il software GenePix PRO 6.0 (Molecular Devices, Sunnyvale, CA, USA).

9.1.1 Esperimento E1: disegno sperimentale

La progettazione del disegno di questo esperimento è stata vincolata dal numero limitato di campioni disponibili.

Poiché il numero dei campioni era dispari è stata automaticamente esclusa la possibilità di realizzare un Balanced Block Design (BBD). Dato, poi, l'esiguo numero di campioni e la necessità di quantificare in maniera efficace la differenza di espressione media fra le classi è stato escluso il Reference Design (RD) e si è optato per un confronto diretto fra campioni.

E' stato realizzato un Loop Design (LD) per ciascuna delle due aree cerebrali indagate, per un totale di dodici array ibridizzati.

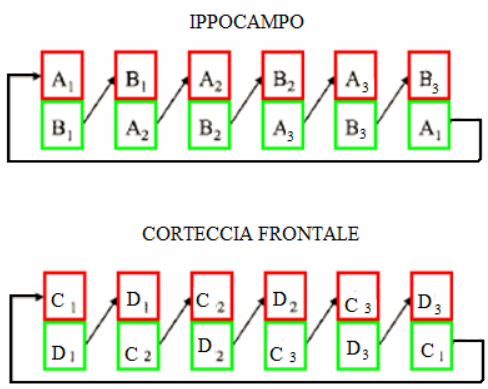


Figura 9.1: Disegno dell'esperimento E1. Le frecce servono ad evidenziare il "loop".

9.1.2 Esperimento E1: sottrazione del "background"

I metodi di sottrazione del "background" presi in considerazione nell'esperimento E1 per ripulire i dati grezzi dal segnale di intensità, dovuto a fenomeni non riconducibili ad ibridizzazione specifica, sono *subtract* e *minimum*, poiché al momento della realizzazione dell'esperimento non era ancora disponibile il metodo *normexp+offset*.

Utilizzando lo *scatterplot* dei dati di intensità è stato possibile identificare quali sono i contributi dei due canali all'intensità globale degli spot. Nella figura 9.2 sono presentati lo *scatterplot* e un ingrandimento alle basse intensità dei dati grezzi di uno dei 12 array utilizzati nell'esperimento E1.

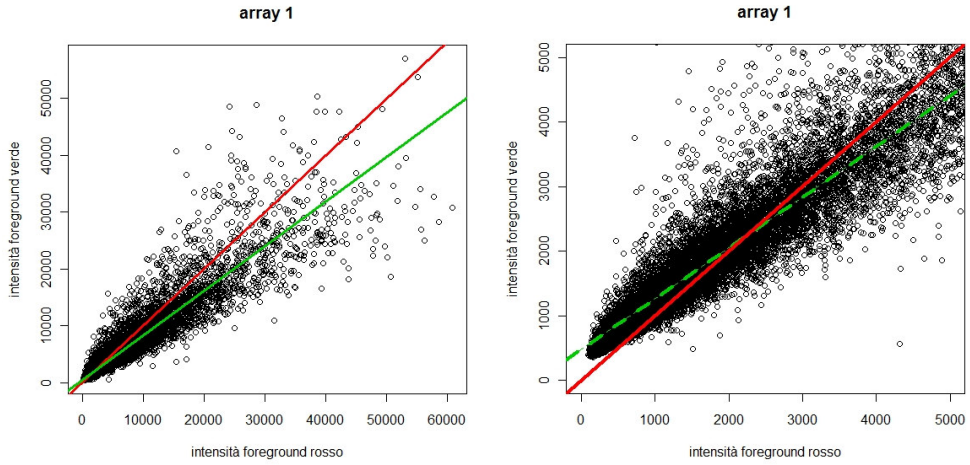


Figura 9.2: *Scatterplot delle intensità del "foreground" dei due canali nell'array 1(sx) e ingrandimento alle basse intensità di segnale (dx)*

La linea rossa in figura 9.2 rappresenta l'andamento ideale della retta di interpolazione dei dati grezzi, mentre quella verde è la retta di interpolazione reale. Come è possibile osservare, alle basse e medie intensità è preponderante il segnale di intensità verde, mentre alle alte intensità è maggiore il contributo del canale rosso.

Un maggior dettaglio del diverso contributo dei due canali al segnale globale degli spot è stato ottenuto visualizzando sullo *scatterplot* il valore

logaritmico delle intensità (figura 9.3). In questo caso si nota la tipica forma a banana con la punta rivolta verso l'alto, cioè verso le intensità verdi.

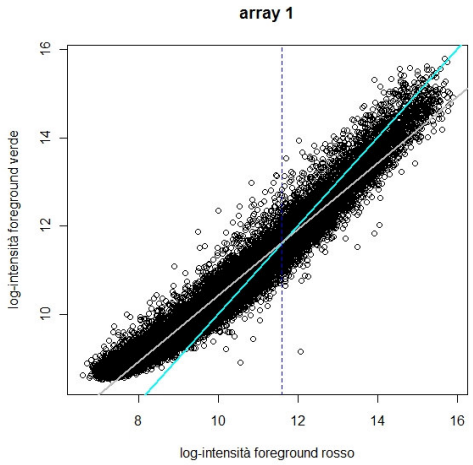


Figura 9.3: Scatterplot del logaritmo delle intensità

In figura 9.3 la linea tratteggiata rappresenta il livello di intensità logaritmica vicino al quale l'intensità del rosso inizia ad essere comparabile con quella del verde, mentre la linea grigia e quella azzurra sono l'equivalente rispettivamente delle linee verde e rossa in figura 9.2.

Per valutare l'effetto dell'operazione di sottrazione, i dati grezzi sono stati confrontati con quelli ripuliti attraverso l'uso di *MA plot* (figura 9.4).

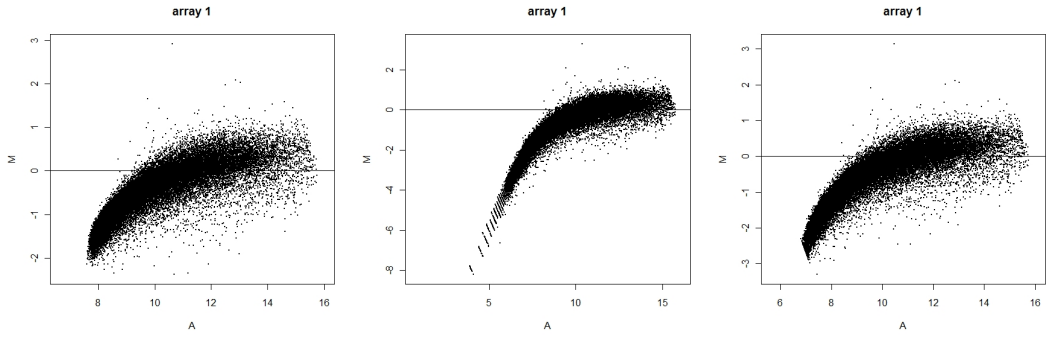


Figura 9.4: MA plot dei dati grezzi (sx), dei dati ripuliti col metodo subtract (centro) e dei dati ripuliti col metodo minimum (dx)

Il grafico MA dei dati grezzi (figura 9.4 sx) evidenzia ancora una volta la preponderanza del canale verde alle basse intensità. Detto in altri termini, poiché le intensità del "background" sono confrontabili fra i due canali, come evidenziato nella figura 9.5, è possibile ipotizzare che la sottrazione del "background" produca un effetto canale-dipendente e, in particolare, che gli effetti negativi di destabilizzazione della varianza alle basse intensità siano maggiormente presenti nel canale rosso piuttosto che in quello verde.

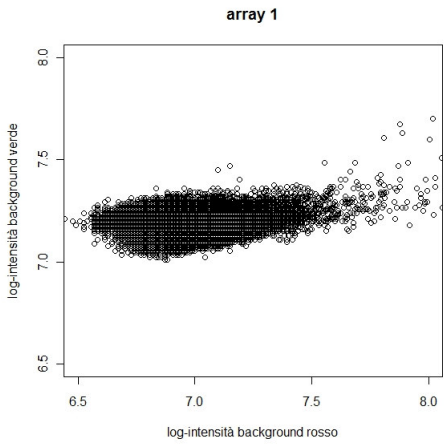


Figura 9.5: Scatterplot dei dati logaritmici di intensità del "background" dell'array 1

La frastagliatura della nuvola dei dati visualizzabile in figura 9.4 (centro) e lo spostamento verso valori di M negativi molto più alti in modulo rispetto a quelli dei dati grezzi, sono causati proprio dal problema appena illustrato e l'uso del metodo *subtract* per ripulire i dati dal rumore lo evidenzia maggiormente.

Per ovviare a questi inconvenienti, dovuti alla concomitanza del basso segnale nel canale rosso e all'applicazione di un metodo di sottrazione che non tiene conto della problematica, è stato deciso di utilizzare il metodo di sottrazione *minimum*.

Come è possibile osservare in figura 9.4 dx, la destabilizzazione della varianza alle basse intensità è molto più contenuta, anche se si osserva uno spostamento verso valori più negativi di M rispetto a quelli rilevati sui dati grezzi.

9.1.3 Esperimento E1: normalizzazione

Gli *MA plot* presentati nel precedente paragrafo mostrano chiaramente la presenza di una tendenza non lineare nella nuvola dei dati di intensità. Tale andamento implica la necessità di dover utilizzare metodi di interpolazione polinomiali per la normalizzazione dei dati.

Il numero e il tipo di spot presenti sugli array utilizzati per l'esperimento E1 hanno consentito di poter effettuare la normalizzazione utilizzando tutti i geni sull'array allo scopo di costruire la curva normalizzatrice o curva di "smoothing".

I dati sono stati normalizzati in maniera diversa a seconda del metodo statistico utilizzato per la successiva analisi e questo al fine di poter confrontare i risultati nella maniera più indipendente possibile e condurre una cross-validazione informatica.

Per realizzare l'analisi statistica con i pacchetti *LIMMA* e *SAM* (quest'ultimo non mette a disposizione alcun metodo di normalizzazione) i 12 array sono stati normalizzati *within-array* utilizzando il metodo *LOESS* ed interpolando un polinomio di secondo grado. La normalizzazione "paired-slide" per l'eliminazione dell'effetto-dye gene-specifico è stata effettuata in fase di analisi statistica aggiungendo al modello lineare un coefficiente che lo rappresenta e ne consente la valutazione.

Il risultato della normalizzazione sull'array 1 è visibile in figura 9.6.

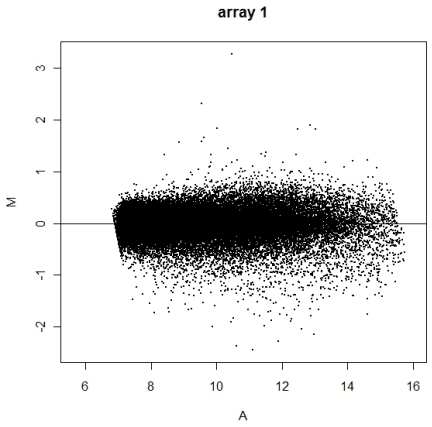


Figura 9.6: MA plot dell'array 1 normalizzato utilizzando il metodo LOESS

L'effetto complessivo della normalizzazione sui dodici array utilizzati nell'esperimento E1 può essere osservato nei boxplot in figura 9.7.

