UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA SPECIALISTICA
IN STATISTICA E INFORMATICA

TESI DI LAUREA

APPROCCIO BAYESIANO EMPIRICO PER L'IDENTIFICAZIONE DELLA PROPORZIONE DI GENI DIFFERENZIALMENTE ESPRESSI IN ESPERIMENTI DI *MICROARRAY*: STUDIO EMPIRICO DELLE PROPRIETÁ DELLE STIME

RELATORE: CH.MO PROF. ALBERTO ROVERATO

LAUREANDA: DANIELA DAMETTO

ANNO ACCADEMICO 2005-06

Indice

Introduzione

Una grande sfida che si presenta oggi ai ricercatori nel campo della biologia molecolare e della genetica è la caratterizzazione delle malattie genetiche, più precisamente delle anomalie del codice genetico che portano all'insorgenza di diverse patologie (ad esempio le leucemie, i tumori). Gli studiosi sono concentrati in questa direzione con l'obiettivo di mettere a punto terapie geniche, in grado di contrastare il progredire delle malattie attraverso un'azione mirata ai geni responsabili delle disfunzioni.
Le competenze fondamentali per raggiungere tali obiettivi partono dalla conoscenza approfondita dei protagonisti delle malattie: i geni. Al giorno d'oggi sono disponibili tecnologie avanzate di analisi, che permettono di monitorare il loro funzionamento in varie condizioni sperimentali. Queste tecnologie prendono il nome di *microarray* a DNA (microgriglie a DNA).

La tecnologia dei DNA *microarray* è in rapida crescita; dalla sua nascita negli anni Novanta, ha subìto uno sviluppo velocissimo, sia per quanto riguarda lo svolgimento tecnico dell'esperimento, sia riguardo l'analisi statistica dei dati. Di fatto, diversi metodi classici non portano a risultati significativi in questo campo, per questo sono attivi molti progetti di ricerca che estendono tecniche classiche e propongono nuovi espedienti finalizzati all'ottenimento di metodologie più robuste e specifiche. Lo stimolo creato dalla possibilità di studiare il comportamento di migliaia di geni simultaneamente ha dato un forte impulso di collaborazione a diverse scienze, quali biologia, genetica, informatica, statistica; sono richieste competenze sempre più complete e variegate per realizzare un esperimento così complesso, e queste esigenze danno origine a nuove professionalità, quali ad esempio il *bioinformatico, il biostatistico*.

Sull'onda di queste innovazioni, nel 1990 è stato ufficialmente inaugurato il **Progetto Genoma Umano**. Nato dalla collaborazione di diversi laboratori di ricerca in tutto il mondo, è finalizzato al sequenziamento del genoma umano;

per **sequenza del genoma umano** si intende la sequenze completa del DNA presente nei 22 autosomi più la coppia di cromosomi sessuali, sui quali sono distribuite $3{,}2 \times 10^9$ coppie nucleotidiche. Molte persone hanno fornito il loro DNA per il progetto, e ognuna differisce dall'altra, in media, per un nucleotide su 1000 (Alberts *et al.*, 2005), per cui la sequenza pubblicata del genoma umano è un mosaico di molte successioni individuali. Nel 2003, è terminato il sequenziamento, ed è stata fornita una mole di informazioni impressionante, che non ha precedenti in biologia. Il genoma umano ha 25 volte le dimensioni di ogni altro genoma sequenziato prima. Ci vorranno decenni per analizzare compiutamente l'informazione contenuta nel nostro genoma.

Un dato sorprendente è il numero di geni codificati: le prime stime si aggiravano nell'ordine dei 100.000 geni (Walter Gilbert, metà anni Ottanta), mentre nuove stime collocano il numero di geni vicino a 30.000, benché non sia ancora certo. È evidente, perciò, l'importanza che il fenomeno assume, e assumerà nel futuro, per la vita e la prosperità delle specie.

I motivi dell'applicazione di metodi statistici a questi esperimenti sono diversi: tra gli altri, l'identificazione di geni differenzialmente espressi sotto diverse condizioni sperimentali o tra soggetti che presentano varie forme della stessa patologia, l'individuazione di gruppi di geni co-regolati, la classificazione di campioni biologici.

Obiettivi dell'elaborato

La stima della proporzione di geni differenzialmente espressi - tra due o più campioni di cellule - è un obiettivo importante di alcuni studi recenti sull'analisi di dati da *microarray*. In particolare, nell'approccio bayesiano empirico sviluppato in Smyth (2004) e in Lönnstedt & Speed (2002), tale stima assume importanza cruciale nella definizione della statistica B, il cosiddetto *log posterior odd*. Lo scopo di questo elaborato è comprendere il ruolo del parametro che rappresenta questa quantità, che è denominato p, e, più nello specifico, valutare empiricamente le proprietà di due stimatori: uno proposto da Smyth (2004) e l'altro da Lönnstedt & Britton (2005). Questo si ottiene attraverso dati simulati da un'esemplificazione del modello bayesiano empirico sopra menzionato.
A completamento dell'analisi, si propone un'applicazione a dati che provengono da un esperimento reale.

Capitolo 1

La biologia molecolare e i *microarray*

In questo capitolo si presenta una descrizione degli esperimenti di *microarray*. Nei paragrafi 1.1 e 1.2 vengono esposte, rispettivamente, alcune nozioni sulla cellula e di biologia molecolare utili alla comprensione degli esperimenti, mentre nel par. 1.3 se ne tratteggiano le fasi principali. Nel par. 1.4 si descrivono alcuni degli sviluppi più recenti, in particolare il software LIMMA, ed infine, nel par. 1.5, si propone una riflessione sulla normalità dei dati in questione.

1.1 La cellula

La cellula è la più piccola unità di un organismo in grado di funzionare in modo autonomo. Tutti i viventi sono costituiti da una o più cellule: in base a questa caratteristica, possono essere suddivisi, rispettivamente, in organismi unicellulari e pluricellulari. Al primo gruppo appartengono, ad esempio, archebatteri, eubatteri, alghe azzurre; il secondo comprende le piante, gli animali e i funghi pluricellulari. Tutte le cellule sono accomunate da "orfanelli" (particolari organuli) e strutture tra cui la membrana esterna, il citoplasma e la molecola di DNA, che contiene il codice genetico dell'organismo a cui la cellula appartiene. Le cellule furono osservate per la prima volta nel 1665 da Robert Hooke, che

studiò con un microscopio rudimentale sottili fettine di sughero e vide che esse erano formate da elementi di forma regolare. Egli chiamò cellule questi elementi (dal latino cellula "piccola stanza"), perché esse avevano l'aspetto di piccole scatole. Nel 1830 Theodor Schwann compì studi al microscopio sulla cartilagine di animali e vide che questa era formata da cellule simili a quelle delle piante, e ipotizzò che le cellule fossero gli elementi costitutivi fondamentali di piante e animali; analoghe conclusioni trasse nel 1839 Matthias Schleiden. Nel 1860 Rudolf Virchow affermò che le cellule devono essere le "unità vitali" di tutti gli organismi, e che ogni cellula deriva da un'altra cellula.

Le cellule possono avere dimensioni e forme molto diverse. Quelle batteriche sono le più piccole, avendo una lunghezza dell'ordine di $1\mu m$ (un milionesimo di metro). Quelle dei tessuti animali hanno forma estremamente varia, a seconda del tipo e della funzione. In tutti i viventi, le cellule condividono alcune caratteristiche fondamentali; sono tutte delimitate da una membrana che racchiude il citoplasma. Questo è formato da un componente semifluido, il "citosol", contenente acqua, sali minerali e molecole organiche, in cui si trovano immerse strutture dette organuli o orfanelli, ciascuno preposto a una particolare funzione.

Le cellule sono capaci di riprodursi: ciascuna di esse si divide in due cellule figlie mediante un processo che prende il nome di "mitosi". La capacità di dividersi è differente in base al tipo cui esse appartengono. Si possono riconoscere tre categorie: cellule soggette al rinnovamento, che per tutta la vita dell'individuo vengono continuamente sostituite da cellule nuove; cellule in espansione, che smettono di dividersi quando l'individuo ha completato la sua crescita, ma che possono occasionalmente riprendere a dividersi come conseguenza di ferite o traumi; cellule statiche, che perdono la capacità di dividersi prima ancora che l'accrescimento dell'organismo sia completo. Alcune cellule nell'organismo mantengono la capacità di riprodursi per tutta la vita, e rimangono indifferenziate, potendo quindi dare luogo a diversi tipi cellulari: esse sono dette "staminali".

Procarioti ed eucarioti

Le cellule, in base alla loro organizzazione interna, possono essere distinte in due grandi categorie: cellule procariote e cellule eucariote. Il termine procariote deriva dal greco e significa "prima del nucleo"; il termine eucariote significa "vero nucleo".

Cellule procariote: struttura delle alghe azzurre

Le cellule dei procarioti (tra cui i batteri) mancano di molte delle strutture interne tipiche di quelle degli organismi eucarioti. Pur essendo dotate di membrana plasmatica ed eventuale parete cellulare, sono prive di membrana nucleare; la molecola di DNA circolare si trova, pertanto, libera nel citoplasma. Le cellule procariote sono tipiche degli archebatteri, degli eubatteri e delle alghe azzurre. Esse sono relativamente piccole (con un diametro generalmente compreso fra 1 e 5 μm) e hanno una struttura interna alquanto semplice; il loro DNA si trova concentrato in una regione del citoplasma, senza essere delimitato da alcuna membrana. Sono prive di organuli, ad eccezione dei ribosomi, le particelle preposte alla sintesi delle proteine. Le funzioni cellulari sono comunque effettuate da complessi enzimatici analoghi a quelli delle cellule eucariote. Gli organismi formati da cellule procariote sono detti procarioti.

Cellula eucariota

Negli eucarioti - ossia animali, piante e funghi - la cellula è caratterizzata da un nucleo, in cui è racchiuso il patrimonio genetico, e da organuli membranosi deputati allo svolgimento di specifiche funzioni. Queste strutture sono protette dalla massa gelatinosa del citoplasma e da un involucro detto "membrana plasmatica". Dunque, la cellula eucariote è suddivisa in zone funzionali in cui possono avvenire contemporaneamente reazioni metaboliche che richiedono differenti condizioni; per tale proprietà, definita "compartimentazione", risulta più efficiente delle cellule dei procarioti (batteri e alghe azzurre), prive di organuli e di nucleo. Rispetto al modello cellulare qui illustrato, tra gli eucarioti si possono

riscontrare diversità nel numero e nella effettiva presenza di tutti gli organuli: ad esempio, molte cellule fungine, così come le fibre muscolari umane, possiedono numerosi nuclei; cellule dotate di mobilità, come molti protisti e gameti, sono dotate di flagelli e ciglia; le cellule vegetali, inoltre, possiedono alcune strutture caratteristiche (parete, cloroplasti e vacuoli). Le cellule eucariote che costituiscono tutti gli altri organismi viventi (i protisti, le piante, i funghi e gli animali) sono molto più grandi (solitamente il loro asse maggiore è compreso fra i 10 e i 50 μm). Queste cellule possiedono organuli immersi nel citoplasma, ognuno deputato a svolgere una particolare funzione. Gli organismi formati da cellule eucariote sono detti eucarioti.

Nucleo cellulare

L'organulo di maggiori dimensioni all'interno di gran parte delle cellule vegetali e animali è il nucleo: è delimitato da una membrana e ha forma e dimensioni variabili a seconda del tipo cellulare. All'interno del nucleo si trovano il DNA, che costituisce il materiale genetico della cellula, e proteine (dette istoni) solitamente presenti in coppie, in un numero variabile e caratteristico di ciascuna specie.

1.2 Nozioni di biologia molecolare

La biologia molecolare è la disciplina che studia le molecole organiche complesse presenti nella cellula, in particolare DNA (acido desossiribonucleico), RNA (acido ribonucleico) e proteine, le principali protagoniste della vita organica, allo scopo di mettere in relazione la struttura di ciascuna con la funzione che essa svolge all'interno della cellula stessa e nell'ambito dell'organismo. Le modalità d'indagine della biologia molecolare applicano tecniche che provengono da varie altre discipline, quali la biochimica, la fisica e la genetica, e le scoperte trovano applicazione nell'ingegneria genetica (detta anche tecnologia del DNA ricombinante) e nella biotecnologia.

La fondamentale scoperta che ha segnato la nascita della biologia molecolare fu l'elaborazione, nel 1953, del modello tridimensionale dell'acido desossiribonucleico (DNA), a opera del biologo statunitense James Watson e del biofisico britannico Francis Crick. Il DNA ha due funzioni fondamentali: da un lato presiede alla conservazione e alla trasmissione dell'informazione genetica da una generazione alla successiva, dall'altro dirige la sintesi delle proteine, molecole necessarie sia alla costruzione che al funzionamento delle cellule. I meccanismi in base ai quali questa molecola presiede a questi processi fondamentali risultano evidenti dall'analisi della struttura.

Il DNA è una molecola a doppia elica, formata da due filamenti uniti l'uno all'altro da legami fra quattro subunità, le quali si ripetono in una sequenza variabile lungo tutta la molecola; queste subunità, dette azotate, comprendono l'adenina, la guanina, la citosina e la timina. Non tutti gli appaiamenti tra basi sono possibili: una A su un filamento si appaia sempre con una T sull'altro, mentre una G si appaia sempre con una C. Sequenze lineari di basi formano i geni, serie di geni formano i cromosomi. Prima di ciascuna divisione cellulare il DNA si duplica, in modo tale da trasmettere a ciascuna delle due cellule figlie una copia fedele del patrimonio genetico parentale. Nel corso di questo processo, l'informazione contenuta nella molecola viene trasportata in modo estremamente preciso: i due filamenti si separano e ciascuno di essi funge da stampo per la costruzione di un nuovo filamento appaiato al primo. Grazie a una procedura sperimentale messa a punto da Frederick Ranger nel 1977, è oggi possibile leggere la sequenza lineare delle basi di un frammento di DNA. Il metodo è stato utilizzato anche da tutti i laboratori coinvolti nel Progetto Genoma Umano che, nel 2003, hanno raggiunto l'obiettivo di identificare tutti la sequenza di nucleotidi presenti nel patrimonio genetico della nostra specie.

La biologia molecolare ha permesso di chiarire come avviene la sintesi delle proteine, che è un processo molto complesso e richiede due fasi fondamentali. La prima di queste fasi prende il nome di trascrizione e consiste nella copiatura di un gene (una porzione di DNA), contenente le istruzioni necessarie alla costruzione della specifica proteina, su un'altra molecola di acido nucleico a

singolo filamento, detta mRNA (RNA messaggero). Come nella duplicazione del DNA, anche il gene viene copiato fedelmente mediante l'appaiamento delle basi dell'mRNA sullo stampo fornito dalla porzione di DNA. La seconda fase, chiamata traduzione, prevede l'utilizzazione delle informazioni contenute nella molecola di mRNA per la sintesi di una proteina, che avviene nei ribosomi tramite l'unione di diversi amminoacidi. Il cosiddetto "dogma centrale della biologia molecolare" (fig. 1.1) stabilisce che il flusso delle informazioni passa dal DNA all'mRNA alle proteine. In particolare, capire quando, dove e in quale quantità ogni gene produce proteine equivale a studiare la sua "espressione genica".

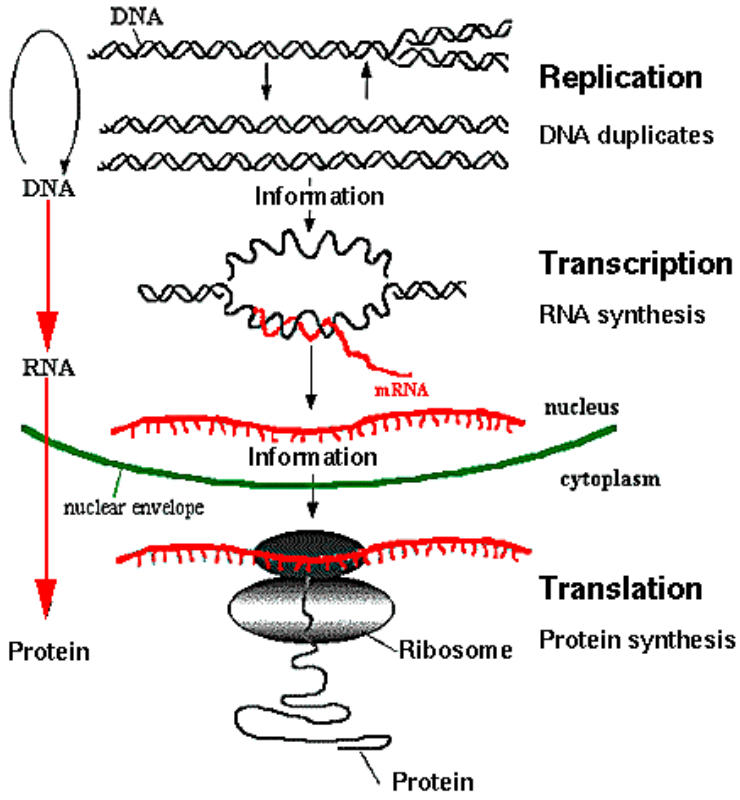

Figura 1.1: Dogma centrale della biologia molecolare.

Tutte le cellule di un organismo, di fatto, contengono lo stesso patrimonio genetico. È evidente, però, che non siano tutte uguali; basti pensare alle cellule cutanee, muscolari, del sangue: deve esserci un meccanismo interno che le rende così diverse. La chiave sta proprio nella produzione delle proteine: cellule differenti sono costituite da proteine differenti, a loro volta sintetizzate da geni differenti. Conoscendo la sequenza di basi del DNA, si può definire ciascun gene come sequenza lineare di basi, che a sua volta indica, in base al codice genetico, la sequenza di amminoacidi della proteina corrispondente.

Gli studi di biologia molecolare hanno permesso di stabilire l'esistenza di un codice genetico, ovvero di una serie di combinazioni di tre basi azotate (triplette) ciascuna corrispondente a un particolare amminoacido. Ad esempio, le triplette ACC e CCC sull'mRNA corrispondono rispettivamente agli amminoacidi treonina e prolina. Va sottolineato che le possibili triplette sono 64, mentre gli amminoacidi sono 20; ciò significa che triplette diverse codificano lo stesso amminoacido. Risulta ora chiaro come sia molto più facile ordinare le sequenze di basi del DNA che non quelle degli amminoacidi delle proteine; per questo motivo, di solito la sequenza amminoacidica di una proteina viene identificata indirettamente, a partire dalla sequenza del gene corrispondente. Quando si specifica la sequenza di un gene, che in alcuni individui è presente in una forma mutata e responsabile di una specifica malattia, dal confronto tra le sequenze del gene normale e del gene mutato è possibile individuare qual è la tripletta alterata. Le eventuali mutazioni della sequenza di basi di un gene, di conseguenza, vengono riportate anche nella struttura delle proteine. Ad esempio, una mutazioni da A a C nella tripletta ACC porta all'aggiunta nella proteina nascente di prolina al posto di treonina. Poiché proteine specifiche hanno effetti biologici specifici, le alterazioni di una sequenza amminoacidica che vanno ad interferire con la funzione della proteina possono riflettersi in modificazioni strutturali o funzionali a livello cellulare o dell'organismo. Quando le mutazioni avvengono nel DNA delle cellule germinali, esse vengono anche trasmesse alle generazioni successive; alcune, comunque, sono fisiologiche e sono responsabili di differenze innocue tra individui, come il colore degli occhi, della pelle o dei

capelli. Quando, invece, a causa di una mutazione genetica viene prodotta una proteina difettosa, l'individuo può essere affetto da una malattia genetica, come l'emofilia, che può essere trasmessa alla prole.

Le mutazioni genetiche

Una mutazione è una modificazione della normale struttura di un gene o di un cromosoma o di un cariotipo, che si verifica in modo improvviso e imprevedibile. Una mutazione può essere spontanea o indotta; in quest'ultimo caso, essa è determinata da fattori che prendono il nome di agenti mutageni. Sono agenti mutageni, ad esempio, fattori fisici come le radiazioni, fattori chimici come verie sostanze chimiche, e fattori biologici come alcuni retrovirus. Il primo scienziato che utilizzò il termine mutazione fu, nel 1901, il botanico olandese Hugo De Vries che, insieme ad altri, ebbe anche il merito di riportare alla luce il lavoro del monaco austriaco Gregor J. Mendel sulla trasmissione dei caratteri ereditari. Nel 1929 il biologo statunitense Hermann J. Muller osservò che i raggi X possono aumentare la frequenza delle mutazioni spontanee. In seguito, la lista delle sostanze che hanno questo effetto si allargò ad altre forme di radiazioni, a valori particolarmente elevati della temperatura e ad un gran numero di composti chimici.

Sebbene la duplicazione del DNA avvenga con un meccanismo estremamente preciso, essa non è sempre perfetta. Possono insorgere, infatti, degli errori, per cui il nuovo frammento di DNA contiene uno o più nucleotidi diversi dall'originale. Questi errori, che rappresentano appunto le mutazioni, possono avvenire in qualunque punto del DNA: se avvengono in una sequenza di DNA codificante per un particolare polipeptide (qualsiasi sostanza costituita da amminoacidi), nella catena polipeptidica si può avere la variazione di un singolo amminoacido o anche un'alterazione più grave della proteina risultante. L'anemia falciforme, ad esempio, è causata da una mutazione genetica che determina la sintesi di una molecola di emoglobina mutante, la quale differisce dalla forma normale per un singolo amminoacido. La frequenza di mutazione aumenta quando alcuni geni che codificano fattori proteici responsabili della fedeltà della duplicazione del

DNA, o della correzione degli errori, sono mutati a loro volta. La maggior parte delle mutazioni genetiche osservabili è silente, ossia non produce alcuna variazione che si manifesti a livello del fenotipo, cioè nell'aspetto esterno dell'individuo. Raramente le mutazioni causano, invece, effetti a livello cellulare, che possono alterare in modo drammatico le funzioni generali dell'organismo. Le mutazioni non silenti compaiono generalmente in geni recessivi e, quindi, i loro effetti nocivi non sono osservabili se non sono presenti due geni mutati contemporaneamente, cioè se l'individuo non è omozigote per la mutazione. Questo accade più frequentemente nei casi di incrocio, ossia nell'accoppiamento di organismi strettamente imparentati, che possono aver ereditato lo stesso gene mutante recessivo da un comune antenato. Per questa ragione, le malattie ereditarie sono più frequenti nei bambini i cui genitori sono cugini o parenti stretti, che non nella popolazione umana generale.

1.3 L'esperimento di *microarray*

Nel campo dei *microarray* ci sono due grandi tipologie di esperimenti: a canale singolo e a due canali (cDNA). L'obiettivo comune è misurare quanto mRNA gene-specifico c'è in una cellula, per ogni gene d'interesse. Si suppone, infatti, che la quantità di mRNA presente nella cellula sia direttamente associabile al numero di proteine sintetizzate. Più mRNA è presente, più sarà indice che il gene è espresso, attivo in quella cellula; viceversa, meno mRNA è presente, meno sarà verosimile che il gene sia espresso; riuscire ad individuare esattamente quali geni possano essere messi in relazione con determinate patologie potrà portare alla preparazione di terapie geniche *adhoc*, mirate alla compensazione delle anomalie.

1.3.1 I *microarray* a cDNA

Un *microarray* (microgriglia) a cDNA è una piastrina di vetro o silicio costituita da moltissime sonde (*spot*), ognuna delle quali rappresenta un gene prestabilito. Ogni *spot* consiste di molte copie di singoli filamenti di DNA complementare

(cDNA), che viene preparato attraverso una trascrizione inversa dall'mRNA. In un *microarray* ci possono essere da 5.000 a 30.000 spots; negli esperimenti odierni, diversi *slides* vengono preparati allo stesso modo per ottenere osservazioni ripetute sullo stesso gene. La deposizione viene effettuata da sistemi robotizzati, che mediante l'utilizzo di pennini prelevano le sonde direttamente dalle piastre utilizzate per la PCR (reazione polimerasica a catena) e le depositano sul vetrino, formando *spots* di circa 100-150 μm di diametro, distanziati l'uno dall'altro 200-250 μm. Durante la collocazione, il sistema di controllo del robot registra automaticamente tutte le informazioni necessarie alla caratterizzazione ed alla completa identificazione di ciascun punto della matrice (identità del cDNA, coordinate sul supporto, ecc.). In ogni esperimento, diversi *slides* identici vengono preparati per ottenere osservazioni ripetute dello stesso gene. Purtroppo non è possibile controllare l'esatto numero di copie di cDNA stampate in ogni *spot* del *microarray*; per questo ed altri motivi, è impossibile misurare i livelli di espressione assoluta dei geni durante l'esperimento. Di conseguenza è necessario ottenere livelli di espressione relativa, monitorando il comportamento di due campioni di cellule (ad esempio trattate verso non trattate, malate verso sane). Per ognuno dei due campioni da studiare, viene isolato l'mRNA ed ogni sequenza viene trascritta in cDNA; questi frammenti vengono poi tinti con due colori fluorescenti (generalmente rosso, Cy5, per il campione trattato e verde, Cy3, per il non trattato) e vengono aggiunti sul *microarray* in uguale quantità. A questo punto, i campioni tenderanno ad "ibridarsi" con il cDNA preparato sul *microarray*; se un gene ha un livello di espressione maggiore nel campione trattato rispetto al non trattato, lo *spot* tenderà a presentarsi rosso (prevalenza di Cy5, gene "sovraespresso"), in caso contrario tenderà a presentarsi verde (prevalenza di Cy3, gene "sottoespresso"). Con il termine "espressione differenziale" ci si riferisce proprio a queste differenze nell'espressione dei geni. Le intensità verde e rossa vengono successivamente captate con un analizzatore di immagini ("*laser scanner*"), dal quale si ottengono due file diversi, uno per ognuna delle due tinte. Questi file vengono sovrapposti in modo da ottenere un risultato visibilmente interpretabile (fig. 1.2).