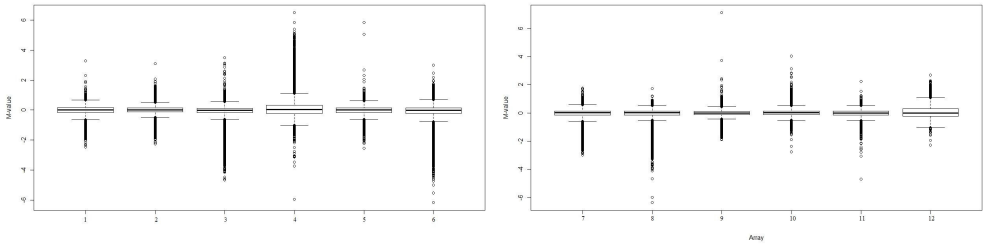


Figura 9.7: Boxplot dei docici array normalizzati con il metodo LOESS. A sinistra i sei array sui quali sono stati ibridizzati i campioni di ippocampo. A destra i sei array sui quali sono stati ibridizzati i campioni di corteccia frontale.

Le scatole contenenti il 50% dei dati visualizzate nei due boxplot sono molto simili fra loro: non è sembrato pertanto necessario procedere con una normalizzazione *between-array* per ciascuna area.

Per condurre l'analisi con il pacchetto *MAANOVA* sono stati sfruttati i metodi *linlog* e *rlowess*.

Il risultato dell'applicazione di questi metodi all'array 1 sono visualizzabili in figura 9.8.

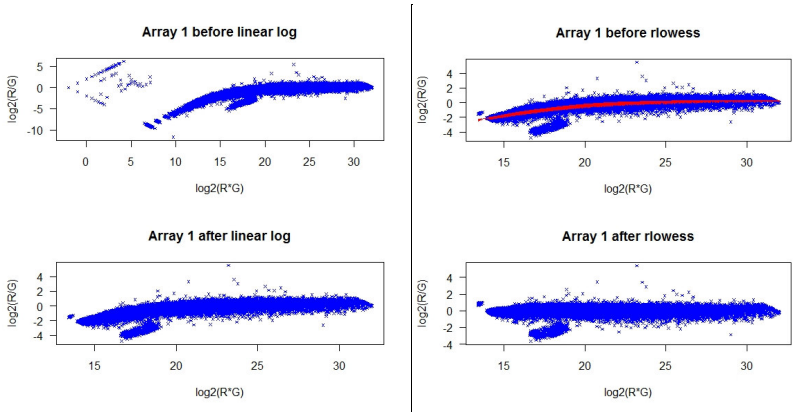


Figura 9.8: Applicazione della trasformazione lin-log (sx) e del metodo rlowess (dx) all'array 1

Come si può osservare nella figura 9.8 in basso, l'applicazione della trasformazione lineare-logaritmica ha prodotto un consistente effetto di stabilizzazione della varianza alle basse intensità (sx), mentre il metodo *rlowess* ha mitigato consistentemente la tendenza non lineare dei dati (dx) .

9.1.4 Esperimento E1: analisi statistica e risultati

Per realizzare l'analisi statistica dell'esperimento E1 e ricavare le due liste di geni differenzialmente espressi relative alle due aree cerebrali indagate sono stati utilizzati tutti e tre i metodi descritti nel capitolo 7. Per ciascuno di essi sono state imposte delle soglie alle statistiche utilizzate.

In *LIMMA* un gene è stato considerato differenzialmente espresso se presentava contemporaneamente B-statistic>0 e adj-P-value ≤ 0.01 per la t-statistic moderata. In *SAM* è stato utilizzato un FDR<0.01. In *MAANOVA* è stato interpolato un modello lineare per valutare il "dye-effect" come effetto globale, oltre al "dye-effect" gene-specifico e all'effetto di interesse relativo al trattamento con la fenitoina. Con quest'ultimo metodo un gene è stato considerato differenzialmente espresso se la sua F-statistic era minore di 0.01 in almeno una delle tre formulazioni del F-test.

Per ottenere la lista definitiva per ciascuna area cerebrale è stata fatta l'intersezione (evidenziata in rosso in figura 9.9) fra le liste fornite dai tre metodi statistici e un gene è stato dichiarato differenzialmente espresso se presente nell'intersezione fra *LIMMA* e almeno uno degli altri metodi. Mentre per l'ippocampo tutti e tre i metodi hanno fornito una lista, per la corteccia frontale *SAM* non ha identificato geni differenzialmente espressi.

La sovrapposizione fra le liste è illustrata in figura 9.9.

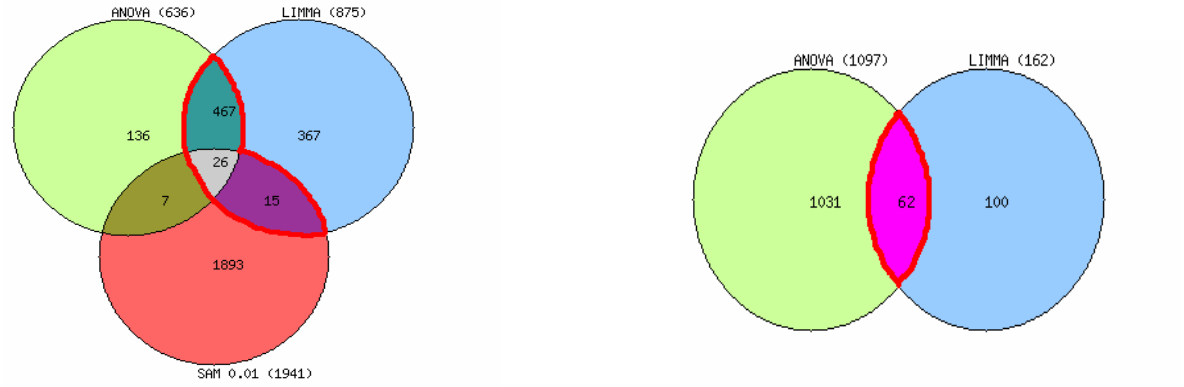


Figura 9.9: Diagrammi di Venn della liste di geni differenzialmente espressi ricavate utilizzando i tre metodi statistici

La somministrazione cronica di fenitoina ha alterato l'espressione di 508 geni nell'ippocampo, 465 sovraespressi e 43 sottoespressi, e di 62 geni nella corteccia frontale, 56 sovraespressi e 6 sottoespressi.

Le informazioni complete relative a questi risultati e all'esperimento E1 sono state rese disponibili in ArrayExpress e sono identificate dal codice E-MEXP-1728.

Al fine di studiare in maniera più approfondita i risultati delle due aree cerebrali è stata realizzata una PCA dei profili di espressione genica dell'ippocampo e della corteccia frontale (vedi figura 9.10).

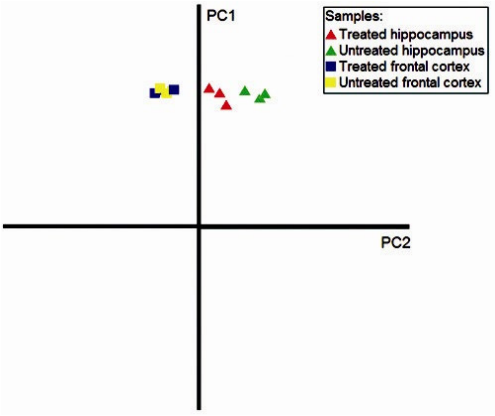


Figura 9.10: *Prime due componenti principali dei profili di espressione di ippocampo e corteccia frontale*

La prima componente principale spiega il 92% della varianza e separa i campioni in base all'area cerebrale. La distanza fra i due gruppi di campioni non trattati indica una considerevole differenza nell'espressione genica basale delle due aree cerebrali. Si può osservare una quasi completa sovrapposizione fra i campioni trattati e non trattati di corteccia frontale, che implica che la fenitoina non esplica un'azione sostanziale su questa area cerebrale. Al contrario, si osserva una maggior distanza fra i campioni trattati e non trattati dell'ippocampo, che può essere spiegata con una maggior ricettività di quest'area nei confronti del farmaco.

Questa analisi esplorativa dell'espressione genica globale nelle due aree concorda con il risultato dell'analisi statistica e fornisce una spiegazione intuitiva della notevole differenza di numerosità delle due liste di geni differenzialmente espressi.

9.1.5 Esperimento E1: Validazione in real time RT- PCR

I geni *Akt1*, *Impa1*, *Mapk10*, *Fyn*, *Rapgef4*, *Prkce*, *Frap1*, *Cap1*, *Gad1*, *Grina* e *Gclc,* risultati differenzialmente espressi nell'ippocampo, sono stati scelti per la validazione dei dati microarray, utilizzando come metodica alternativa la real time RT-PCR. Essi sono stati selezionati sulla base della significatività statistica e della rilevanza biologica. L'unico fra essi il cui valore di espressione non è stato confermato è *Rapgef4.*

Poichè la maggior parte delle sequenze differenzialmente espresse nella corteccia risultavano ancora non annotate come geni nelle banche dati, nessuna di esse è stata scelta per la validazione.

9.1.6 Esperimento E1: Analisi di "pathway" e interpretazione dei dati

L'analisi di "pathway" mediante *Pathway Express* e *Pathway Explorer* ha posizionato alcuni dei geni differenzialmente espressi in differenti "pathway" cellulari, fra i quali il metabolismo del glutammato, il metabolismo del

glutatione, la rete delle MAPK (Mitogen-Activated Protein Kinase), il sistema del fosfatidilinositolo (che ha ottenuto l'"Impact factor" più alto in *Pathway Express)*, la cascata Wnt, le "tight junction", i canali di trasporto ionici, il metabolismo degli acidi grassi e la neuroprotezione.

In particolare, *Pathway Express* ha localizzato 23 geni in 17 "pathway" nell'ippocampo e un gene in tre "pathway" nella corteccia frontale, questi ultimi differenzialmente espressi anche nell'ippocampo. *Pathway Explorer* ha localizzato 37 geni in 46 "pathway" per l'ippocmapo e quattro geni in dieci "pathway" nella corteccia frontale, questi ultimi compresi fra quelli dell'ippocampo.

Poiché molti dei 62 geni differenzialmente espressi nella corteccia frontale erano a funzione sconosciuta, non sono stati presi in considerazione per l'interpretazione biologica dei dati.

Data l'attuale scarsità di informazioni catalogate in maniera ordinata nelle banche dati disponibili, in particolare in KEGG, alla quale attingono tutti gli strumenti di analisi di "pathway", l'interpretazione dei dati di espressione differenziale relativi all'ippocampo è stata completata utilizzando *GeneCards*® e realizzando un'accurata ricerca in letteratura mediante *PubMed*. L'elenco dei geni interpretati con questo metodo e i processi biologici all'interno dei quali essi sono stati inseriti è consultabile in tabella 9.1.

NEURONAL EXCITABILITY	Modulation of GABAergic and Glutamatergic neurotransmission	*Gabra5*
		Gad1
		Glud1
		Grina
NEUROPROTECTIVE EFFECT	Regulation of cell proliferation and survival	*Akt1*
		Frap1
		Prkce
		Junb
		Mapk10
	Antioxidant action	*Gsr*
		Gclc
MEMBRANE TRAFFICKING		*Rab5a*
		Rab11b
		Rasa2
		Ap1s1
UNDEREXPRESSED GENES IN MOOD DISORDERS		*Gfap*
		Cap1

	Pdyn
	Ube2g1
	Uba5
	Ubac1
MYO-INOSITOL REGULATION	*Impa1*

Tabella 9.1: Riassunto dei geni interpretati e regolati dalla fenitoina nell'ippocampo

In conclusione con l'esperimento E1 è stato possibile rilevare che la somministrazione cronica di fenitoina regola geni coinvolti nella neurotrasmissione GABAergica e glutamatergica, implicati nella patofisiologia dei disturbi dell'umore. Inoltre, questa sostanza sembra esercitare un'azione neuroprotettiva in maniera simile al litio e al valproato, tradizionalmente utilizzati come stabilizzanti dell'umore, e modificare l'espressione di geni coinvolti nel meccanismo di regolazione dell'umore. Alcuni dei geni presenti in tabella 9.1 sono già riconosciuti come "target" degli stabilizzanti dell'umore, mentre altri, quali *Frap1*, *Gsr*, *Gfap*, *Pdyn* e *Cap1*, costituiscono nuovi tasselli nella comprensione dell'intricata patofisiologia della malattia bipolare.