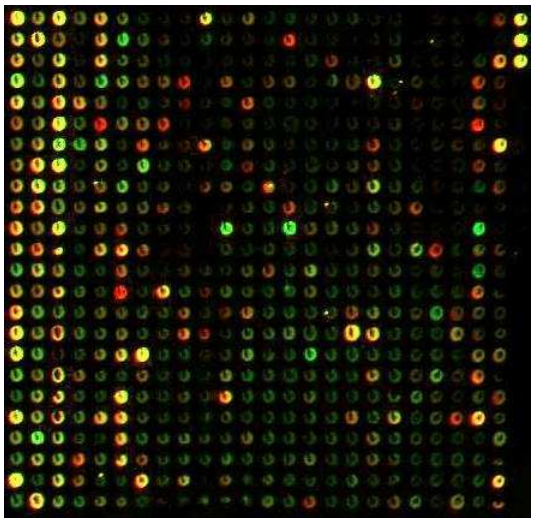

Figura 1.2: Esempio d'immagine ottenuta con i *microarray*.

Da un punto di vista statistico, l'analisi di un esperimento è divisa in varie fasi: la programmazione del disegno sperimentale, secondariamente, una volta scannerizzati i vetrini, l'analisi d'immagine, la normalizzazione, e infine si considera il problema biologico di fondo: testare, attraverso varie tecniche, quali geni sono differenzialmente espressi. Nel seguito, si farà riferimento specificatamente a questo tipo di esperimento.

1.3.2 I *microarray* a canale singolo

L'uso di *microarray* a canale singolo è in fase di grande sviluppo; in particolare, la ricerca è molto fertile sulla piattaforma *Affymetrix*. Si tratta di una tecnologia alternativa a quella dei cDNA, molto più sofisticata e costosa.

Diversamente dagli esperimenti a due canali, viene ibridato un solo campione. I vetrini vengono poi scannerizzati per ottenere un valore d'intensità per ogni *probe*, che misura l'ibridazione del corrispondente oligonucleotide. Questa tipologia di esperimenti non è, tuttavia, oggetto di questa tesi.

1.3.3 Il disegno sperimentale

Il compito della pianificazione statistica del disegno è minimizzare le fonti di variabilità nei dati per aumentare la precisione delle stime delle quantità d'interesse: in base al numero di *slides* disponibili e alla quantità di mRNA che si possiede, la pianificazione del disegno serve per trasformare, quanto possibile, errori sistematici non controllabili in errori pianificati ed eliminabili. La tabella 1.1 descrive le fonti di errore tipiche di un esperimento di *microarray* a due canali.

Nello svolgimento di un esperimento ci si trova di fronte ad alcune questioni pratiche, quali ad esempio; è meglio utilizzare tutti *slides* dello stesso lotto oppure no? È meglio utilizzare diversi metodi di scansione dell'immagine o uno solo? Per ottenere le risposte è necessario basarsi su principi generali di costruzione del disegno; prima, però, occorre chiarire il significato di alcuni concetti.
Il termine "trattamento" o "condizione" indica qualsiasi attributo di interesse primario nell'esperimento. L'unità statistica è ogni ripetizione indipendente soggetta al trattamento; è comune considerare i geni come unità statistiche. Un fattore "blocco" è una condizione che ha effetto sull'esperimento, ma non è di particolare interesse.

Tornando ai principi di costruzione del disegno sperimentale, Cobb (1998) e Draghici *et al.* (2001) suggeriscono le seguenti idee generali:

- suddividere le unità in blocchi, ossia gruppi di unità simili, poi assegnare i trattamenti separatamente all'interno di ciascun blocco;
- assegnare i trattamenti alle unità statistiche in modo casuale;
- incrociare le condizioni sperimentali sulle unità statistiche per poterne confrontare gli effetti.

Qualsiasi disegno sperimentale di *microarray* dev'essere basato su confronti a coppie di campioni di cellule, poichè, come già sottolineato, può essere misurato

Fonti di variabilità nei campioni di mRNA
Differenze nelle condizioni dei campioni
Differenze tra i soggetti raggruppati nello stesso blocco
Differenze dello stesso gene tra i campioni
Variazioni nei metodi di estrazione dell'mRNA
Variazioni durante la trascrizione
Differenze nell'applicazione delle tinte
Fonti di variabilità nella produzione dei microarray
Anomalie degli aghi di stampa
Variazioni nelle quantità stampate anche con lo stesso ago
Variazioni tra lotti di vetrini
Differenze nella lunghezza dei frammenti di DNA
Variazioni del livello di attaccamento allo *slide* fra i vari geni
Fonti di variabilità nel processo di ibridazione
Differenze tra le tinte
Disuguaglianze nell'applicazione dell'mRNA agli *slides*
Altre differenze in parametri come temperatura, sperimentatore, ora del giorno
Fonti di variabilità nella scansione
Differenti laser-scanner
Differenti software di analisi
Diverso allineamento della griglia di punti

Tabella 1.1: Fonti di variabilità in un esperimento di *microarray*

solamente il livello relativo di espressione. È comune utilizzare la notazione di Kerr & Churchill (2001) per descrivere il disegno (fig. 2.1).

Una questione importante da discutere riguarda i confronti diretti e indiretti tra campioni; si tratta, come spesso accade, di trovare un compromesso accettabile tra risorse disponibili ed esigenze informative. Solitamente, se si dispone di due campioni da confrontare e si possiede mRNA sufficiente per due ibridazioni, è opportuno scegliere il disegno diretto, ossia utilizzare due *slides* invertendo le tinte. Se, invece, si ha la possibilità di affrontare una sola ibridazione, o se si vogliono ottenere risultati direttamente confrontabili con altri esperimenti attraverso un campione di riferimento, è chiaro che si sceglierà un disegno indiretto.
Altri riferimenti sul disegno sperimentale si trovano in Glonek & Solomon (2004), Wolfinger *et al.* (2001).

1.3.4 L'analisi d'immagine

I primi esperimenti di *microarray* venivano effettuati con un singolo vetrino; l'analisi statistica si concentrava sul confronto tra le intensità rossa e verde, spesso indicate con R_g e G_g per il gene g. Oggi si utilizzano diverse replicazioni degli stessi geni e quindi vari *slides*: di conseguenza, per poterli confrontare, è necessario prestare particolare attenzione in questa fase, cercando quanto possibile di mantenere costanti le condizioni di lettura delle immagini. L'obiettivo è estrarre una coppia di intensità per ogni tinta e per ogni gene: quella di primo piano (*foreground*) e quella di fondo (*background*). In realtà, il valore che interessa per l'analisi è quello che emerge in superficie del vetrino; l'intensità sottostante è di disturbo e va stimata per essere in qualche modo eliminata.

La prima fase dell'analisi, detta di "filtraggio", serve a rimuovere piccole contaminazioni dell'immagine, dovute a polvere, ad esempio, e ad eliminare eventuali rumori di fondo o interfenze. Successivamente, si determina, almeno approssimativamente, il centro di ogni *spot* dell'immagine. Questa fase è detta "indirizzamento". Segue la "segmentazione", che ha come obiettivo l'identificazione dei bordi delle sonde: i pixel vengono classificati come oggetto dell'ana-

lisi o come sfondo a seconda che siano parte di uno *spot* oppure no. Attraverso un processo di "quantificazione", si ricavano le coppie di valori d'intensità (superficiale e di fondo) per ogni gene. Per il calcolo delle intensità in primo piano, generalmente, si utilizza la media dei pixel appartenenti agli *spots*, mentre per quelle di fondo si usa la più robusta mediana. Si ottiene così un valore tra 0 e 2^{16}; per comodità si considera il logaritmo in base 2, in modo da riportare i valori in una scala da 0 a 16.

1.3.5 La normalizzazione

Per poter confrontare diversi *slides* tra di loro, e rimuovere eventuali errori sistematici nelle misurazioni, si rende necessario un processo di "aggiustamento" dei dati. Tale processo prende il nome di normalizzazione.

Un problema comune nelle misurazioni di quantità ottiche è l'esistenza di un segnale di fondo permanente. Sono stati proposti vari metodi per contrastare questo effetto (Wit & McClure, 2004; Yang *et al.*, 2002a).
La maggior parte degli autori assume che il segnale sia di tipo additivo, ossia che il segnale osservato S si possa scrivere come somma dell'effetto *background* B e del vero segnale T,

$$S = B + T.$$

Le intensità finali si otterrebbero, quindi, sottraendo il valore di *background* da quello di superficie, ossia attraverso il procedimento di *local background substraction*; questo dovrebbe consentire di correggere variazioni locali dovute alla non uniformità dei vetrini. Sfortunatamente, però, questo effetto di fondo non può essere misurato, se non nelle vicinanze degli *spots*. Si possono utilizzare vari espedienti per risolvere il problema, ma la questione principale rimane l'assunzione che questi due segnali siano additivi. È stato infatti riconosciuto che il DNA negli spot può effettivamente mascherare questo rumore di fondo.
Wit & McClure consigliano di non sottrarre il *background* al segnale osservato, in quanto il valore ricavato dalla differenza di due stime (spesso distorte) ha

maggiore variabilità del segnale stesso.

La necessità di un'ulteriore normalizzazione deriva dalla tecnologia odierna dei *microarray* a due canali, che si basa sul confronto fra intensità di frammenti di cDNA tinto; i due colori più comunemente usati sono il rosso (Cy5) e il verde (Cy3). Queste tinte hanno proprietà leggermente differenti; dalla grandezza delle molecole, alla capacità di reazione allo sbiancamento fotografico, un effetto che consegue allo *scanning* multiplo degli *array*. Il risultato è una diversità nell'efficienza dei due canali di un *array*, che rende difficili i confronti dei dati di espressione.
Principalmente, sono stati suggeriti due metodi: seguendo il primo, si stima l'efficienza relativa di ogni tinta e la si sottrae dai dati, mentre il secondo, che è conosciuto come *dye-swap*, consiste nel ripetere due volte un'ibridazione, scambiando le tinte, e considerare la media - o qualche altra statistica - delle due espressioni di colore diverso, per ogni gene. Purtroppo non ci sono garanzie che tale metodo rimuova completamente la distorsione.

Lönnstedt & Speed suggeriscono un'analisi dei dati basata principalmente su metodi grafici, e propongono di utilizzare come misura del livello di espressione relativa il log-rapporto (o valore M) dato da

$$M_g = \log_2 \frac{R_g}{G_g}.$$

Nelle analisi successive, per comodità, si considera il logaritmo in base 2 dei valori ottenuti dall'analisi d'immagine.
Si procede analizzando il cosiddetto MA-plot, con

$$A = (\log_2 R_g + \log_2 G_g) \times \frac{1}{2},$$

che rappresenta l'intensità media. Anche con questo metodo emerge la distorsione dovuta alle tinte, attraverso una chiara dipendenza tra i valori M ed A (fig. 4.1). La soluzione proposta è la costruzione di una funzione di lisciamento l

(*lowess*) sulla nuvola di punti nel grafico MA e la sottrazione dei valori ottenuti dagli M. I nuovi log-rapporti normalizzati sono dati da

$$M'_g = M_g - l(A_g).$$

Questo metodo prende il nome di *global lowess normalization* (si veda Yang *et al.*, 2002b).

Un altro tipo di errore che può influenzare l'esperimento è il cosiddetto *bias* spaziale, che condiziona i dati quando una tinta si ibrida meglio in una zona piuttosto che in un'altra del vetrino, oppure se qualche robot per la produzione degli *slides* viene danneggiato: Yang *et al.* (2002b) propongono una normalizzazione specifica anche in questo caso. Altre proposte per la normalizzazione si trovano in Kerr *et al.* (2000), Wolfinger *et al.* (2001), Huber *et al.* (2002), Yang & Thorne (2003).

1.4 Gli sviluppi più recenti

Al giorno d'oggi, i *microarray* sono utilizzati come strumenti all'interno di esperimenti complessi. La ricerca si sta largamente occupando, ad esempio, di espressione di gruppi di geni, piuttosto che di singoli. In queste analisi si seguono due strade, principalmente: ci si chiede se geni dichiarati differenzialmente espressi negli esperimenti singoli possano essere raggruppati secondo la locazione o attraverso qualche altra funzione, oppure si analizzano direttamente insiemi di geni predefiniti, selezionati, ad esempio, dal *database* GO, `www.geneontology.org`. Questo ricco *database* genetico è stato costruito seguendo sistemi di classificazione dei geni a diversi livelli, quali la loro funzione, il tessuto nel quale sono espressi, eccetera. La sua creazione ha richiesto una grossa concentrazione di competenze bioinformatiche negli ultimi anni; parallelamente, ci sono stati molti sforzi nell'implementazione di metodi d'analisi.

Oltre a software specifici, molte procedure sono state fornite come librerie

del software S. S è un linguaggio molto intuitivo e funzionale, che forma le basi del progetto gratuito **R** e del sistema commerciale **S-plus**. Ulteriore supporto all'analisi di *microarray* è disponibile nel progetto **Bioconductor** (`http://www.bioconductor.org`).

Tra le varie librerie disponibili, è nota LIMMA (*LInear Models for MicroArray data*), le cui funzionalità hanno sostituito la più datata libreria "sma" (*Statistical Microarray Analysis*).

L'idea dei *microarray* si sta espandendo verso nuovi obiettivi, come ChIP-chip (Buck & Lieb, 2004) per l'identificazione dei siti di assemblaggio delle proteine, lo studio del processo di traduzione, eccetera. Ognuno di essi offre nuovi spunti per le analisi statistiche, dalla normalizzazione, ai test per la verifica d'ipotesi. Per una rassegna dei software disponibili per l'analisi di *microarray* si veda Parmigiani *et al.* (2003), che fornisce riferimenti per tutte le fasi degli esperimenti, quali l'analisi d'immagine, la normalizzazione, i test, la classificazione, eccetera.

1.4.1 Il software LIMMA

LIMMA è un pacchetto per l'analisi di espressioni geniche provenienti da *microarray*, in particolare per l'uso dei modelli lineari nell'analisi degli esperimenti e l'identificazione dei geni differenzialmente espressi (mediante un metodo bayesiano empirico). Ha caratteristiche che rendono ogni analisi stabile, anche per un numero esiguo di *slides*, poiché sfrutta informazioni provenienti dall'intero insieme dei geni. Alcune funzioni, tra cui quelle di normalizzazione ed analisi esplorativa, sono costruite per esperimenti a cDNA; i modelli lineari e le funzioni di determinazione dell'espressione differenziale, invece, si applicano anche ad esperimenti con un unico canale (o *Affymetrix*). È possibile scaricare il pacchetto all'indirizzo *web*: `http://bioinf.wehi.edu.au/limma`. Per un'introduzione alle principali peculiarità del programma, è disponibile una guida all'URL `http://cran.r-project.org/doc/packages/limma.pdf`. Inoltre, si

può consultare il file di *help* in linea di R, per una descrizione dettagliata delle funzioni.

In questo elaborato viene usato questo software principalmente perchè l'analisi si basa sul modello di Smyth (2004), che è stato da lui stesso implementato nella libreria. Non di meno, è uno dei pacchetti più usati per dati da *microarray*.

1.5 I dati da *microarray*: sono gaussiani oppure no?

Gran parte della teoria statistica è costruita attorno all'idea che i dati abbiano una componente di casualità, ossia, seguano una certa distribuzione. La più comune distribuzione di probabilità è la normale, o gaussiana; se i dati sono normali, è possibile sfruttare utili risultati teorici che semplificano lo studio. Per questo motivo, molto spesso si assume che i dati da *microarray* siano normali e, allo stesso tempo, sono state spese moltissime energie per dimostrarlo, in particolare per la distribuzione del logaritmo del rapporto di espressione genica. Spesso, a supporto di questa tesi, sono stati forniti dei diagrammi quantile-quantile (*QQ-plot*) dei dati di un singolo *slide*; a giudicare dal grafico, in alcuni casi, si potrebbe dare considerazione a questa affermazione, se non fosse per un difetto di interpretazione. In questi esperimenti, l'obiettivo è studiare come si comporta ogni gene in vari *slides* (le replicazioni), non come si comportano diversi geni nello stesso. Il diagramma quantile-quantile menzionato sopra, di conseguenza, non può essere visto come una vera distribuzione di probabilità.

Wit & McClure (2004) osservano, d'altra parte, come l'esperienza suggerisca che molte distribuzioni di espressioni di geni, dopo la trasformazione logaritmica, siano indistinguibili dalla normale, e l'assunzione di normalità sia, di fatto, irrilevante ai fini dell'analisi.

Il modello studiato in questa tesi assume la normalità, nella speranza che, anche se non verificata, sia effettivamente poco influente sui risultati raggiunti.

Capitolo 2

I metodi bayesiani empirici

In questo capitolo si vuole riproporre il modello bayesiano empirico di Smyth (2004), che sta alla base di tutte le analisi che seguiranno. A questo scopo, si ripercorre lo sviluppo delle metodologie statistiche per l'identificazione dei geni differenzialmente espressi, a partire dai test t fino all'applicazione di impostazioni bayesiane empiriche (par. 2.1). Successivamente si presenta il modello nella sua formulazione completa (par. 2.2) e in una esemplificazione utile allo scopo di questa tesi (par. 2.3). Nel paragrafo 2.4 si descrive la rappresentazione grafica del modello in questione, che è il punto di partenza delle simulazioni effettuate. A conclusione del capitolo, nel paragrafo 2.5 vengono esposti alcuni commenti al modello.

2.1 Gli studi precedenti

I primi procedimenti statistici per l'identificazione di geni differenzialmente espressi attraverso *microarray* includevano test t gene-specifici e/o metodi di permutazione (Dudoit *et al.*, 2002), tanto quanto metodi di massima verosimiglianza (Ideker *et al.*, 2000). Due erano, e sono tuttora, i principali problemi comuni ai vari criteri: il controllo dell'errore di primo tipo (α) e la stima della varianza gene-specifica (σ_g^2). Per quanto riguarda il primo punto, Dudoit *et al.* (2003) propongono diverse procedure di aggiustamento dei valori-p stimati, ad

esempio attraverso il controllo del *family-wise error rate* (Westfall & Young, 1993), o, in alternativa, del *false discovery rate*, FDR (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001). Riguardo al secondo punto, l'uso del test t si è rivelato inadatto sin dall'inizio a questo tipo di studi. A causa dell'elevato numero di geni coinvolti in un esperimento, accade di frequente che alcuni presentino somme dei quadrati molto piccole, e di conseguenza valori del test grandi, pur essendo espressi in quantità scarse. È facile, perciò, che tali geni, nelle analisi, diano luogo ad errori di identificazione. Per ovviare a questi inconvenienti, sono state proposte statistiche alternative, che considerino stimatori delle varianze in grado di limitare i valori esageratamente piccoli ed esageratamente grandi. Sono stati così introdotti gli stimatori *shrinkage*, che vengono utilizzati in statistiche cosiddette t-moderate, e hanno lo scopo di ridurre la distorsione delle stime attraverso diverse strategie.

Alcune delle soluzioni proposte sono influenzate dalla teoria bayesiana e bayesiana empirica. In questo contesto, è possibile sfruttare tutta l'informazione portata dai geni, considerando le varianze come realizzazioni di un'unica variabile casuale, per migliorare le stime a livello globale. Ne è un esempio la statistica *SAM* (Tusher *et al.*, 2001), che modifica leggermente il test t aggiungendo a denominatore una costante appropriata. Può essere pensata come una statistica-test bayesiana, nella quale il denominatore è dato dalla somma di una deviazione standard "a priori" e della deviazione standard gene-specifica del log-rapporto.

In questa tesi si tratta un modello bayesiano empirico nel quale si effettuano tanti test t-moderati quanti geni sono presenti nell'esperimento; si trasformano i risultati in modo da ottenere una statistica B per ordinare i geni sulla base dell'espressione differenziale ed infine si stima la percentuale p di geni da considerarsi effettivamente differenzialmente espressi.

I metodi bayesiani si adattano bene a studiare problemi di inferenza multidimensionale, di conseguenza si applicano naturalmente a dati di *microarray*. Contrariamente a metodi che applicano inferenza classica separatamente per ogni gene, nell'analisi bayesiana è presente una sorta di condivisione delle informazioni tra geni. Questa conoscenza globale viene riassunta in distribuzioni a

priori per alcuni parametri, che si combinano con medie e deviazioni standard a livello di geni. Può sembrare paradossale che l'analisi dell'espressione genica differenziale di un gene venga in qualche modo influenzata dell'espressione di altri geni; a grandi linee, l'idea è che tutti i dati forniscano informazioni circa la variabilità tipica del sistema.

Fra i metodi bayesiani parametrici proposti in letteratura, Baldi & Long (2001) modellano ogni canale ($\log_2 R$ e $\log_2 G$) separatamente, utilizzando distribuzioni normali e a priori coniugate. Il risultato è dato da due modelli a posteriori per ogni gene, e dalla probabilità a posteriori che i due parametri di intensità siano uguali.

In Lönnstedt & Speed (2002) viene presentato, in un contesto bayesiano empirico, il logaritmo della quota a posteriori B, che verrà esposto più avanti nella riparametrizzazione di Smyth (2004). Le distribuzioni campionarie di migliaia di geni vengono riassunte in distribuzioni a priori per migliorare l'inferenza su ognuno di essi.

Broët *et al.* (2002) propongono un modello gerarchico completamente bayesiano dove i geni vengono assegnati ad uno di N (un numero sconosciuto) livelli di espressione. Il modello mistura che ne deriva, di conseguenza, ha un numero ignoto di componenti. Ogni gene assegnato allo stesso livello viene modellato con la stessa media e varianza.

Newton *et al.* (2001) presentano un metodo basato su modelli gerarchici di livelli di espressione, nei quali si tiene conto di due fonti di variabilità. La prima è l'errore di misurazione, la seconda è dovuta al fatto che sullo stesso *microarray* vengano analizzati diversi geni, che hanno valori di espressione diversi. Combinando queste due fonti, ottengono informazioni sulla probabilità di espressione differenziale. I dati analizzati provengono da un unico *array*.

Kendziorski *et al.* (2003) estendono la modellazione bayesiana empirica parametrica del precedente articolo al caso di più replicazioni di espressioni geniche in diverse condizioni.

Esistono altre formulazioni: Long *et al.* (2001), Gottardo *et al.* (2003), Ishwaran & Rao (2003) sono solo alcuni esempi della letteratura crescente che

si sviluppa attorno a questa metodologia.

Fra i metodi bayesiani empirici non parametrici (NEBM), Efron *et al.* (2001) presentano un modello mistura che con un minimo di assunzioni a priori produce probabilità di espressione differenziale a posteriori per ogni gene. In realtà, il metodo in questione aiuta piuttosto a selezionare tra diversi schemi di riduzione dei dati, un punto cruciale nel trattare la mole di informazioni che i *microarray* producono.

Inferenza bayesiana empirica

La precisione delle stime raggiunta attraverso l'applicazione di metodi bayesiani empirici ha uno svantaggio nella specificazione delle distribuzioni a priori sui parametri. Spesso le distribuzioni a priori sono difficili da verificare, ed è naturale aspettarsi che la loro specificazione abbia effetti sull'inferenza. A questo proposito, nei casi particolari come quello esaminato in cui risulta attuabile, si considera la possibilità di stimare gli iperparametri direttamente dai dati.
Si consideri il modello

$$y_1, ..., y_n|\theta_1, ..., \theta_n \overset{ind}{\sim} f(y_1|\theta_1), ..., f(y_n|\theta_n), \quad \theta_1, ..., \theta_n \overset{iid}{\sim} \pi(\theta|\gamma).$$

Un metodo bayesiano aggiungerebbe, a questo punto, una densità a priori $\pi(\gamma)$ per γ, e baserebbe l'inferenza per i θ_j sulla densità marginale a posteriori $\pi(\theta_j|y)$. Se, invece, si evita di aggiungere questo ulteriore livello di complessità, i dati avranno densità marginale

$$f(y_1, ..., y_n|\gamma) = \prod_{j=1}^{n} \int f(y_j|\theta_j)\ \pi(\theta_j|\gamma)\ d\theta_j,$$

dalla quale si può stimare γ. Un approccio naturale è usare lo stimatore a massima verosimiglianza $\hat{\gamma}$ da questa densità, e successivamente basare l'inferenza sulle densità a posteriori $\pi(\theta_j|y, \hat{\gamma})$. Generalmente, l'integrale viene risolto mediante metodi numerici.

2.2 L'approccio di Smyth

Il modello di Smyth (2004) presentato in questo paragrafo è una prosecuzione di quello proposto in Lönnstedt & Speed (2002). In esso appare per la prima volta la statistica B come alternativa per l'identificazione dei geni differenzialmente espressi. Smyth riprende l'impostazione e la completa suggerendo alcuni metodi di stima per gli iperparametri. È stato scelto questo modello per l'analisi perchè è molto usato nella pratica: la sua diffusione è sicuramente influenzata dal fatto che esista un software dedicato (LIMMA), e che questo sia la naturale estensione della libreria SMA, una delle prime disponibili per le analisi. Nondimeno, è in grado di gestire con facilità i disegni sperimentali più diffusi.

2.2.1 Introduzione al modello

Il metodo proposto si prefigge di identificare i geni differenzialmente espressi sulla base di probabilità a posteriori. Considerando il modello gerarchico parametrico sviluppato in Lönnstedt & Speed (2002), Smyth rielabora l'espressione del logaritmo della quota a posteriori (*log posterior odds* B) per generalizzarla ad un approccio da applicare ai tipi di disegni sperimentali più diffusi nella pratica.
Il contesto considerato è quello dei modelli lineari; vengono ricavati stimatori consistenti per tutti gli iperparametri del modello.