9.2 Esperimento E2: Caratterizzazione dei profili di espressione di cellule di lievito trasfettate con cinque varianti missenso del gene BRCA1 [71].

Cinque varianti missenso del gene BRCA1: Y179C, S1164I, I1766S, M1775R e A1789T e la sequenza "wild-type" del gene sono state trasfettate stabilmente in cellule di lievito *S. cerevisiae* allo scopo di confrontare i profili di espressione genica indotti, nelle cellule ospiti, dalle mutazioni rispetto alla sequenza "wild type".

L'esperimento è stato realizzato in collaborazione con il laboratorio di Terapia Genica e Molecolare dell'Istituto di Fisiologia Clinica del CNR di Pisa e con la Sezione di Oncologia Genetica, Divisione di Patologia Chirurgica, Molecolare e Ultrastrutturale del Dipartimento di Oncologia dell'Università di Pisa.

Prima di analizzare i profili di espressione, le mutazioni sono state classificate sulla base del fenotipo evidenziato durante i saggi funzionali effettuati sulle cellule trasfettate. In particolare, si è osservato che l'inserzione di BRCA1 "wild-type" provoca nel lievito una soppressione della crescita che non si osserva in presenza delle mutazioni I1766S, M1775R e A1789T. Le mutazioni Y179C, S1164I, I1766S e M1775R, rispetto al "wild-type", determinano un'induzione della ricombinazione omologa. Le mutazioni sono state, quindi, suddivise in tre insiemi, corrispondenti ad altrettanti contrasti valutati attraverso la successiva analisi statistica dei dati microarray: Y179C e S1164I costituivano l'insieme R (Ricombinazione), I1766S e M1775R l'insieme RP (Ricombinazione e Proliferazione), mentre A1789T è l'unica mutazione dell'insieme P (Proliferazione).

Tranne che per la M1775R, che è stata già classificata come deleteria attraverso saggi di attivazione trascrizionale [72], non si hanno informazioni riguardo alle altre varianti missenso considerate. Tuttavia, le analisi *in silico* con Sorting Intolerant From Tolerant (SIFT) (http://blocks.fhcrc.org/sift/SIFT.html) e Polymorphism Phenotyping (PolyPhen) (http://tux.-embl-heidelberg.de/ramensky/polyphen.cgi) hanno evidenziato che tutte e cinque le varianti probabilmente inattivano la proteina. Esistono alcuni dati isolati di letteratura che descrivono la I1766S come deleteria [73], mentre risultati inconcludenti sono stati riportati sulla Y179C. Le varianti S1164I e A1789T sono state studiate solo nel lavoro di Caligo et al [74].

La concentrazione e la purezza di ciascun campione di RNA totale sono state misurate utilizzando lo spettrofotometro NanoDrop ND-1000 (NanoDrop Technologies, Wilmington, Del, USA), valutando l'assorbanza a 260 e a 280 nm e il rapporto di questi due valori, che è risultato superiore a 1.9 per tutti i campioni.

L'integrità dell'RNA è stata valutata su elettroforesi in gel di agarosio-formaldeide al 1.2%.

Gli RNA sono stati amplificati, marcati con i fluorofori Alexa 555 e Alexa 647 e ibridizzati su vetrini Yeast Oligo 2x11k Microarray G4140B (Agilent Technologies, Palo Alto, CA, USA). Ogni microarray è costituito da due sub-array all'interno dei quali sono immobilizzati circa 11000 oligonucleotidi rappresentativi delle 6256 Open Reading Frame di *S. cerevisiae*.

Le immagini dei microarray sono state acquisite a 5 µm e PMT variabile utilizzando lo scanner Axon 4000B (Axon Instruments, USA) e l'estrazione dei dati grezzi è stata effettuata utilizzando il software GenePix PRO 6.0 (Molecular Devices, Sunnyvale, CA, USA).

9.2.1 Esperimento E2: disegno sperimentale

Per l'esperimento E2 è stato scelto un confronto diretto dei campioni. Per eliminare la fonte di confondimento derivante dall'effetto "dye" gene-specifico è stato utilizzato un Dye Swap Design (DSD). Ogni campione di RNA proveniente dalle cellule trasfettate con BRCA1 mutato è stato confrontato due volte, su due array diversi e scambiando la marcatura, con quello delle cellule trasfettate con BRCA1 "wild-type": sono state così ottenute due repliche sperimentali per campione.

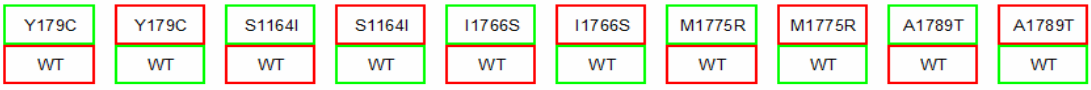


Figura 9.11: Disegno dell'esperimento E2

9.2.2 Esperimento E2: sottrazione del "background"

I metodi valutati per la sottrazione del "background" dei dati dell'esperimento E2 sono *subtract* e *minimum*.

Dallo *scatterplot* dei dati si osserva una prevalenza di intensità del canale verde alle basse intensità, seppure meno marcata rispetto a quella rilevata nell'esperimento E1, come è deducibile dalla minor curvatura della nuvola dei

dati, dalla maggior dispersione dei dati intorno alle rette d'interpolazione (figura 9.12) e, soprattutto, dal *density plot* (figura 9.13 (dx)).

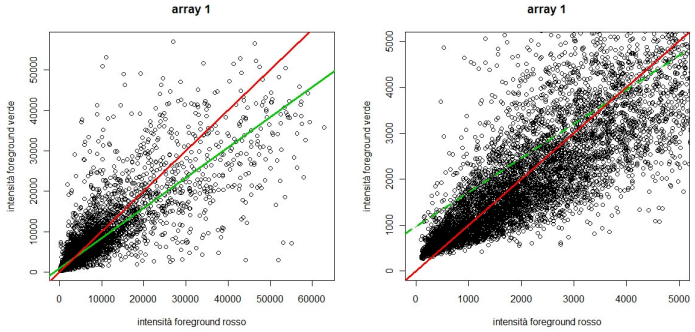


Figura 9.12: Scatterplot delle intensità del "foreground" dei due canali nell'array 1(sx) e ingrandimento alle basse intensità di segnale (dx)

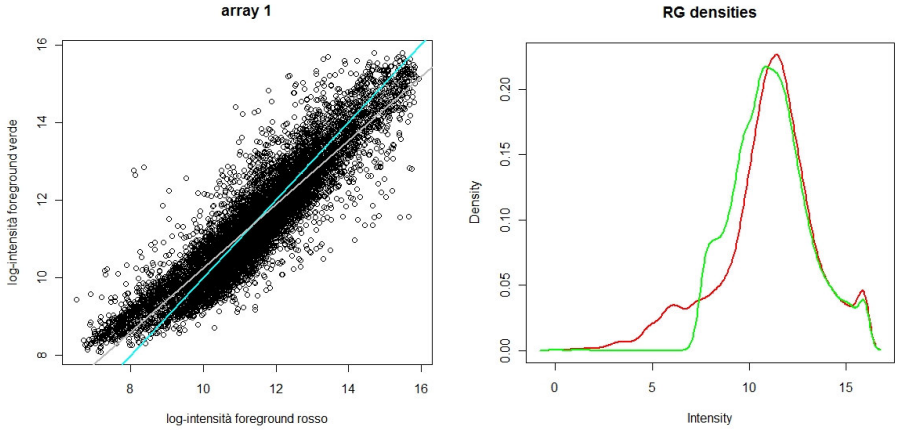


Figura 9.13: Scatterplot del logaritmo delle intensità (sx) e Density plot dei dati grezzi (dx)

Dalla visualizzazione su *MA plot* dell'effetto dell'applicazione dei due metodi di sottrazione del "background" (vedi figura 9.14) è stato possibile determinare che l'uso del metodo *minimum* sui dati di questo esperimento è più efficace del metodo *subtract* per la rimozione del rumore senza che questa operazione comporti un sostanziale innalzamento della varianza dei dati alle basse intensità.

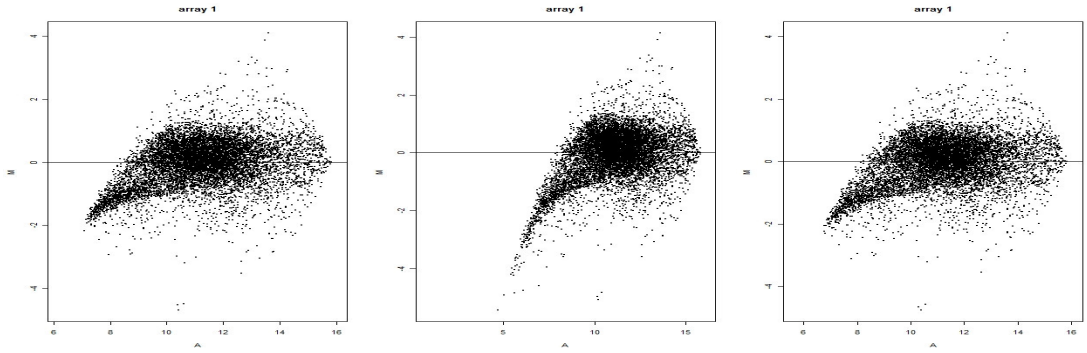


Figura 9.14: MA plot dei dati grezzi (sx), dei dati ripuliti col metodo subtract (centro) e dei dati ripuliti col metodo minimum (dx)

9.2.3 Esperimento E2: Normalizzazione

Come per l'esperimento E1, i dati sono stati normalizzati con metodi diversi a seconda dello strumento utilizzato per la successiva analisi statistica.

Per l'analisi con *LIMMA* e *SAM*, i dati ripuliti dal rumore sono stati normalizzati in prima istanza utilizzando il metodo *LOESS*. Il risultato della sua applicazione è visibile in figura 9.15 (centro). La normalizzazione "paired-slide" per l'eliminazione dell'effetto "dye" gene-specifico è stata effettuata in fase di analisi statistica aggiungendo al modello lineare un coefficiente che lo rappresenta e ne consente la valutazione.