Smyth (2004) estende il modello di Lönnstedt & Speed (2002) nelle seguenti direzioni:

- reindirizza il contesto ai modelli lineari;
- ottiene stimatori consistenti ed in forma chiusa;
- affianca a B (*log posterior odds*) una statistica t moderata, nella quale vengono usate deviazioni standard a posteriori al posto di quelle ordinarie.

L'uso della statistica t si dimostra (in un certo senso) preferibile al rapporto B, in quanto il numero di parametri da stimare risulta inferiore.

Un approccio simile è stato seguito, come già menzionato, in Tusher *et al.* (2001), che propone l'uso di una statistica t con deviazione standard "compensata". In pratica il denominatore della t è costituito dalla deviazione standard più una costante, calcolata in modo da minimizzare il coefficiente di variazione.

2.2.2 L'uso dei modelli lineari in esperimenti di *microarray*

Smyth considera esperimenti che possano essere rappresentati in termini di un modello lineare per ogni gene,

$$y_g = X\alpha_g + \epsilon_g$$

dove y_g è la variabile risposta del gene g, X è una matrice nota, α_g è il vettore di coefficienti del modello gene-specifico ed $\epsilon_g \sim N(0, \sigma_g^2)$. In fig. 2.1 sono mostrati alcuni esempi di disegni sperimentali nella notazione di Kerr & Churchill (2001), dove A, B e C sono campioni di mRNA da confrontare e ogni freccia rappresenta un *microarray*.

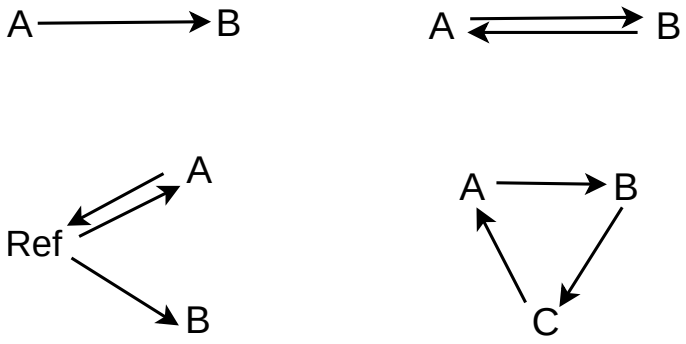

Figura 2.1: Esempi di disegno sperimentale.

Il campione alla base della freccia viene contrassegnato con una tinta verde, mentre l'altro con una tinta rossa. Nel disegno (a), ad esempio, c'è un solo

microarray che mette a confronto i campioni A e B. L'esperimento (b), invece, ripete due volte l'ibridazione di A e B invertendo le tinte (procedimento noto come *dye-swap*). Il modello che ne deriva ha variabili risposte y_{g1} e y_{g2}, che sono i log-rapporti dei due *slides*, e matrice X

$$X = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Il disegno (c) confronta i campioni A e B attraverso un comune campione di riferimento. Una matrice appropriata in questo caso è

$$X = \begin{pmatrix} -1 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix},$$

che produce un modello lineare in cui il primo coefficiente stima la differenza tra A e il riferimento, mentre il secondo stima la differenza di interesse, ossia B - A.

Generalmente, la variabile risposta è il logaritmo del rapporto tra i due campioni e si assume che sia normalizzata; supponendo di disporre di n *slides*, questa viene indicata da $y_g^T = (y_{g1}, y_{g2}, ..., y_{gn})$, per il g-esimo gene. Dati i parametri, tutti i geni e le replicazioni si assumono indipendenti.
Si assume che

$$E(y_g) = X\alpha_g = \mu_g,$$

dove μ_g rappresenta il valore medio dell'espressione del gene g, e che

$$Var(y_g) = W_g\sigma_g^2,$$

dove W_g è una matrice di pesi semi-definita positiva nota. È possibile che vi siano particolari contrasti d'interesse biologico, definiti da $\beta_g = C^T\alpha_g$. Sarà importante verificare se i valori di questi contrasti sono uguali o diversi da zero, $H_0 : \beta_{gj} = 0$.

Per ogni gene è dunque possibile ottenere le stime dei coefficienti $\hat{\alpha}_g$, le stime s_g^2 di σ_g^2 e la stima della matrice di varianze e covarianze

$$Var(\hat{\alpha}_g) = V_g s_g^2,$$

dove V_g è una matrice definita positiva e indipendente da s_g^2. Gli stimatori dei contrasti sono $\hat{\beta}_g = C^T V_g C s_g^2$. Non si assume necessariamente che le risposte siano normalmente distribuite, nè che la stima del modello venga realizzata a minimi quadrati. Tuttavia, si suppone che gli stimatori dei contrasti siano approssimativamente normali con media β_g e matrice di varianze e covarianze $C^T V_g C s_g^2$.
Essendo v_{gj} il g-esimo elemento diagonale della matrice V_g, le assunzioni sulle distribuzioni degli stimatori considerate nel seguito si possono riassumere con

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

e

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

dove d_g rappresenta i gradi di libertà residui del modello lineare per il gene g. Sotto questi presupposti, la statistica t ordinaria

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g\sqrt{v_{gj}}}$$

si distribuisce approssimativamente come una t di Student con d_g gradi di libertà.

2.2.3 Il modello gerarchico

Come già sottolineato, un punto fondamentale per l'analisi è la necessità di trarre vantaggio dalla struttura parallela dei dati, per la quale lo stesso modello viene stimato per ogni gene. La chiave è capire come i coefficienti ignoti β_{gj} e le varianze ignote σ_g^2 varino tra geni, ponendo distribuzioni a priori sui parametri.

Si assume che la distribuzione a priori per σ_g^2 sia un χ^2 inverso, ovvero

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2}\chi_{d_0}$$

dove s_0^2 è l'iperparametro della distribuzione per σ_g^2 e d_0 indica i gradi di libertà. Riguardo β_{gj}, si suppone che sia diverso da zero con probabilità p, ossia

$$P(\beta_{gj} \neq 0) = p_j.$$

In questo modo p_j indica la proporzione attesa di geni effettivamente differenzialmente espressi. Per questi geni, l'informazione a priori è l'equivalente di un'osservazione a priori uguale a zero, con varianza $v_{0j}\sigma_g^2$,

$$\beta_{gj}|\beta_{gj} \neq 0, \sigma_g^2 \sim N(0, v_{0j}\sigma_g^2).$$

Questa espressione descrive la distribuzione attesa dei cambiamenti nei geni che sono differenzialmente espressi. A prescindere dalla proporzione p_j, l'equazione descritta costituisce una a priori coniugata per il modello normale presentato nel paragrafo precedente. Con questo modello gerarchico, la media a posteriori di $\sigma_g^2|s_g^2$ è

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}.$$

Il valore a posteriori avvicina le varianze osservate ai valori a priori (stimatore *shrinkage*).
A questo punto, si definisce la statistica t moderata come

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v}_g}.$$

Questa statistica rappresenta un approccio ibrido classico/bayesiano, nel quale le usuali varianze campionarie della statistica t ordinaria vengono sostituite dalle varianze a posteriori.

Smyth (2004) dimostra che $\tilde{t}_g$ e s_g^2 sono indipendenti. In particolare, la statistica t moderata, sotto $H_0 : \beta_{gj} = 0$, segue una t di Student con $d_g + d_0$ gradi di libertà. I gradi di libertà aggiuntivi rispetto alla statistica ordinaria, riflettono l'informazione in più che si ottiene, sulla base del modello gerarchico, dall'insieme dei geni.

2.2.4 Il logaritmo della quota a posteriori

Date le distribuzioni marginali di $\tilde{t}_{gj}$ ed s_g^2, si ottiene facilmente la quota a posteriori O: questo rapporto è un utile strumento di ordinamento dei geni sulla base dell'evidenza di espressione genica differenziale.
Si definisce come

$$O_{gj} = \frac{p(\beta_{gj} \neq 0|\tilde{t}_{gj}, s_g^2)}{p(\beta_{gj} = 0|\tilde{t}_{gj}, s_g^2)} = \frac{p(\beta_{gj} \neq 0, \tilde{t}_{gj}, s_g^2)}{p(\beta_{gj} = 0, \tilde{t}_{gj}, s_g^2)} = \frac{p_j}{1-p_j} \frac{p(\tilde{t}_{gj}|\beta_{gj} \neq 0)}{p(\tilde{t}_{gj}|\beta_{gj} = 0)},$$

dato che $\tilde{t}_{gj}$ ed s_g^2 sono indipendenti e la distribuzione di s_g^2 non dipende dai β_{gj}. Sostituendo la densità di $\tilde{t}_{gj}$, si ottiene l'espressione della quota

$$O_{gj} = \frac{p_j}{1-p_j} \left(\frac{v_{gj}}{v_{gj}+v_{0j}} \right)^{1/2} \left(\frac{\tilde{t}_{gj}^2 + d_0 + d_g}{\tilde{t}_{gj}^2 \frac{v_{gj}}{v_{gj}+v_{0j}} + d_0 + d_g} \right)^{(1+d_0+d_g)/2}$$

L'espressione è in accordo con l'equazione (7) di Lönnstedt & Speed (2002). Seguendo il loro procedimento, la statistica

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

è in una scala più comoda, perciò di uso preferibile rispetto ad O.

Una volta ottenuta la statistica B, si pone il problema dell'interpretazione: un traguardo importante per l'analisi è riuscire a fissare una soglia al di sopra della quale dichiarare i geni differenzialmente espressi. Un valore di $B = 0$, ad esempio, rappresenta il caso di maggiore incertezza, in quanto deriva da $O = 1$, ossia dall'uguaglianza delle probabilità p e $1 - p$. Si potrebbe quindi pensare di fissare la soglia di identificazione in questo punto; purtroppo, però, non si può fare affidamento sul valore di B, poiché dipende dal parametro p, sul quale, sinora, si sono sempre effettuate congetture a priori. Di conseguenza, la statistica B viene usata in particolar modo come metodo di ordinamento dei geni sulla base dell'espressione differenziale, ma ancora non è chiaro il significato che assuma. In alternativa, può essere usato il test t-moderato con correzione

del FDR per i test multipli, che viene fornito assieme a B nelle analisi con LIMMA.

Si è accennato al ruolo di p nella determinazione della statistica B; il parametro rappresenta la proporzione di geni che si suppone, a priori, siano differenzialmente espressi, e solitamente viene fissato ad un valore arbitrario. Di fatto, sono stati proposti alcuni metodi di calcolo parametrici e non, ma nella pratica, spesso, ci si dimentica di questo parametro. In LIMMA, ad esempio, il valore è impostato a $p = 0.01$ nella funzione `eBayes()`, di conseguenza, se non specificato altrimenti, si assume che la proporzione di geni differenzialmente espressi sia 1%.

2.2.5 La stima degli iperparametri

Nel lavoro di Smyth (2004) gli iperparametri del modello gerarchico vengono stimati dai dati; si calcolano stime consistenti ed in forma chiusa per d_0, s_0 e v_{0j} partendo dalle varianze campionarie s_g^2 e dalle statistiche t moderate $\tilde{t}_{gj}$. Più in dettaglio, d_0 ed s_0 vengono stimati mediante il metodo dei momenti, eguagliando i valori empirici dei primi due momenti della quantità $\log s_g^2$ a quelli attesi. Si è scelto di utilizzare $\log s_g^2$ al posto di s_g^2 perché i momenti di $\log s_g^2$ sono finiti per qualsiasi grado di libertà e perchè la distribuzione di $\log s_g^2$ è più prossima alla normale. Le stime di v_{0j}, invece, si ottengono eguagliando le statistiche ordinate $|\tilde{t}_{gj}|$ ai loro valori nominali. Gli stimatori ricavati potrebbero essere utilizzati come valori iniziali per eventuali stime di massima verosimiglianza; tuttavia, si è osservato che, solitamente, gli stimatori calcolati in questo modo sono sufficientemente precisi.

Stima di d_0 ed s_0

Poniamo

$$z = \log s_g^2;$$

ogni s_g^2 segue una distribuzione F scalata, di conseguenza ogni z_g si distribuisce come una z di Fisher più una costante (Johnson & Kotz, 1970). Dunque la

distribuzione degli z_g è approssimativamente normale e ha momenti finiti di ogni ordine, incluso

$$E(z_g) = \log s_0^2 + \psi(d_g/2) - \psi(d_0/2) + \log(d_0/d_g)$$

e

$$Var(z_g) = \psi'(d_g/2) + \psi'(d_0/2),$$

dove $\psi()$ e $\psi'()$ sono, rispettivamente, le funzioni *digamma* e *trigamma*. Fissata la quantità

$$e_g = z_g - \psi(d_g/2) + \log(d_g/2),$$

si ottengono il suo valore atteso

$$E(e_g) = \log s_0^2 - \psi(d_0/2) + \log(d_0/2)$$

e la sua varianza

$$E\left\{(e_g - \bar{e})^2 n/(n-1) - \psi'(d_g/2)\right\} \approx \psi'(d_0/2);$$

a questo punto si deriva la stima di d_0 risolvendo

$$\psi'(d_0/2) = \text{mean}\left\{(e_g - \bar{e})^2 n/(n-1) - \psi'(d_g/2)\right\}.$$

Questa funzione può essere risolta numericamente (si veda Smyth, 2004, Appendice). Data la stima di d_0, si ricava quella di s_0^2

$$s_0^2 = \exp\left\{\bar{e} + \psi(d_0/2) - \log(d_0/2)\right\}.$$

Stima di v_{0j}

In questa sezione si stima v_{0j} per un dato j. L'indice j verrà quindi omesso nelle quantità $\tilde{t}_{gj}$, v_{gj}, v_{0j} e p_j.