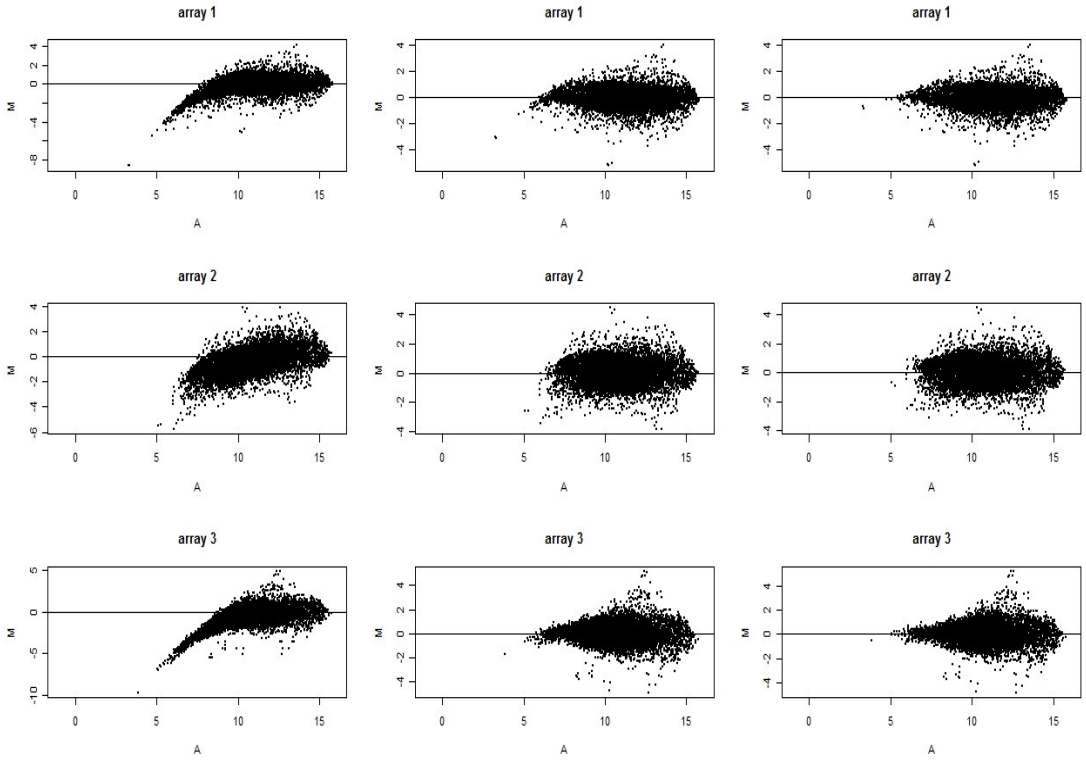


Figura 9.15: MA plot dei dati non normalizzati (sx), normalizzati con il metodo LOESS (centro) e con il metodo Aquantile (dx) per tre array utilizzati nell'esperimento E2

Gli array 1 e 3 in figura 9.15 (centro) evidenziano una persistenza della tendenza non lineare alle basse intensità. La successiva applicazione del metodo *Aquantile* ha consentito di eliminare completamente la non linearità.

La normalizzazione dei dati da sottoporre a *MAANOVA* è avvenuta utilizzando i metodi *linlog* e *rlowess*. Sono stati così moderati sia l'effetto di incremento della varianza alle basse intensità che la tendenza non lineare residua. *MAANOVA* non mette a disposizione metodi di normalizzazione *between array* per cui non è stato possibile uniformare le copie biologiche fra loro al fine di migliorare la normalizzazione.

9.2.4 Esperimento E2: analisi statistica e risultati

L'analisi statistica dei dati di espressione dell'esperimento E2 è stata realizzata utilizzando tutti e tre i pacchetti. I criteri per definire un gene differenzialmente espresso sono quelli utilizzati per l'esperimento E1.

I metodi statistici hanno fornito una lista di geni differenzialmente espressi per tutti i gruppi fenotipici, tranne che nel caso del gruppo P quando analizzato con *SAM*.

Le tre liste di geni differenzialmente espressi, corrispondenti ai gruppi di fenotipo P, R ed RP, sono state ricavate con *LIMMA* impostando una matrice dei contrasti nel modello statistico e interpolando un contrasto per ciascun gruppo fenotipico. Il livello di espressione differenziale assegnato è stato ottenuto dalla media dei livelli di espressione delle mutazioni simili per fenotipo, rispetto al *"wild-type"*.

In *MAANOVA* è stato interpolato un modello lineare per valutare il "dye-effect" come effetto globale, oltre al "dye-effect" gene-specifico e all'effetto che identifica l'appartenenza di una variante a uno dei gruppi fenotipici.

La sovrapposizione fra i risultati dei tre metodi per i gruppi R ed RP è mostrata in figura 9.16.

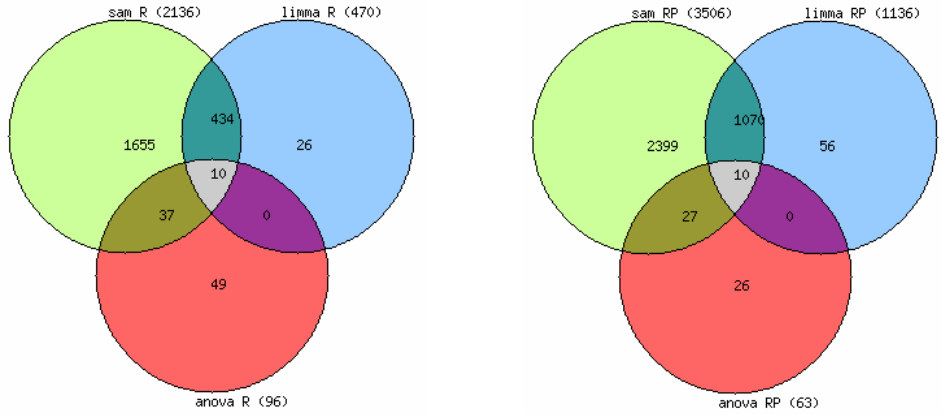


Figura 9.16: *Sovrapposizione fra le liste di geni differenzialmente espressi ricavate utilizzando SAM, LIMMA e MAANOVA per i gruppi fenotipici R (sx) e RP (dx)*

Poiché la sovrapposizione fra i risultati, in particolare fra quelli di *LIMMA* e *SAM*, è stata quasi totale, per la successiva fase di interpretazione sono state considerate le tre liste complete ricavate utilizzando *LIMMA*.

LIMMA ha prodotto 470, 740 e 1136 geni differenzialmente espressi rispettivamente nei gruppi R, P ed RP. Il livello di sovrapposizione delle tre liste è osservabile in figura 9.17.

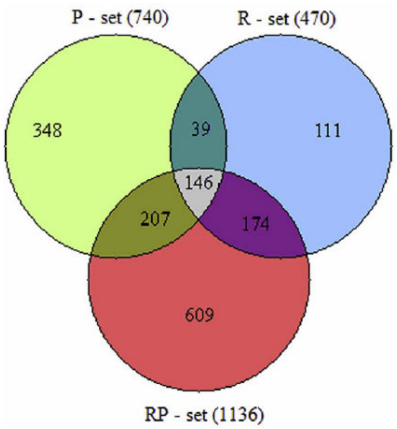


Figura 9.17: Intersezione delle tre liste di geni differenzialmente espressi ricavate nell'esperimento E2 utilizzando LIMMA

Le informazioni complete riguardo a questi risultati e all'esperimento E2 sono state rese disponibili in ArrayExpress e sono identificate dal codice E-MEXP-1867.

La visualizzazione su *heatmap* dei dati di espressione dei geni appartenenti alle quattro intersezioni ha evidenziato una concordanza della direzione del "fold-change" pari a 99.64%.

9.2.5 Esperimento E2: Validazione in real time RT- PCR

I geni *RNR1, POL30, SKM1, HHF2* e *ADE1* sono stati selezionati per la successiva validazione in real time RT-PCR. Gli esperimenti di validazione sono stati condotti separatamente per ciascuna delle varianti analizzate e il valore di espressione dei geni selezionati per ciascuna di essi è stato confrontato con quello ricavato dai microarray per il relativo gruppo fenotipico.

Tutti e cinque i geni hanno confermato il dato di espressione ricavato dai microarray.

9.2.6 Esperimento E2: analisi di "pathway" e interpretazione

L'analisi di "pathway" dell'esperimento E2 è stata condotta utilizzando esclusivamente *Pathway Explorer* poiché in *Pathway Express* sono assenti i dati relativi alle reti di co-regolazione dell'organismo *S.cerevisiae*.

Nei "pathway" sono stati posizionati circa il 20% gei geni differenzialmente espressi. Ciò è dovuto allo scarso livello di dettaglio dei "pathway" di lievito in KEGG, per la maggior parte appartenenti alla prima sezione dedicata ai processi metabolici.

Questo è il motivo per cui la maggior parte dei geni differenzialmente espressi piazzati da *Pathway Explorer* è stata assegnata a reti di tipo metabolico (circa il 90% dei geni mappati). Altri "pathway" interessanti emersi e non appartenenti alla sessione dei "pathway" metabolici sono il ciclo cellulare, la replicazione e la riparazione del DNA, anche se l'analisi di "pathway" non ha consentito una loro sufficiente caratterizzazione.

L'interpretazione è stata completata attraverso la consultazione della letteratura e, in particolare, utilizzando le informazioni presenti in Saccharomyces Genome Database (SGD) (http://www.yeastgenome.org), Ensembl (http://www.ensembl.org), information Hyperlinked Over Proteins (iHOP) (http://www.ihop-net.org/UniPub/iHOP/), Munich Information centre for Protein Sequences (http://MIPS.gsf.de).

L'uso approfondito di questi strumenti di navigazione delle informazioni ha consentito di individuare numerosi "pathway" correlabili ai fenotipi analizzati (tabella 9.2).

	Induzione della ricombinazione omologa						Recupero della proliferazione					
	R		P		RP		R		P		RP	
	Up	Down	Up	Down	Up	Down	Up	Down	Up	Down	Up	Down
Assemblaggio della cromatina		*HHF2 HTA2 HTB2 HAT1*		*HTB2 HAT1*		*HHF2 HTA2 HTB2 HIF1 HAT1*						
Metabolismo dei nucleotidi		*ADE1 ADE13 ADE17 ADE4 DUT1*		*ADE4 URA3 DUT1*		*ADE1 ADE13 ADE17 ADE6 ADE4 URA2 URA3 DCD1 PRS4 DUT1*		*RNR1*		*RNR1*	*RNR2 RNR4*	*RNR1*
Ciclo cellulare									*CLNG1*	*CDC6 CLN1 CLB6 RFC5 DRC1 DDC1 IPL1*		
Crescita invasiva e pseudo-ifale							*SKM1*		*FLO11 MEP2 GPA2 HMS1 ASH1 SKM1*	*DIG2*	*SKM1*	
Rimodellazione della cromatina						*ARP7 ARP9 SFH1*						*ARP7 ARP9 SFH1*
Controllo del ciclo cellulare		*MSH2*		*MSH2*		*TOP2 MSH2*		*MSH2*		*MSH2*		*TOP2 MSH2*

Tabella 9.2: Tabella riassuntiva dei geni coinvolti nei fenotipi analizzati nell'esperimento E2

9.3 Esperimento E3: Caratterizzazione dei profili di espressione di due varianti missenso di BRCA1 trasfettate in cellule HeLa

Questo esperimento rappresenta un ulteriore approfondimento della caratterizzazione molecolare di due delle cinque mutazioni di BRCA1 precedentemente trasfettate in cellule di lievito, la M1775R e la A1789T,

entrambe posizionate sul dominio BRCT del gene a 14 amminoacidi di distanza l'una dall'altra.

Anche questo esperimento è realizzato in collaborazione con il laboratorio di Terapia Genica e Molecolare dell'Istituto di Fisiologia Clinica del CNR di Pisa e con la Sezione di Oncologia Genetica, Divisione di Patologia Chirurgica, Molecolare e Ultrastrutturale del Dipartimento di Oncologia dell'Università di Pisa.

La scelta di M1775R è stata suggerita dall'assenza di informazioni di tipo molecolare a supporto del carattere patogenetico di questa mutazione, ipotizzato attraverso saggi trascrizionali. La A1789T è stata scelta, data la sua vicinanza alla M1775R, nell'ipotesi di poter osservare un profilo di espressione simile per entrambe le mutazioni.

Le due varianti e la sequenza "wild-type" di BRCA1 sono state trasfettate in maniera transiente in cellule HeLa, che possiedono una copia endogena del gene BRCA1, ma espressa a livello basso rispetto a quella trasfettata e quindi trascurabile.

La concentrazione e la purezza dell'RNA totale sono state misurate utilizzando lo spettrofotometro NanoDrop ND-1000 (NanoDrop Technologies, Wilmington, Del, USA), valutando l'assorbanza a 260 e a 280 nm e il rapporto di questi due valori, che è risultato superiore a 1.9 per tutti i campioni.

L'integrità dell'RNA è stata valutata utilizzando lo strumento Bioanalyzer 2100 di Agilent (Agilent Technologies, Palo Alto, CA, USA) e per tutti i campioni l'RNA Integrity Number è risultato superiore a 9.

I campioni sono stati amplificati, marcati con Cy3 e Cy5 ed ibridizzati su vetrini Whole Human Genome 4x44k Microarray G4112F (Agilent Technologies, Palo Alto, CA, USA). Ciascun microarray è costituito da quattro sub-array sui quali sono presenti oltre 45000 sonde rappresentative di circa 41000 trascritti umani.

Le immagini dei microarray sono state acquisite, a 5 μm e PMT fissato a 100%, utilizzando lo scanner Agilent G2565BA (Agilent Technologies, Palo Alto, CA, USA) e l'estrazione dei dati grezzi è stata effettuata con il software Feature Extraction 10.5.1.1 (Agilent Technologies, Palo Alto, CA, USA).

9.3.1 Esperimento E3: disegno sperimentale

Per contenere la variabilità biologica, rappresentata in questo esperimento da un'ipotetica differente risposta delle cellule alla trasfezione, è stato deciso di realizzare più trasfezioni in parallelo della stessa sequenza.

Prove di trasfezione hanno evidenziato la presenza di una certa variabilità nei risultati dei test funzionali, per cui si è deciso di utilizzare, per ogni trasfezione, un numero di cellule sufficiente ad ottenere tutto il materiale necessario per realizzare le ibridazioni microarray, la successiva validazione dei risultati in real time RT-PCR e i saggi funzionali.

Non esistono dati di letteratura che forniscono informazioni dettagliate sulla varianza biologica media di esperimenti di trasfezione effettuati su cellule in coltura, utili a calcolare il numero di copie biologiche necessarie. E', tuttavia, verosimile che tale varianza possa essere considerata più bassa di quella riportata per organismi superiori come il ratto, pari a 0.065 [7]. Sulla base di questo dato, se si sceglie di optare per un RD, sono necessari sei ratti per detettare una variazione di "fold-change" pari a due con una potenza del 95% e

una quota di falsi positivi pari a 1‰. E' stato, quindi, possibile ipotizzare di poter collezionare un numero inferiore di copie biologiche per detettare lo stesso "fold-change" utilizzando RNA proveniente da cellule trasfettate.

Il numero di copie biologiche per ciascuna sequenza trasfettata è stato, quindi, fissato a cinque, in base al numero totale di piastre di cellule gestibili in un unico esperimento di trasfezione.

Il numero dispari di campioni non ha consentito l'uso di un BBD. Ipotizzando una bassa variabilità biologica, non c'è molta differenza fra l'uso di un confronto diretto o di un confronto indiretto. Per questo motivo è stato scelto il RD, utilizzando come reference un campione ottenuto dall'unione di tutti gli RNA provenienti dalle cinque copie trasfettate con la sequenza "wild type" (pool).

M1775R	M1775R	M1775R	M1775R	M1775R	WT	WT	WT	WT	WT	A1789T	A1789T	A1789T	A1789T	A1789T
POOL	POOL	POOL	POOL	POOL	POOL	POOL	POOL	POOL	POOL	POOL	POOL	POOL	POOL	POOL

Figura 9.17: Disegno dell'esperimento E3

9.3.2 Esperimento E3: sottrazione del "background"

L'andamento dello *scatterplot* dei dati d'intensità evidenza un'inversione di tendenza rispetto ai dati degli esperimenti E1 ed E2. Quello che si osserva, infatti, è la predominanza del canale rosso rispetto a quello verde.

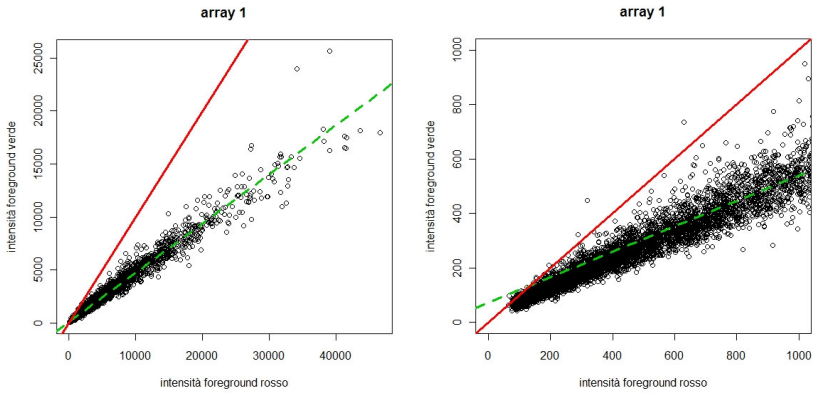


Figura 9.19: : Scatterplot delle intensità del "foreground" dei sue canali nell'array 1(sx) e ingrandimento alle basse intensità di segnale (dx)

Questo è stato determinato da una particolare efficienza di marcatura con il fluorocromo rosso del "kit" utilizzato, che è diverso da quello impiegato negli esperimenti E1 ed E2.

In maniera del tutto speculare alla figura 9.5, si può osservare che la variabilità maggiore è presente sui valori di "background" verde. E' possibile, inoltre, affermare che il livello di rumore sugli array dell'esperimento E3 è inferiore a quello rilevato sugli array dell'esperimento E1: la maggior parte dei valori di intensità logaritmiche si mantiene nell'intervallo 7-7.3 e 6.4-8 rispettivamente per i canali verde e rosso nell'esperimento E1, mentre è fra 4-6.5 e 5.2-6 nell'esperimento E3 (figura 9.20).

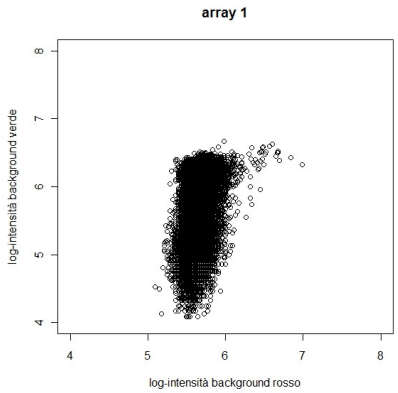


Figura 9.20: Scatterplot dei dati logaritmici di intensità del "background" dell'array 1

Applicando il metodo *subtract* per la sottrazione del "background" si osserva, infatti, un aumento della variabilità alle basse intensità con andamento inverso rispetto a quello degli esperimenti E1 ed E2.

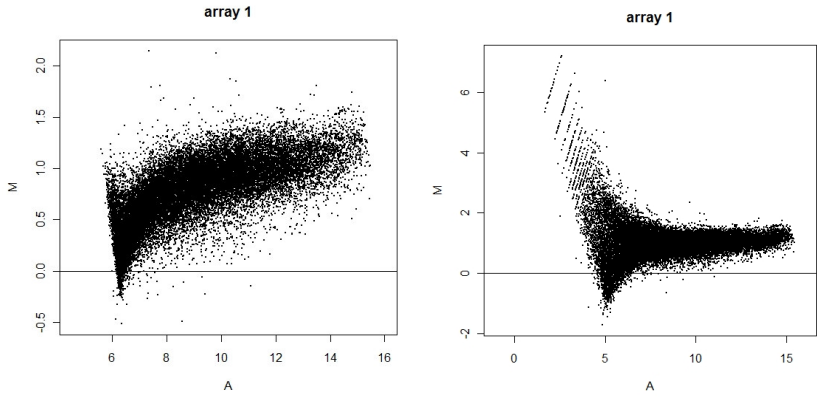


Figura 9.21: MA plot dei dati d'intensità dell'array 1 senza sottrazione del "background" (sx) e utilizzando per la sottrazione il metodo subtract (dx)

Questo marcato effetto ventaglio prevalente sulle intensità rosse può essere parzialmente corretto utilizzando il metodo *minimum*, ma è stato completamente eliminato dai dati di questo esperimento utilizzando il metodo *Normexp+offset*.

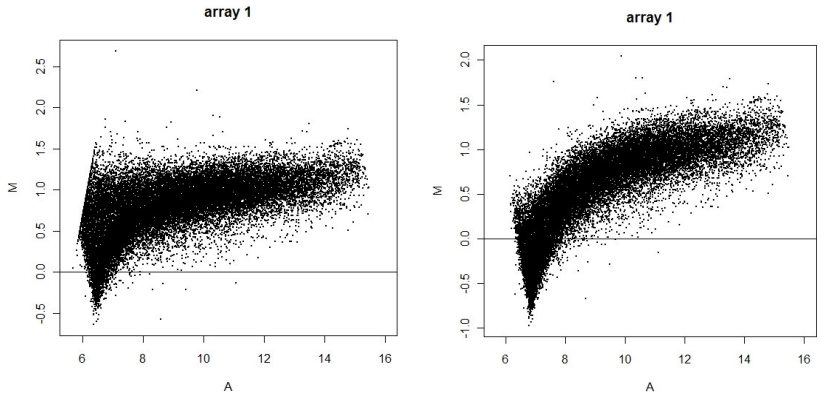


Figura 9.22: MA plot dei dati d'intensità dell'array utilizzando il metodo minimum (sx) e il metodo Normexp+offset (dx)

9.3.3 Esperimento E3: normalizzazione

Il grafico in figura 9.22 (dx) mostra chiaramente il permanere della tendenza non lineare dopo l'operazione di sottrazione del "background". Per eliminarla è stato necessario utilizzare il metodo LOESS, che ha avuto l'effetto di riequilibrare i due canali su ciascun vetrino. L'effetto della sua applicazione è stato visualizzato utilizzando l'*Image-plot*, che evidenza l'eliminazione dello sbilanciamento globale fra i due canali, e il grafico MA, che sottolinea la correzione delle differenze intensità-dipendenti fra i due canali.

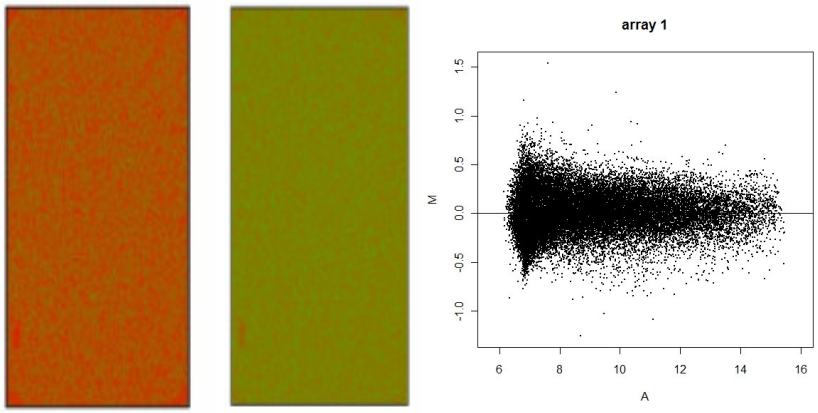


Figura 9.23: Image-plot prima(sx) e dopo (centro) dell'applicazione del metodo LOESS e MA plot (dx) dell'array 1

La visualizzazione dei dati così normalizzati su un *boxplot* ha evidenziato una lieve disomogeneità fra le scatole dei dati relative alle copie biologiche di ciascun gruppo. Per eliminare questa fonte di variabilità è stata applicata una normalizzazione sul solo canale rosso, che rappresenta il campione di riferimento di tutto l'esperimento E3, al fine di uniformarlo fra i vetrini.