Ogni gene g produce una statistica t moderata $\tilde{t}_g$. La funzione di ripartizione di $\tilde{t}_g$ è

$$F(\tilde{t}_g; v_g, v_0, d_0 + d_g) = p \cdot F\left(\tilde{t}_g \left\{\frac{v_g}{v_g + v_0}\right\}^{1/2}; d_0 + d_g\right) + (1-p)F(\tilde{t}_g; d_0 + d_g)$$

dove $F(\cdot; k)$ è la funzione di ripartizione di una t con k gradi di libertà.

Sia r la posizione del gene g nella lista ordinata in modo decrescente dei $|\tilde{t}_g|$. Lo scopo è associare i valori-p di ogni $|\tilde{t}_g|$ al rispettivo valore nominale, dato il rango. Per ogni g ed r bisogna quindi risolvere

$$2F(-|\tilde{t}_g|; v_g, v_0, d_0 + d_g) = \frac{r - 0.5}{G}.$$

Il membro di sinistra è il valore-p dati i parametri, mentre il membro di destra è il valore nominale corrispondente al valore-p di rango r. L'interpretazione grafica che si può dare chiarisce l'idea di fondo: se si costruisse un *probability*-plot con i $|\tilde{t}_g|$ verso i quantili teorici corrispondenti alle probabilità $(r - 0.5)/G$, la condizione appena definita rappresenterebbe il requisito per cui un valore di $|\tilde{t}_g|$ sia collocato esattamente sulla linea di uguaglianza. Il valore di v_0 che soddisfa l'equazione è

$$v_0 = v_g \left(\frac{\tilde{t}_g^2}{q_{target}^2} - 1\right)$$

con

$$q_{target} = F^{-1}(p_{target}; d_0 + d_g)$$

e

$$p_{target} = \frac{1}{p}\left\{\frac{(r - 0.5)}{2G} - (1-p)F(-|\tilde{t}_g|; d_0 + d_g)\right\},$$

posto che $0 < p_{target} < 1$ e $q_{target} \leq |\tilde{t}_g|$.

Per ottenere uno stimatore di v_0 combinato, Smyth (2004) propone di utilizzare la media delle stime individuali per $r = 1, ..., Gp/2$. Esso sarà positivo, a meno che nessuno dei primi $Gp/2$ valori di $|\tilde{t}_g|$ ecceda il corrispondente valore nominale della t, nel qual caso lo stimatore sarà zero.

Purtroppo le informazioni a disposizione per la stima dei v_0 sono poche, dato che questo parametro compare solo nella distribuzione dei geni differenzialmente espressi; una strategia pratica per superare questi problemi, che è stata implementata in LIMMA, è porre dei limiti alla stima di $v_0^{1/2}\sigma_g$. Questa quantità rappresenta la deviazione standard dei *log-fold-changes* (ossia i cambiamenti percentuali dei valori di y) per i geni differenzialmente espressi, e quindi non può assumere valori esageratamente elevati o piccoli. In LIMMA i limiti sono fissati a 0.1 e 4, ma possono essere scelti.

Stima di p

Nel suo articolo, Smyth accenna ad un metodo di stima iterativo di p, tuttavia non dà molta importanza a questa proposta e non ne analizza le caratteristiche. Alla stima di questo iperparametro è dedicato interamente il capitolo 3.

2.3 Un'esemplificazione del modello di Smyth

Il modello che verrà analizzato specificatamente in questo elaborato è un caso particolare di quello di Smyth. Essendo l'obiettivo di questo elaborato studiare gli effetti della struttura gerarchica sulla stima di p, risulta conveniente considerare il modello nella sua versione più semplice, dato che in questo si eliminano effetti indotti da particolari piani sperimentali.

Il punto di partenza è sempre rappresentato dal modello lineare per ogni gene,

$$y_g = \mu_g + \epsilon_g.$$

In particolare, nel seguito, si farà riferimento ad un esperimento del tipo A $\rightarrow$ B, con n *slides* e matrice

$$X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

La variabile risposta è la stessa, $y_g^T = (y_{g1}, y_{g2}, ..., y_{gn})$, per il g-esimo gene. In questo caso, nell'espressione

$$E(y_g) = X\alpha_g,$$

α_g è uno scalare, il coefficiente del modello gene-specifico, ed anche in

$$Var(y_g) = W_g\sigma_g^2$$

W_g è uno scalare noto. Il parametro di interesse è μ_g e quindi non è necessario introdurre contrasti.
Essendo v_g pure uno scalare, le assunzioni sulle distribuzioni degli stimatori si riassumono con

$$\hat{\alpha}_g|\alpha_g, \sigma_g^2 \sim N(\alpha_g, \sigma_g^2)$$

e

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

dove d_g rappresenta i gradi di libertà residui del modello lineare per il gene g. La statistica t ordinaria si distribuisce approssimativamente come una t di Student con d_g gradi di libertà. Nel modello gerarchico i parametri rimangono invariati, l'unica differenza riguarda l'indice j, che, non essendoci contrasti da analizzare, sparisce. In questo modo p è unico, e indica la proporzione attesa di geni effettivamente differenzialmente espressi. Per questi geni, l'informazione a priori è l'equivalente di un'osservazione a priori uguale a zero, con varianza $v_0\sigma_g^2$,

$$\alpha_g|\alpha_g \neq 0, \sigma_g^2 \sim N(0, v_0\sigma_g^2).$$

La statistica t moderata diventa

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g\sqrt{v_g}}.$$

Anche la quota a posteriori è unica per ogni gene,

$$O_g = \frac{p(\beta_g \neq 0|\tilde{t}_g, s_g^2)}{p(\beta_g = 0|\tilde{t}_g, s_g^2)} = \frac{p(\beta_g \neq 0, \tilde{t}_g, s_g^2)}{p(\beta_g = 0, \tilde{t}_g, s_g^2)} = \frac{p}{1-p}\frac{p(\tilde{t}_g|\beta \neq 0)}{p(\tilde{t}_g|\beta = 0)}$$

La stima degli iperparametri avviene nel modo descritto al paragrafo precedente, ad eccezione di v_0 che è unico, quindi non richiede stime combinate.

2.4 Il modello in un grafo

La costruzione di un modello grafico, spesso, è il primo passo di un'analisi. In particolare, è utile rappresentare un modello gerarchico mediante un grafo, innanzitutto perché chiarisce le relazioni di dipendenza tra le variabili, secondariamente perché ad esso si associa una fattorizzazione ricorsiva delle densità efficace nel momento in cui si vogliano simulare realizzazioni del modello. A questo proposito, va segnalato il software `BUGS`, che è disponibile all'URL (`http://www.mrc-bsu.cam.ac.uk/bugs/`), un programma per l'analisi bayesiana di modelli statistici complessi mediante l'uso di catene markoviane (metodi *MCMC*), che si basa proprio su grafi di questo tipo.

Con riferimento al modello descritto, la fig. 2.2 mostra il grafo in cui ogni variabile viene rappresentata tramite *nodi*; ogni nodo può avere "genitori" e "figli" collegati da frecce secondo un rapporto di dipendenza. I piatti (*plates*) sono un espediente grafico finalizzato a semplificare e rendere più chiara la presentazione. Se, ad esempio, vi sono G variabili σ_g^2 ma un solo d_0, la struttura di fig. 2.3(a) viene rappresentata, per praticità, come in fig. 2.3(b). Le variabili Y_{gi} sono, di conseguenza, $G \times n$.

Per comprendere le relazioni di dipendenza nel modello, si fa riferimento alla convenzione che ogni nodo, dati i suoi genitori, è indipendente da tutti gli altri nodi del grafo, eccetto i suoi discendenti.

Il grafo in questione è strutturato come segue. All'esterno dei piatti ci sono i quattro iperparametri del modello: v_0, p, d_0 e s_0^2. Essi non hanno genitori. Il nodo p ha un figlio $[D_g = \ ?]$, che rappresenta una variabile binomiale $Bi(1, p)$; in questa fase, si simula un gene differenzialmente espresso con probabilità p, oppure non differenzialmente espresso con probabilità $1 - p$. A sua volta, il

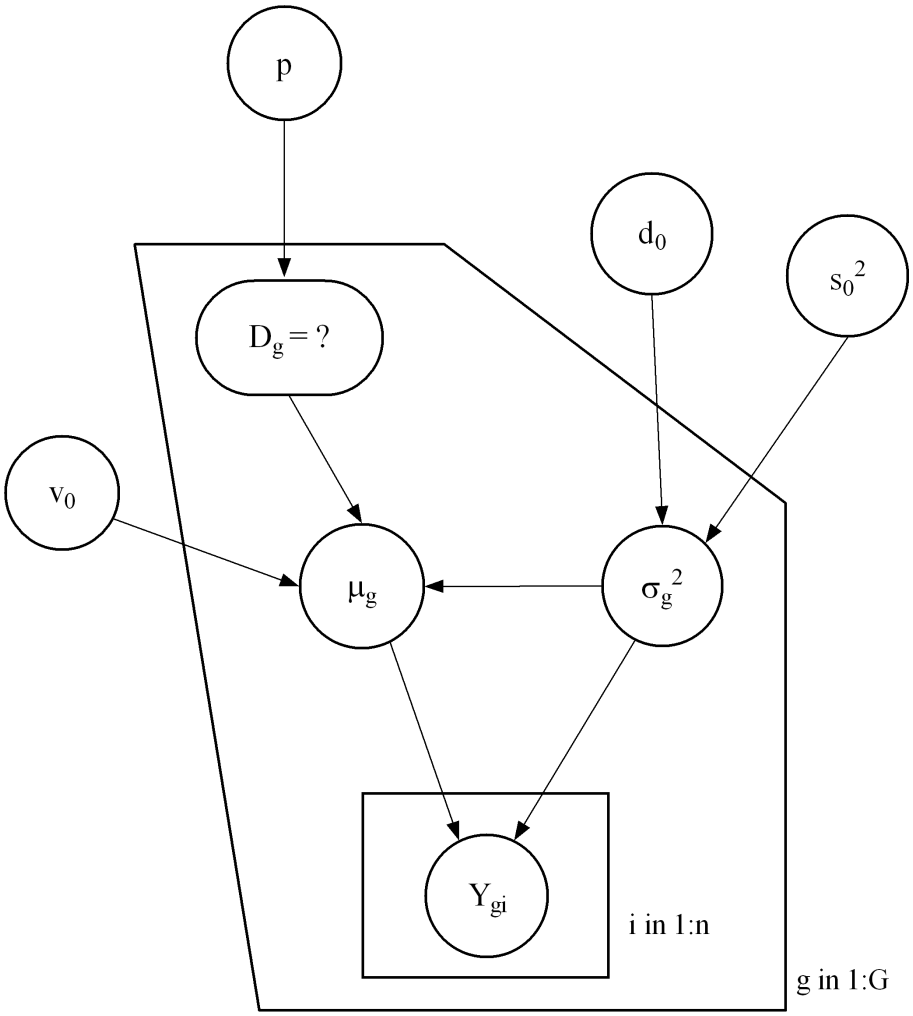

Figura 2.2: Rappresentazione grafica del modello di Smyth.