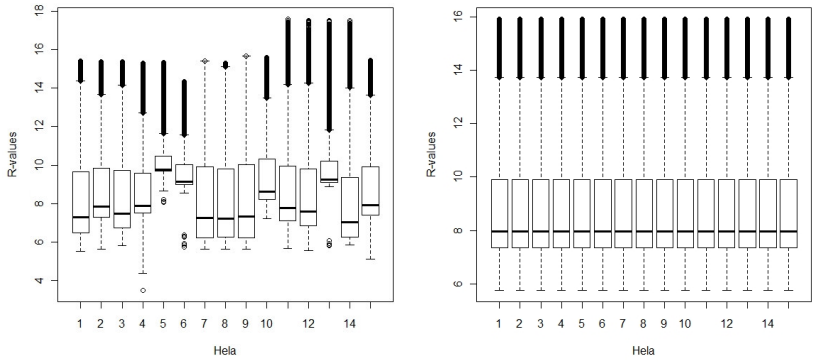


Figura 9.24: Logaritmo dei dati di intensità relativi al canale rosso prima (sx) e dopo (dx) la normalizzazione su singolo canale.

Successivamente è stata applicata una normalizzazione *quantile* a ciascuna delle tre classi separatamente, che ha compensato lo sbilanciamento fra le scatole dei dati d'intensità.

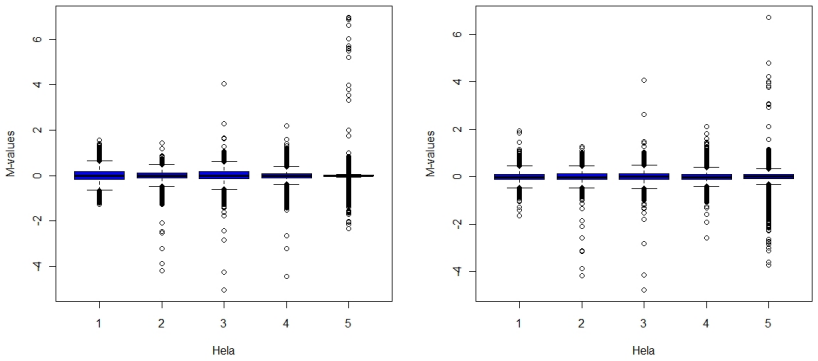


Figura 9.25: Boxplot dei dati relativi alla classe di campioni di HeLa trasfettate con la mutazione M1775R prima (sx) e dopo (dx) la normalizzazione con il metodo quantile

L'analisi statistica di questi dati è ancora in corso.

9.4 Esperimento E4: analisi dell'espressione genica in tessuti di ratti trattati con T_1AM.

Obiettivo di questo esperimento è indagare il meccanismo molecolare di azione della 3-Iodotironamina (T_1AM), un composto endogeno simile, per struttura chimica, agli ormoni tiroidei, in particolare all'ormone T_4 [75, 76]. Studi recenti [77, 78] hanno osservato che T_1AM è un ligando ad alta affinità di TAAR1, un recettore accoppiato alla proteina G e che l'iniezione di T_1AM in ratti provoca un rapido abbassamento della temperatura corporea e alterazioni nel ritmo e nella forza della contrazione cardiaca [79-81]. In aggiunta all'effetto ipometabolico è stato osservato che T_1AM aumenta il metabolismo dei lipidi a scapito di quello dei carboidrati [82].

Il progetto è svolto in collaborazione con il Laboratorio di Biochimica del Dipartimento di Scienze dell'Uomo e dell'Ambiente dell'Università di Pisa e prevede il confronto dei profili di espressione genica in tessuto adiposo e tessuto cardiaco di otto ratti trattati con T1AM a confronto con altrettanti ratti non trattati.

La concentrazione e la purezza dell'RNA totale sono state misurate utilizzando lo spettrofotometro NanoDrop ND-1000 (NanoDrop Technologies, Wilmington, Del, USA), valutando l'assorbanza a 260 e a 280 nm e il rapporto di questi due valori, che è risultato superiore a 1.9 per tutti i campioni.

L'integrità dell'RNA è stata valutata utilizzando lo strumento Bioanalyzer 2100 di Agilent (Agilent Technologies, Palo Alto, CA, USA) e per tutti i campioni l'RNA Integrity Number è risultato superiore a 9.

I campioni sono stati amplificati, marcati con Cy3 e Cy5 ed ibridizzati su vetrini Whole Rat Genome 4x44k Microarray G4131F (Agilent Technologies, Palo Alto, CA, USA). Ciascun microarray è costituito da quattro sub-array sui quali sono presenti oltre 45000 sonde rappresentative di circa 41000 tracritti di ratto.

Le immagini dei microarray sono state acquisite, a 5 µm in modalità XDR (eXtended Dynamic Range), utilizzando lo scanner Agilent G2565BA (Agilent Technologies, Palo Alto, CA, USA) e l'estrazione dei dati grezzi è stata effettuata

con il software Feature Extraction 10.5.1.1 (Agilent Technologies, Palo Alto, CA, USA).

9.4.1 Esperimento E4: disegno sperimentale

Il disegno dell'esperimento E4 è stato condotto in assenza di vincoli fissati prima della progettazione dello stesso. Poiché lo scopo principale dell'esperimento è stato il confronto dell'espressione fra tessuti di soggetti trattati e di controllo, è stato scelto a priori di realizzare un BBD, al fine di quantificare in maniera efficace ed efficiente la differenza di espressione media fra le classi di soggetti.

Per determinare quanti soggetti sarebbe stato necessario trattare per strutturare in maniera statisticamente robusta l'esperimento è stata sfruttata la formula 1.2 illustrata nel capitolo 1.

Se si ammette che nel BBD la variabilità biologica per il ratto sia otto volte quella sperimentale e, considerato che esiste una relazione matematica che lega la varianza biologica σ^2, rilevata in un esperimento realizzato utilizzando il RD, alla τ^2, ricavata utilizzando un BBD, come riportato in [7], è possibile calcolare il numero di campioni che servono per osservare un'espressione differenziale di due "fold-change" fra le classi con α e β desiderati.

In particolare, ciascuna delle due varianze è stata scomposta in due componenti [7]: ξ, che identifica la varianza biologica pura di ciascuna classe, cioè senza considerare la componente additiva data dall'errore sperimentale della metodica, e ν, che invece quantifica proprio quest'ultimo.

La scomposizione delle varianze rilevate con i due disegni sperimentali risulta in:

$$\sigma^2 = \xi^2 + 2\,\nu^2 \qquad (9.1)$$

$$\tau^2 = 2\xi^2 + 2\,\nu^2 \qquad (9.2)$$

Per la soluzione di questo sistema di due equazioni nelle quattro incognite è possibile utilizzare l'ipotesi sul rapporto fra varianza biologica pura e varianza sperimentale:

$$\xi^2 / \nu^2 = 8 \qquad (9.3)$$

Risolvendo il sistema col metodo di sostituzione si è ottenuto che ad una variabilità σ pari a 0.25 corrisponde una variabilità τ pari a 0.33.

Per determinare la numerosità campionaria è stato necessario utilizzare la formula 1.2 inserendola in un processo iterativo di aggiornamento che sfrutta le tabelle della distribuzione t di Student e i dati numerici appena illustrati. La risonanza fra i due risultati più probabili, evidenziata negli ultimi passaggi di iterazione, è stata risolta utilizzando come punto di equilibrio per il processo di aggiornamento il dato di letteratura relativo alla numerosità campionaria valutata per campioni umani da osservare in BBD.

Il calcolo ha suggerito di fissare a otto il numero minimo dei ratti per classe.

Figura 9.19: Disegno sperimentale dell'esperimento E4

Questo esperimento è ancora in corso e non sono al momento disponibili dati preliminari riguardanti l'espressione genica differenziale delle classi confrontate.

Capitolo 10

Conclusioni

Gli esperimenti di espressione genica realizzati con microarray sono estremamente complessi e la loro esecuzione richiede molto impegno e competenza da parte di tutte le figure coinvolte.

La fase sperimentale riguardante l'estrazione, la marcatura e l'ibridizzazione su vetrino dei campioni è regolata da protocolli abbastanza standardizzati, seppure diversi a seconda dell'azienda costruttrice dei vetrini e non sempre sufficientemente dettagliati.

Lo stesso non si può dire dell'analisi dei dati, che necessita di uno studio *ad hoc* per ciascun esperimento. Il diagramma di flusso presentato in figura I2 dell'Introduzione rappresenta, infatti, un tentativo di schematizzazione di questo articolato processo. Ciascuno dei pass in esso illustrati deve essere compiuto solo dopo avere verificato se è realmente necessario e i modi per realizzarlo sono molteplici a seconda delle peculiarità dell'esperimento in corso.

Molti errori possono essere evitati progettando adeguatamente l'esperimento prima di eseguirlo, il che richiede un attento studio *a priori*, che si articola in:

- ricerca approfondita di tutti gli elementi che possono contribuire alla progettazione e alla realizzazione
- formulazione di ipotesi biologiche dettagliate e di una loro scala di priorità
- valutazione di adeguate procedure di reclutamento dei campioni
- scelta attenta dello schema di confronto per contenere le fonti di variabilità.

Solo alla fine di questa rigorosa valutazione è possibile stabilire quali possono essere i margini per una deviazione dalla teoria, calcolare in che termini questa si abbatte sul risultato ed, eventualmente, accettare il rischio.

Una volta realizzato l'esperimento è necessario verificare sui risultati se le ipotesi formulate *a priori* in fase di progettazione, in particolare quelle riguardanti l'entità dell'espressione genica differenziale che si sarebbe dovuta rilevare, si sono attuate e prevedere eventualmente una successiva fase di ampliamento del numero dei campioni per consentire la generalizzazione del risultato.

Le tecnologie costruttive dei microarray e le procedure sperimentali sono sempre più capaci di ottenere risultati ottimali sia per ciò che riguarda l'allestimento del vetrino che la scansione dell'immagine. Tuttavia, è ancora necessario l'uso, seppure dibattuto, degli indicatori per l'esclusione di quegli spot che non soddisfano i parametri di qualità fissati dall'analizzatore dei dati. Da questo punto di vista, la progettazione dei filtri informatici che realizzano questa operazione è strettamente dipendente dalle scelte dell'operatore e la sua conservatività è regolata dall'uso di soglie il cui valore deve essere legato, per esempio, all'efficienza di marcatura delle sonde, allo scanner utilizzato per la rivelazione del segnale e ai software di estrazione dei dati. L'inclusione nell'analisi di "spot", il cui segnale non è determinato dalla corretta ibridizzazione delle sequenze di mRNA provenienti dai campioni, può determinare artefatti. L'applicazione di un controllo di qualità eccessivamente severo, d'altro canto, può ridurre al minimo, o, addirittura rendere insufficiente, l'informazione utile al processo di modellazione statistica che produce i risultati.

Un processo di valutazione equilibrato della qualità di un vetrino deve consentire di individuare tutti e soli gli "spot" la cui intensità non è il risultato dell'ibridizzazione specifica, ma anche di eliminare quegli array che risultano scadenti.

L'operazione di sottrazione del "background" è un passaggio estremamente delicato che ha due effetti collaterali importanti: l'incremento della variabilità alle basse intensità e l'esclusione di geni poco espressi, ma biologicamente importanti, dalla successiva analisi.