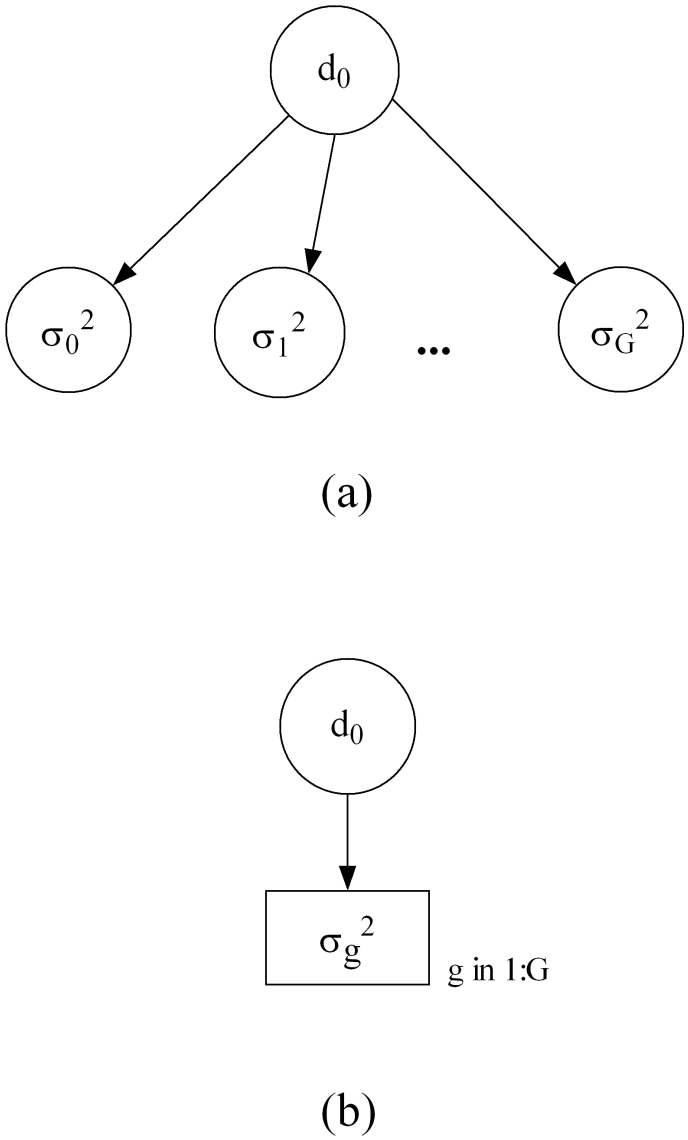

Figura 2.3: Esempio di utilizzo dei piatti.

nodo $[D_g = \ ?]$ ha un figlio μ_g, in comune con i nodi v_0 e σ_g^2. Questo indica che la variabile μ_g dipende da tre genitori: infatti, la media di ogni gene viene determinata in modo differente se questo è differenzialmente espresso o no, con varianza dipendente da v_0 e σ_g^2.
Dall'altro lato, ci sono i nodi d_0 ed s_0^2 che hanno un figlio comune, σ_g^2, ossia la varianza campionaria gene-specifica. Esso, a sua volta, è genitore di due nodi: μ_g e Y_{gi}, in quanto necessario per la definizione della varianza sia delle medie di espressione, sia della variabile risposta. Infine, il nodo Y_{gi}, che non ha figli, dipende da due genitori: μ_g e σ_g^2, rispettivamente la media e la varianza della distribuzione da cui vengono generati i valori.

2.5 Alcune considerazioni sul modello di Smyth

È necessario, a questo punto, sottolineare alcuni aspetti del modello utilizzato. Innanzitutto una considerazione sull'impostazione bayesiana empirica; come già accennato al paragrafo 2.1, l'idea di base è che l'intero sistema di geni fornisca informazioni utili per le stime dei parametri, in particolare riguardo alla varianza della variabile risposta. Risulta più intuitivo pensare che siano le replicazioni dello stesso gene ad essere informative, ed è indubbio, naturalmente; anche la numerosità dei geni, però, porta conoscenze utili per le stime. Un semplice ragionamento può illustrare il principio. Nel modello, si considera il valore di espressione di ogni gene come una realizzazione da una variabile casuale con una certa media e una certa varianza. Avendo informazioni per migliaia di geni contemporaneamente, si può pensare che ci sia una variabile che genera tutte le varianze gene-specifiche, e di conseguenza utilizzare le conoscenze nel loro complesso per migliorare le stime a livello di singoli geni.

Sempre con riferimento alla variabilità, va notato che un parametro dal significato apparentemente poco chiaro è v_0. Esso compare nella varianza dello stimatore della media (per i geni differenzialmente espressi), ed è una sorta di

fattore di scala che mette in relazione tale varianza con quella della variabile risposta y_g. La connessione tra la varianza campionaria attesa e la varianza nella distribuzione a priori di μ_g è difficile da giustificare dal punto di vista teorico, ma è dettata da esigenze di tipo pratico; in questo modo si riescono ad ottenere stime dei parametri in forma chiusa. Lönnstedt (2005) propone un metodo di stima di v_0 (c nell'articolo) alternativo a quello di Smyth, basato sul metodo dei momenti. La difficoltà che si incontra nella stima di questo parametro è dovuta al fatto che compare solamente nei geni differenzialmente espressi, che sono ignoti e solitamente sono una piccola parte del totale.

Un'ultima considerazione a proposito dell'indipendenza tra geni. Nonostante i metodi bayesiani empirici siano molto usati in pratica, alcuni autori (Qiu *et al.*, 2005) sollevano delle critiche relative alla validità degli assunti distribuzionali, che, se non verificati, possono portare a risultati di scarsa qualità. Alcuni riferimenti a metodi che considerino una struttura di correlazione si trovano in Dudoit *et al.* (2004a,b), van der Laan *et al.* (2004a,b).

Capitolo 3

Il parametro p

In questo capitolo vengono presentate le simulazioni effettuate per studiare il parametro p. Nel par. 3.1 si trovano alcune considerazioni preliminari, nel par. 3.2 si descrivono i metodi di stima considerati per le analisi e nel par. 3.3 vengono esposti i risultati ottenuti. Nel par. 3.4 si accenna al *false discovery rate* ed infine nel par. 3.5 si trovano alcune considerazioni conclusive sui metodi utilizzati.

3.1 Considerazioni preliminari

Il parametro p, come già menzionato, rappresenta la percentuale di geni differenzialmente espressi in un esperimento di *microarray*. Non si dispone di molta informazione a priori su di esso; probabilmente è per questo motivo che nella maggior parte delle analisi viene fissato ad un valore arbitrario, come ad esempio $p = 0.01$ oppure $p = 0.005$. È molto chiaro, d'altra parte, il suo ruolo nella determinazione del valore di B: sebbene al variare di p l'ordinamento dei geni secondo B non cambi, o cambi solo marginalmente, si osservano evidenti alterazioni della scala. A causa di ciò, non si può fare affidamento a nessun valore di soglia "predefinito", come ad esempio $B = 0$, per la determinazione dei geni differenzialmente espressi. Una procedura di stima specifica per p darebbe significato statistico a B, che di conseguenza, invece di essere solamente uno

strumento di ordinamento, potrebbe essere utilizzato, per ogni gene, anche per il suo valore numerico.

3.2 Metodi di stima proposti

Nel presente elaborato si vuole provare l'efficacia di due metodi di stima di p attraverso simulazioni dal modello descritto nel capitolo 2. Il primo è quello proposto proprio da Smyth (2004), che suggerisce di calcolare iterativamente la quantità

$$\hat{p}^{(i+1)} = \frac{1}{G}\sum_{g=1}^{G}\frac{O_g^{(i)}}{1+O_g^{(i)}},$$

dato che $\frac{O_g}{1+O_g}$ è la probabilità stimata che il gene g sia differenzialmente espresso. In dettaglio, partendo da un qualsiasi valore di $\hat{p}$ fissato, si stimano i valori $O_g^{(i)}$, si calcola il nuovo $\hat{p}^{(i+1)}$, si ristimano i valori $O_g^{(i+1)}$ e si prosegue in questo modo fino a convergenza. Tale proposta, tuttavia, è presentata in modo vago, e non è implementata in LIMMA; come già sottolineato, la funzione `eBayes()` calcola i valori di B con $p = 0.01$, se non specificato altrimenti.
Il secondo metodo è stato proposto da Lönnstedt & Britton (2005), che presentano una procedura basata sul metodo dei momenti. Brevemente, gli autori fissano le quantità

$$M_g = y_{g.}/s_g,$$

dove $y_{g.}$ rappresenta la media dei valori di espressione per ogni gene g, e considerano le equazioni di riferimento $M_g\sqrt{n} \sim t(n-1)$ se il gene non è differenzialmente espresso, oppure $M_g\sqrt{n/(1+nv_0)} \sim t(n-1)$ in caso contrario. La quantità M si può scrivere come prodotto di due variabili indipendenti $M = ZT$, dove $T \sim t(n-1)$ e $Z = \sqrt{1/n}$ con probabilità $1-p$ e $\sqrt{(1+nv_0)/n}$ con probabilità p. Per l'indipendenza, $E(M^r) = E(Z^r)E(T^r)$ per l'r-simo momento di M. I momenti dispari di T sono zero per simmetria della t, per questo si usano le equazioni generate dal primo - in valore assoluto - e dal secondo momento: $E|M| = E|Z|E|T|$ e $E(M^2) = E(Z^2)E(T^2)$. Sostituendo le espressioni di questi momenti nelle equazioni M_g si ottiene

$$\begin{cases} pv_0 = k_1 \\ p(\sqrt{1+nv_0} - 1) = k_2 \end{cases}$$

dove

$$\begin{cases} k_1 = \frac{n-3}{n-1} E(M^2) - \frac{1}{n} \\ k_2 = \frac{(n-2)\sqrt{\pi}\Gamma(\frac{n-1}{2})\sqrt{n}}{2\sqrt{n-1}\Gamma(\frac{n}{2})} E|M| - 1 \end{cases}$$

che ha soluzione $\hat{p} = \frac{k_2^2}{nk_1 - 2k_2}$.

3.3 Simulazioni

3.3.1 Validazione dei dati

Con riferimento al grafo di fig. 2.2, sono stati simulati diversi insiemi di dati, al fine di valutare il rendimento dei metodi proposti per la stima di p. Sono stati fissati:

- gli iperparametri $p = 0.03$, $s_0^2 = 1$, $v_0 = 4$ e $d_0 = 3$;
- il numero di geni $G = 5000$;
- il numero di replicazioni per ogni gene $n = 6$;
- il numero di simulazioni svolte con questi parametri $L = 300$.

Con questa impostazione si sono ottenuti dati abbastanza realistici, che, almeno apparentemente, presentano le caratteristiche tipiche di esperimenti veri. A questo proposito, il grafico di fig. 3.1 mostra i valori della variabile risposta per tutti i geni analizzati in un ipotetico *array*; i punti di colore diverso dal nero rappresentano i geni differenzialmente espressi. La fig. 3.2, invece, riporta un istogramma della variabile risposta, sempre con riferimento allo stesso *array*. È evidente che la variabile y per i geni differenzialmente espressi non assume necessariamente valori elevati, nè, viceversa, valori grandi indicano con certezza espressioni geniche differenziali; inoltre, la maggior parte di essi si distribuisce tra -5 e 5. Si può presumere che i dati rappresentino verosimilmente un esperimento.

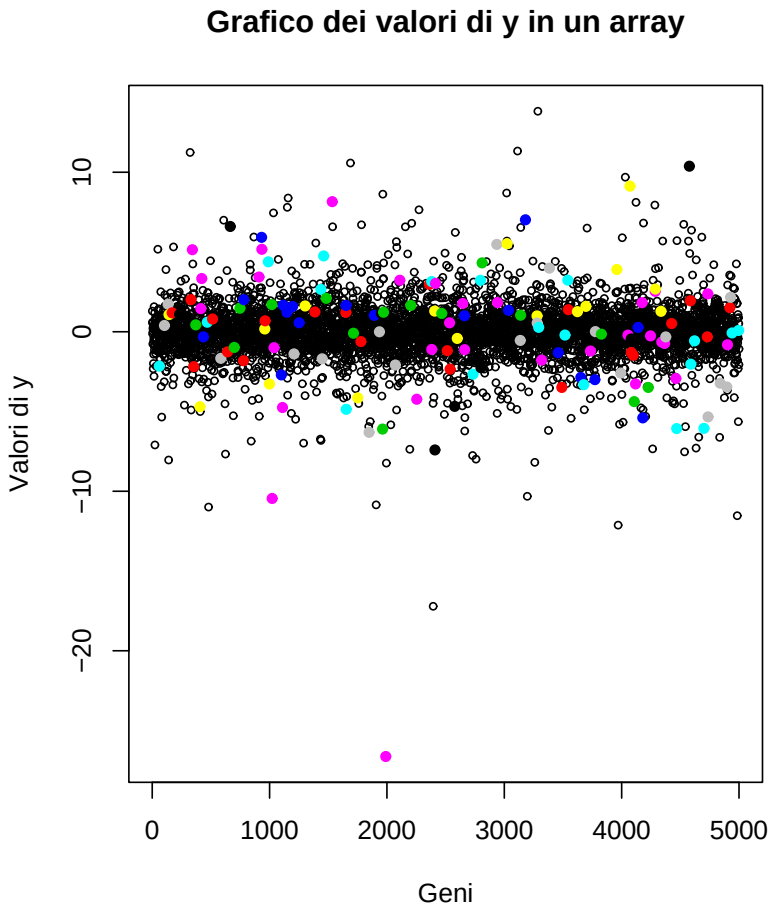

Figura 3.1: Valori di y per un *array*.

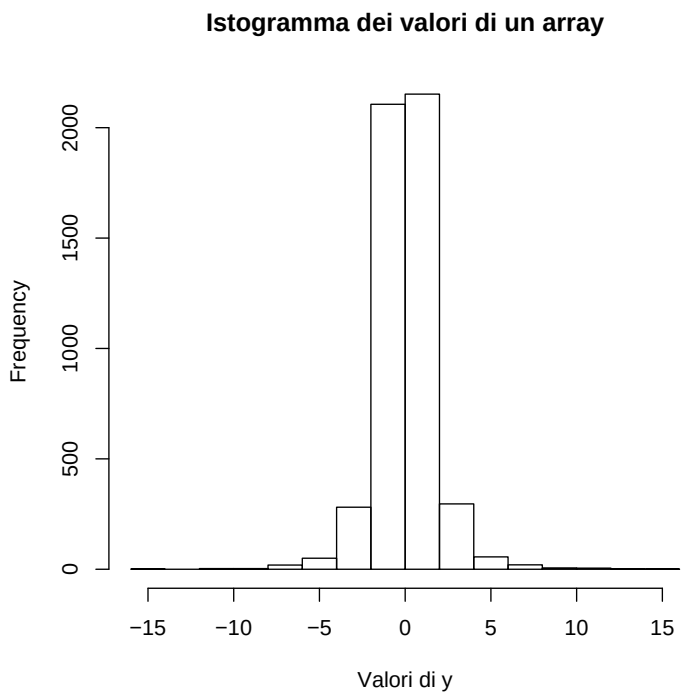

Figura 3.2: Istogramma dei valori di y per un $array$.