Esistono dei metodi grafici, gli *M*-plot diagnostici illustrati nel capitolo 3, per valutare se l'operazione di sottrazione sia indispensabile, ma per ammissione degli stessi autori [27] tali metodi riescono a diagnosticare male il caso in cui ci siano effetti spaziali di accumulo del "background", che sono proprio le situazioni in cui l'uso appropriato delle tecniche di sottrazione può contribuire a migliorare notevolmente il dato.

Alcuni autori [83] sono arrivati a dimostrare che su alcune piattaforme tecnologiche di microarray, quali quella Agilent, la sottrazione del "background"

è addirittura deleteria, quando effettuata con metodi che incrementano la varianza alle basse intensità.

L'uso di nuovi e più complessi metodi di sottrazione, quali *Normexp + offset*, ha consentito di limitare l'effetto di incremento di varianza, come si è potuto verificare nell'esperimento E3, contenendo in questo modo gli effetti collaterali di questa operazione.

Inoltre, l'uso di tecniche di scansione del vetrino quali la XDR di Agilent, ha ampliato notevolmente il range dinamico di intensità rilevabili sull'array, consentendo di amplificare il segnale di geni poco espressi senza peggiorare la rivelazione alle alte intensità.

L'uso dei metodi di normalizzazione ha consentito di eliminare gli effetti dovuti ad efficienze non ottimali di marcatura dei campioni o al non perfetto bilanciamento del sistema di scansione, o, ancora, ad una resa di fluorescenza differente fra i due fluorofori, cioè a tutti quei fenomeni che modificano l'intensità del segnale in maniera non dipendente dalla concentrazione di mRNA marcato nel campione.

Negli esperimenti descritti in questa tesi questi effetti hanno sempre mostrato un andamento non lineare e una prevalenza di un canale sull'altro.

L'applicazione del metodo *LOESS* ha migliorato notevolmente i dati, eliminando l'andamento non lineare pressochè prevalente alle basse intensità e riportando la tendenza centrale della distribuzione logaritmica dei dati sullo zero.

L'uso di metodi di normalizzazione *between array* è stato limitato ai casi in cui si è osservata una reale discrepanza fra le distribuzioni dei dati relativi alle diverse copie biologiche e ha prodotto l'effetto di rendere i dati maggiormente confrontabili durante la successiva analisi statistica. Questo ulteriore grado di normalizzazione ha migliorato la valutazione della varianza delle osservazioni, consentendo probabilmente di non perdere la significatività statistica dei dati, come osservabile nel caso dell'analisi statistica dell'esperimento E2 effettuata con *MAANOVA*, che non consente di fare normalizzazione *between array*.

La scelta dei metodi statistici è stata guidata dalla loro capacità di gestire dati provenienti da esperimenti di modeste dimensioni.

LIMMA è l'unico fra i tre metodi selezionati a mettere a disposizione numerose funzioni per la sottrazione del "background", efficaci strumenti per la normalizzazione *within* e *between array* e per la gestione dei più disparati disegni sperimentali. Proprio per la sua completezza è stato utilizzato come metodo principale di analisi. Inoltre, *LIMMA* fornisce molte possibilità per monitorare gli effetti dell'applicazione dei metodi di pre-trattamento dei dati

grazie ad un'ampia gamma di grafici. Infine, l'applicazione di un valido "framework" bayesiano, che valuta la varianza di ciascun gene sulla base dei dati provenienti dagli altri geni, consente di incrementare i gradi di libertà per il calcolo del p-value e di migliorare l'affidabilità dell'analisi statistica anche in esperimenti con bassa numerosità campionaria [84].

SAM è anch'esso uno strumento valido per l'analisi dei dati di esperimenti di dimensione ridotta, perché compensa l'assenza di dati reali con i dati fittizi derivanti dalle iterazioni.

Il metodo di analisi della significatività statistica utilizza un concetto intuitivo come quello del valore di soglia, irrobustito da una valutazione statistica del risultato ottenuto.

Il criterio che individua i due valori di soglia in *SAM* è di tipo iterativo, permettendo un maggiore controllo sui dati e sulle loro fluttuazioni; queste ultime vengono catturate attraverso le permutazioni.

Il processo di permutazione dei dati consente, infatti, di considerare i due insiemi di geni sovraespressi e sottoespressi come "indipendenti", ossia è possibile ricavare per essi due valori di soglia che possono portare ad avere un intervallo non simmetrico sul *SAM plot*. Ciò è conseguenza del fatto che l'espressione differenziale non si manifesta necessariamente con la stessa intensità relativa sui geni sovraespressi e sottoespressi, ma può succedere che il "fold-change" minimo per affermare la presenza di espressione differenziale non sia lo stesso.

SAM, tuttavia, risulta estremamente conservativo nel caso in cui si hanno a disposizione pochissimi array. Per esempio, *SAM* non ha fornito alcun risultato nella valutazione della lista di geni differenzialmente espressi per l'unica mutazione del gruppo P dell'esperimento E2, dove erano disponibili solo le osservazioni provenienti dai due array sui quali la mutazione A1789T veniva confrontata in "dye-swap" con il "wild-type". *LIMMA*, al contrario, ha identificato 740 geni, di cui alcuni sono stati successivamente validati utilizzando la real time RT-PCR.

Nel caso opposto, ossia con campioni molto numerosi o con molti geni per microarray, il carico computazionale potrebbe divenire estremamente oneroso e risultare ingestibile se non si dispone di un adeguato supporto hardware.

Una grossa limitazione che ha evidenziato *SAM* è l'assenza di metodi per il pre-trattamento e la selezione di qualità dei dati, per i quali è stato necessario avvalersi di quelli forniti da *LIMMA*.

L'utilizzo, infine, di un parametro statistico come il False Discovery Rate permette un'immediata stima del livello di affidabilità dell'insieme di geni

selezionati come differenzialmente espressi, evidenziando la percentuale di errori di tipo I che si commette selezionando soglie diverse.

Infine, il pacchetto statistico *MAANOVA* ha dimostrato di fornire strumenti efficaci per la rimozione degli errori sistematici da ciascun array, sebbene non presenti alcuna possibilità di normalizzazione *between array*. Quest'ultimo aspetto ha probabilmente influenzato i risultati dell'esperimento E2, nel quale le liste di geni differenzialmente espressi ricavate utilizzando *MAANOVA* sono assai poco numerose: l'impossibilità di tener conto delle repliche per la normalizzazione non ha consentito di moderare la varianza dei dati, facendo perdere significatività statistica ai risultati. L'alto numero di osservazioni necessarie all'interpolazione di modelli più completi di quelli utilizzati negli esperimenti illustrati per migliorare lo studio della varianza dei dati e l'eliminazione delle sue fonti, insieme all'eccessiva essenzialità e alla difficoltà nell'uso del pacchetto, ha posto l'utilizzo di *MAANOVA* in coda alle preferenze per l'analisi statistica dei dati.

L'interpretazione dei risultati di un esperimento microarray di espressione genica è ancora oggi un'operazione legata alla capacità del biologo di reperire informazioni spesso frammentate, organizzarle e formulare ipotesi biologiche partendo da esse.

Sebbene molti strumenti di analisi di "pathway" siano stati messi a punto nell'intento di realizzare l'interpretazione automatica dei risultati, nessuno di essi è, allo stato attuale, in grado di sostituire le capacità umane nel portare a termine questo aspetto dell'esperimento. L'uso di *Pathway Express* e *Pathway Explorer* ha consentito di evidenziare in prima istanza alcuni dei temi dell'interpretazione, ma, a causa dei limiti intrinseci degli strumenti e dell'assenza di informazioni catalogate coerentemente in KEGG, è stato fondamentale completare la raccolta di elementi utili alla ricostruzione dei meccanismi molecolari con l'attenta ispezione manuale delle liste di geni differenzialmente espressi e la consultazione delle banche dati geniche e di letteratura specifica.

Ringraziamenti

Un sincero ringraziamento va alla Dottoressa Silvia Pellegrini. Abbiamo intrapreso insieme questo percorso di crescita scientifica e di vita e, grazie al continuo stimolo, alla disponibilità all'ascolto e al confronto che ha sempre dimostrato nei miei confronti, oggi mi sento più "grande" sia come componente del nostro gruppo di ricerca che come persona e "donna".

Insieme alla Dottoressa, non posso che ringraziare anche tutte le sue collaboratrici e mie colleghe. Prima fra tutte voglio ringraziare Veronica, con la quale ormai si è instaurato un sincero e, spero, duraturo rapporto basato sulla stima, l'arricchimento, la comprensione e l'affetto reciproci: usualmente viene da me definita "la mia mente biologica, ma i nostri scambi vanno ben al di là del lavoro. Un affettuoso grazie va anche a Caterina, che con la sua singolare ironia rende speciali i nostri momenti di condivisione, e alla "new-entry" Manuela, con la quale mi auguro di poter costruire una proficua e non soltanto professionale "collaborazione".

Non posso dimenticare Lorenzo, Emiliano e Daniela, per l'immancabile aiuto nella soluzione di innumerevoli problemi e l'affetto che mi dimostrano da sempre.

Ai miei genitori e a mia sorella Noemi va la mia gratitudine per il supporto non soltanto morale durante questo impegnativo percorso: la loro incondizionata fiducia nei miei confronti mi ha sorretto ed accompagnato fino al raggiungimento di questo nuovo traguardo.

A mio marito Alfio, ormai maestro nell'esercizio della virtù della pazienza, che mi ha affiancato con lealtà e amore e mi ha sostenuto nella decisione di intraprendere questo percorso, riservo il mio ringraziamento più speciale: grazie per aver condiviso con me ogni momento!

Bibliografia

1. Collins, F.S., M. Morgan, and A. Patrinos, *The Human Genome Project: lessons from large-scale biology.* Science, 2003. **300**(5617): p. 286-90.
2. Lee, M.L., et al., *Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations.* Proc Natl Acad Sci U S A, 2000. **97**(18): p. 9834-9.
3. Churchill, G.A., *Fundamentals of experimental design for cDNA microarrays.* Nat Genet, 2002. **32 Suppl**: p. 490-5.
4. Dobbin, K. and R. Simon, *Comparison of microarray designs for class comparison and class discovery.* Bioinformatics, 2002. **18**(11): p. 1438-45.
5. Simon, R., M.D. Radmacher, and K. Dobbin, *Design of studies using DNA microarrays.* Genet Epidemiol, 2002. **23**(1): p. 21-36.
6. Yang, Y.H. and T. Speed, *Design issues for cDNA microarray experiments.* Nat Rev Genet, 2002. **3**(8): p. 579-88.
7. Dobbin, K., J.H. Shih, and R. Simon, *Questions and answers on design of dual-label microarrays for identifying differentially expressed genes.* J Natl Cancer Inst, 2003. **95**(18): p. 1362-9.
8. McShane, L.M., J.H. Shih, and A.M. Michalowska, *Statistical issues in the design and analysis of gene expression microarray studies of animal models.* J Mammary Gland Biol Neoplasia, 2003. **8**(3): p. 359-74.
9. Smyth, G.K., Y.H. Yang, and T. Speed, *Statistical issues in cDNA microarray data analysis.* Methods Mol Biol, 2003. **224**: p. 111-36.
10. Yang, M.C., et al., *Microarray experimental design: power and sample size considerations.* Physiol Genomics, 2003. **16**(1): p. 24-8.
11. Dobbin, K. and R. Simon, *Sample size determination in microarray experiments for class comparison and prognostic classification.* Biostatistics, 2005. **6**(1): p. 27-38.
12. Bueno Filho, J.S., S.G. Gilmour, and G.J. Rosa, *Design of microarray experiments for genetical genomics studies.* Genetics, 2006. **174**(2): p. 945-57.
13. Kerr, M.K. and G.A. Churchill, *Statistical design and the analysis of gene expression microarray data.* Genet Res, 2007. **89**(5-6): p. 509-14.
14. Sanchez, P.S. and G.F. Glonek, *Optimal designs for 2-color microarray experiments.* Biostatistics, 2009. **10**(3): p. 561-74.
15. Armstrong, N.J. and M.A. van de Wiel, *Microarray data analysis: from hypotheses to conclusions using gene expression data.* Cell Oncol, 2004. **26**(5-6): p. 279-90.
16. Jafari, P. and F. Azuaje, *An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors.* BMC Med Inform Decis Mak, 2006. **6**: p. 27.
17. Olson, N.E., *The microarray data analysis process: from raw data to biological significance.* NeuroRx, 2006. **3**(3): p. 373-83.
18. Grant, G.R., E. Manduchi, and C.J. Stoeckert, Jr., *Analysis and management of microarray gene expression data.* Curr Protoc Mol Biol, 2007. **Chapter 19**: p. Unit 19 6.

19. Dondrup, M., et al., *An evaluation framework for statistical tests on microarray data.* J Biotechnol, 2009. **140**(1-2): p. 18-26.
20. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
21. Smyth, G., *Limma: linear models for microarray data.*, in. *Bioinformatics and Computational Biology Solutions using R and Bioconductor.* 2005, Springer: New York. p. 397-420.
22. Wu, H., K. Kerr, and G. Churchill, *MAANOVA: a software package for the analysis of spotted cDNA microarray experiments.*, in *The Analysis of Gene Expression Data: An Overview of Methods and Software.* 2003, Springer: New York. p. 313-431.
23. Kendziorski, C.M., et al., *The efficiency of pooling mRNA in microarray experiments.* Biostatistics, 2003. **4**(3): p. 465-77.
24. Kerr, M.K. and G.A. Churchill, *Experimental design for gene expression microarrays.* Biostatistics, 2001. **2**(2): p. 183-201.
25. Dobbin, K., J.H. Shih, and R. Simon, *Statistical design of reverse dye microarrays.* Bioinformatics, 2003. **19**(7): p. 803-10.
26. Cui, X. and G.A. Churchill, *Statistical tests for differential expression in cDNA microarray experiments.* Genome Biol, 2003. **4**(4): p. 210.
27. Scharpf, R.B., et al., *When should one subtract background fluorescence in 2-color microarrays?* Biostatistics, 2007. **8**(4): p. 695-707.
28. Southern, E.M., *DNA microarrays. History and overview.* Methods Mol Biol, 2001. **170**: p. 1-15.
29. Schena, M., *Microarray Biochip Technology*. 2000, Westborough, MA: BioTechniques Press.
30. Hardiman, G., *Microarray technologies -- an overview. The University of California San Diego Extension, Bioscience, Microarray Technologies -- an overview, March 13-15, 2002.* Pharmacogenomics, 2002. **3**(3): p. 293-7.
31. Brown, C.S., P.C. Goodwin, and P.K. Sorger, *Image metrics in the statistical analysis of DNA microarray data.* Proc Natl Acad Sci U S A, 2001. **98**(16): p. 8944-9.
32. Martinez, M.J., et al., *Identification and removal of contaminating fluorescence from commercial and in-house printed DNA microarrays.* Nucleic Acids Res, 2003. **31**(4): p. e18.
33. Kooperberg, C., et al., *Improved background correction for spotted DNA microarrays.* J Comput Biol, 2002. **9**(1): p. 55-66.
34. Ritchie, M.E., et al., *A comparison of background correction methods for two-colour microarrays.* Bioinformatics, 2007. **23**(20): p. 2700-7.
35. Silver, J.D., M.E. Ritchie, and G.K. Smyth, *Microarray background correction: maximum likelihood estimation for the normal-exponential convolution.* Biostatistics, 2009. **10**(2): p. 352-63.
36. Smyth, G., *Limma: linear models for microarray data*, in *Bioinformatics and Computational*
Biology Solutions using R and Bioconductor, Gentleman R, et al., Editors. 2005, Springer: New York. p. 397-420.
37. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol, 2004. **5**(10): p. R80.
38. Schuchhardt, J., et al., *Normalization strategies for cDNA microarrays.* Nucleic Acids Res, 2000. **28**(10): p. E47.

39. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
40. McGee, M. and Z. Chen, *Parameter estimation for the exponential-normal convolution model for background correction of affymetrix GeneChip data.* Stat Appl Genet Mol Biol, 2006. **5**: p. Article24.
41. Yang, Y.H., et al., *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.* Nucleic Acids Res, 2002. **30**(4): p. e15.
42. Yang, Y.H. and N. Thorne, *Normalization for Two-color cDNA Microarray Data*, in *IMS Lecture Notes, Monograph Series*, T. Speed and D. Goldstein, Editors. 2003. p. 403-418.
43. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.
44. Kerr, M.K., M. Martin, and G.A. Churchill, *Analysis of variance for gene expression microarray data.* J Comput Biol, 2000. **7**(6): p. 819-37.
45. Kerr, M., et al., *Sources of Variation in Microarray Experiments*, in *Computational and Statistical Approaches to Genomics*, W. Zhang and I. Shmulevich, Editors. 2002, Kluwer Academic Publishers. p. 41-51.
46. Bayes, T. and R. Price, *An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S.* Philosophical Transactions of the Royal Society of London, 1763. **53**: p. 370-418.
47. Wolfinger, R.D., et al., *Assessing gene significance from cDNA microarray expression data via mixed models.* J Comput Biol, 2001. **8**(6): p. 625-37.
48. Bilofsky, H.S., et al., *The GenBank genetic sequence databank.* Nucleic Acids Res, 1986. **14**(1): p. 1-4.
49. Miller, G., R. Fuchs, and E. Lai, *IMAGE cDNA clones, UniGene clustering, and ACeDB: an integrated resource for expressed sequence information.* Genome Res, 1997. **7**(10): p. 1027-32.
50. Marchler-Bauer, A., et al., *MMDB: Entrez's 3D structure database.* Nucleic Acids Res, 1999. **27**(1): p. 240-3.
51. Hubbard, T., et al., *The Ensembl genome database project.* Nucleic Acids Res, 2002. **30**(1): p. 38-41.
52. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res, 1999. **27**(1): p. 29-34.
53. Brandt, K.A., *The GDB Human Genome Data Base: a source of integrated genetic mapping and disease data.* Bull Med Libr Assoc, 1993. **81**(3): p. 285-92.
54. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2001. **29**(1): p. 11-6.
55. Gene Ontology Consortium, *Creating the gene ontology resource: design and implementation.* Genome Res, 2001. **11**(8): p. 1425-33.
56. Rebhan, M., et al., *GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support.* Bioinformatics, 1998. **14**(8): p. 656-64.
57. Eyre, T.A., et al., *The HUGO Gene Nomenclature Database, 2006 updates.* Nucleic Acids Res, 2006. **34**(Database issue): p. D319-21.
58. Boutet, E., et al., *UniProtKB/Swiss-Prot.* Methods Mol Biol, 2007. **406**: p. 89-112.

59. Hewett, M., et al., *PharmGKB: the Pharmacogenetics Knowledge Base.* Nucleic Acids Res, 2002. **30**(1): p. 163-5.
60. Mlecnik, B., et al., *PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W633-7.
61. Draghici, S., et al., *A systems biology approach for pathway level analysis.* Genome Res, 2007. **17**(10): p. 1537-45.
62. Brazma, A., et al., *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.* Nat Genet, 2001. **29**(4): p. 365-71.
63. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res, 2002. **30**(1): p. 207-10.
64. Brazma, A., et al., *ArrayExpress--a public repository for microarray gene expression data at the EBI.* Nucleic Acids Res, 2003. **31**(1): p. 68-71.
65. Sturn, A., J. Quackenbush, and Z. Trajanoski, *Genesis: cluster analysis of microarray data.* Bioinformatics, 2002. **18**(1): p. 207-8.
66. Mariotti, V., et al., *Effect of prolonged phenytoin administration on rat brain gene expression assessed by DNA microarrays.* Experimental Biology and Medicine, in press.
67. Mishory, A., et al., *Phenytoin as an antimanic anticonvulsant: a controlled study.* Am J Psychiatry, 2000. **157**(3): p. 463-5.
68. Mishory, A., M. Winokur, and Y. Bersudsky, *Prophylactic effect of phenytoin in bipolar disorder: a controlled study.* Bipolar Disord, 2003. **5**(6): p. 464-7.
69. Nemets, B., Y. Bersudsky, and R.H. Belmaker, *Controlled double-blind trial of phenytoin vs. fluoxetine in major depressive disorder.* J Clin Psychiatry, 2005. **66**(5): p. 586-90.
70. Bersudsky, Y., *Phenytoin: an anti-bipolar anticonvulsant?* Int J Neuropsychopharmacol, 2006. **9**(4): p. 479-84.
71. Di Cecco, L., et al., *Characterisation of gene expression profiles of yeast cells expressing BRCA1 missense variants.* Eur J Cancer, 2009. **45**(12): p. 2187-96.
72. Monteiro, A.N., A. August, and H. Hanafusa, *Evidence for a transcriptional activation function of BRCA1 C-terminal region.* Proc Natl Acad Sci U S A, 1996. **93**(24): p. 13595-9.
73. Carvalho, M.A., et al., *Determination of cancer risk associated with germ line BRCA1 missense variants by functional analysis.* Cancer Res, 2007. **67**(4): p. 1494-501.
74. Caligo, M.A., et al., *A yeast recombination assay to characterize human BRCA1 missense variants of unknown pathological significance.* Hum Mutat, 2009. **30**(1): p. 123-33.
75. Zucchi, R., et al., *Trace amine-associated receptors and their ligands.* Br J Pharmacol, 2006. **149**(8): p. 967-78.
76. Scanlan, T.S., *Minireview: 3-Iodothyronamine (T1AM): a new player on the thyroid endocrine team?* Endocrinology, 2009. **150**(3): p. 1108-11.
77. Scanlan, T.S., et al., *3-Iodothyronamine is an endogenous and rapid-acting derivative of thyroid hormone.* Nat Med, 2004. **10**(6): p. 638-42.
78. Hart, M.E., et al., *Trace amine-associated receptor agonists: synthesis and evaluation of thyronamines and related analogues.* J Med Chem, 2006. **49**(3): p. 1101-12.

79. Ghelardoni, S., et al., *Modulation of Cardiac Ionic Homeostasis by 3-Iodothyronamine.* J Cell Mol Med, 2009.
80. Zucchi, R., S. Ghelardoni, and G. Chiellini, *Cardiac effects of thyronamines.* Heart Fail Rev, 2008.
81. Chiellini, G., et al., *Cardiac effects of 3-iodothyronamine: a new aminergic system modulating cardiac function.* Faseb J, 2007. **21**(7): p. 1597-608.
82. Braulke, L.J., et al., *3-Iodothyronamine: a novel hormone controlling the balance between glucose and lipid utilisation.* J Comp Physiol B, 2008. **178**(2): p. 167-77.
83. Zahurak, M., et al., *Pre-processing Agilent microarray data.* BMC Bioinformatics, 2007. **8**: p. 142.
84. Jeffery, I.B., D.G. Higgins, and A.C. Culhane, *Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data.* BMC Bioinformatics, 2006. **7**: p. 359.