3.3.2 Convergenza empirica delle stime

In questa prima fase dell'analisi, è importante verificare la convergenza empirica delle stime iterative proposte da Smyth. Nelle simulazioni, sono state effettuate 15 iterazioni su ogni insieme di dati; la differenza tra le ultime due stime è mediamente di 1/10000. L'andamento che si osserva nei due esempi delle figg. 3.3 e 3.4 mostra una chiara tendenza alla convergenza.

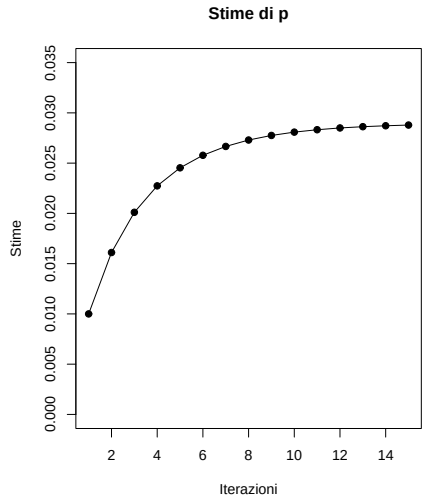

Figura 3.3: Grafico delle stime iterative di p.

3.3.3 Distribuzione campionaria delle stime

A questo punto è opportuno verificare l'accuratezza dei metodi di stima proposti. Sempre con riferimento alle $L = 300$ simulazioni svolte in precedenza, le figg. 3.5 e 3.6 mostrano gli istogrammi dei valori di $\hat{p}$ ottenuti con i due metodi. Nel primo la distribuzione delle stime è abbastanza simmetrica attorno al vero p, mentre nel secondo la distribuzione, oltre ad essere asimmetrica, è completamente spostata rispetto al valore 0.03. Si può affermare che il metodo di Smyth si avvicina molto al valore vero del parametro p (la media delle stime è

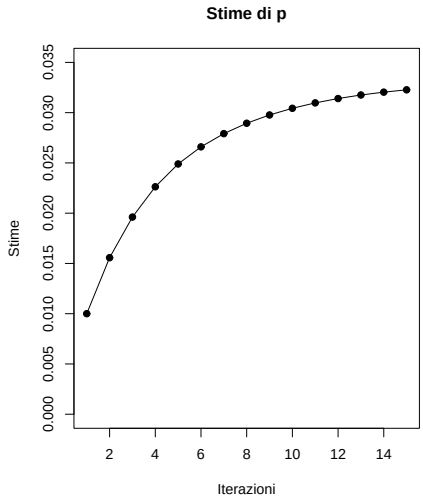

Figura 3.4: Grafico delle stime iterative di p.

infatti 0.0299), mentre quello di Lönnstedt è ben lontano (la media delle stime è 0.5719). Riguardo quest'ultimo, in particolare, nemmeno gli autori dell'articolo garantiscono un buon rendimento.

3.3.4 Distorsione delle stime

Al fine di valutare il comportamento degli stimatori al variare della numerosità campionaria e del numero di geni analizzati, è stato monitorato l'andamento delle stime per vari n e per vari G. Si è cercato di analizzare come influisca l'informazione portata, rispettivamente, da più replicazioni dello stesso gene, e dall'aggiunta di geni nell'esperimento. Se la prima questione è chiara e ragionevole, la seconda lo diventa in un contesto bayesiano empirico, com'è quello di questa tesi. Ci si aspetta, in virtù di quanto detto al par. 2.5, che le stime ottenute sfruttando l'informazione di molti geni siano più accurate di quelle ottenute con pochi; ciò significa che, pur non aumentando la numerosità campionaria n, si auspicano ugualmente miglioramenti nell'analisi. Le simulazioni sono state svolte per vari n e G mantenendo la proporzione di geni differen-

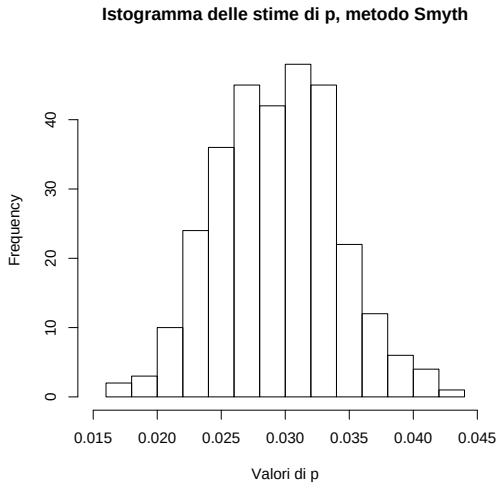

Figura 3.5: Istogramma dei valori di p stimati col metodo di Smyth.

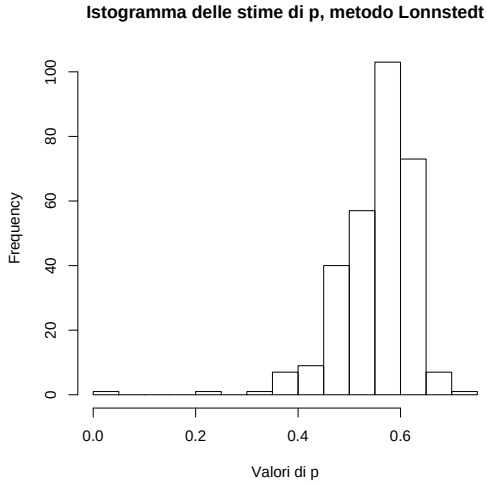

Figura 3.6: Istogramma dei valori di p stimati col metodo di Lönnstedt.

zialmente espressi costante, $p = 0.03$; questo non ha un effetto prevedibile sulla consapevolezza dell'esperimento, perchè l'aggiunta di un gene va a modificare il contesto che è stato analizzato nella simulazione precedente. In sostanza, si aggiunge informazione riguardo ad una situazione che muta, non è ben definita.

Numero di geni costante

In primo luogo, è stato fissato il numero di geni $G = 1000$, al fine di studiare il comportamento delle stime per n che assume i valori 5, 10, 15, 20, 50, 100, 150, 200, 300. Se gli stimatori operano bene, ci si aspetta che all'aumentare della numerosità campionaria le stime di p diventino più precise, ossia non si discostino eccessivamente dal vero p. Nelle figg. 3.7 e 3.8 si osservano i comportamenti dei due stimatori; le stime calcolate con il metodo di Smyth effettivamente diventano molto più accurate all'aumentare della numerosità campionaria; lo stimatore di Lönnstedt & Britton, invece, non si avvicina per nulla al valore reale. Si può dedurre, perciò, che sinora tra i due metodi è preferibile quello di Smyth.
La valutazione si completa prestando attenzione alla deviazione standard delle stime prodotte nelle varie simulazioni; naturalmente, ci si aspetta che l'informazione portata da n grande si rifletta in una diminuzione dell'incertezza della stima. Si può notare nelle figg. 3.9 e 3.10 che entrambi gli stimatori confermano le supposizioni, ma la deviazione standard dello stimatore di Smyth è nettamente più piccola rispetto a quella di Lönnstedt & Britton, o meglio, quest'ultimo non raggiunge il primo stimatore nemmeno con $n = 300$. La fig. 3.11 rappresenta, invece, lo stesso grafico delle stime di p di fig. 3.7 con l'aggiunta delle deviazioni standard.

Numero di replicazioni costante

In secondo luogo, è stato analizzato il comportamento delle stime con la numerosità campionaria fissata $n = 6$; il numero di geni G assume i valori 100, 300, 500, 700, 900, 1000, 2000, 3000, 4000, 5000. La fig. 3.12 riporta l'andamento dello stimatore di Smyth. Si osserva un piccolo miglioramento della stima,

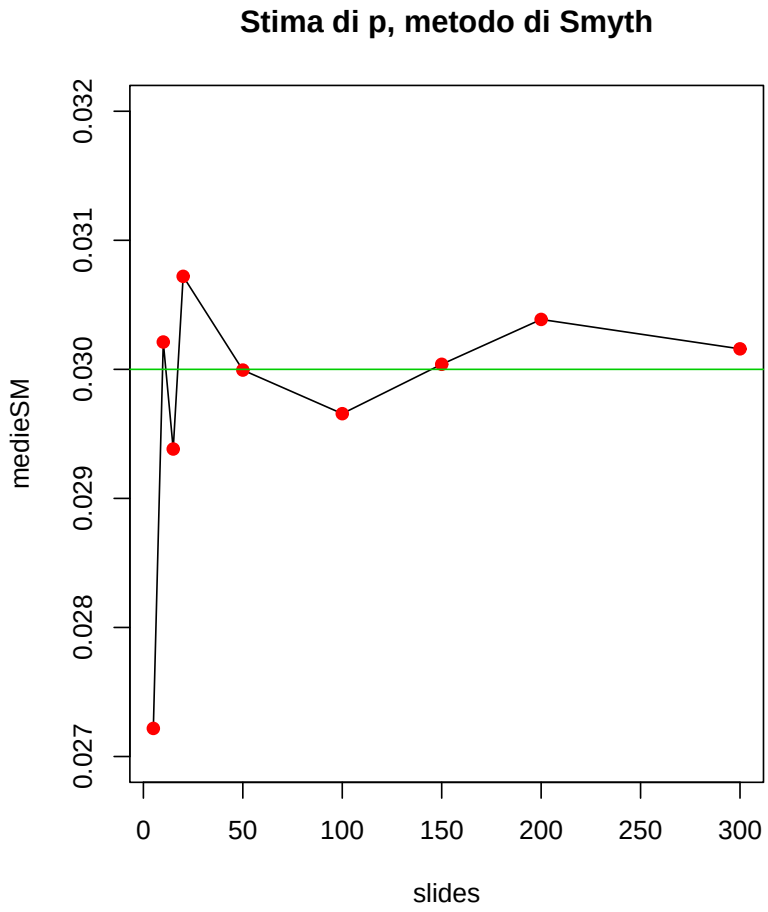

Figura 3.7: Stime di p al variare di n, metodo di Smyth.

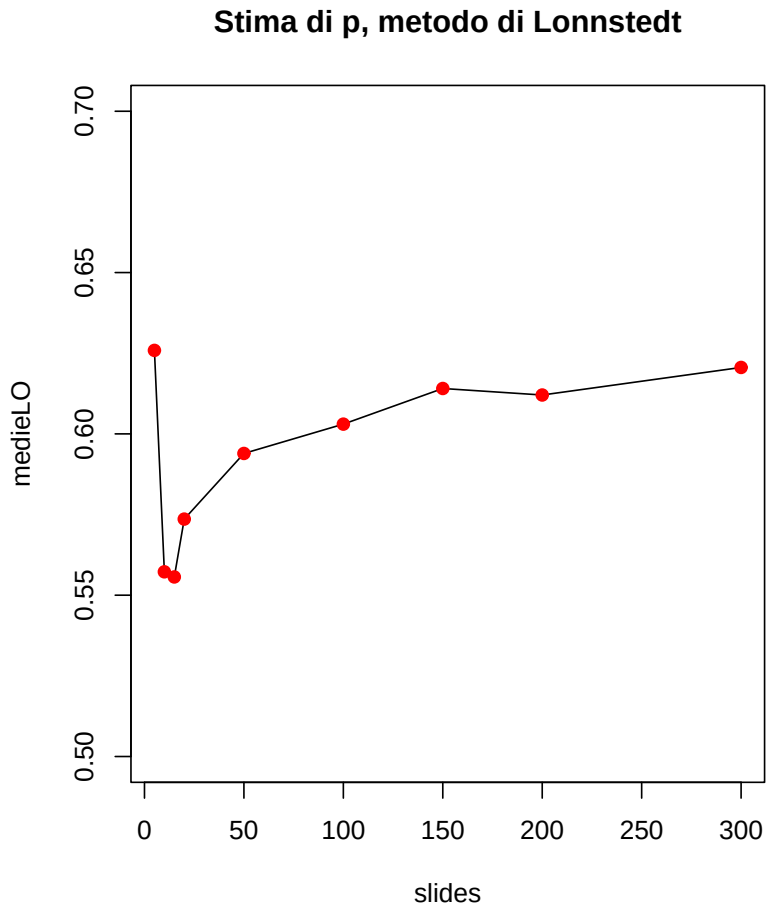

Figura 3.8: Stime di p al variare di n, metodo di Lönnstedt.

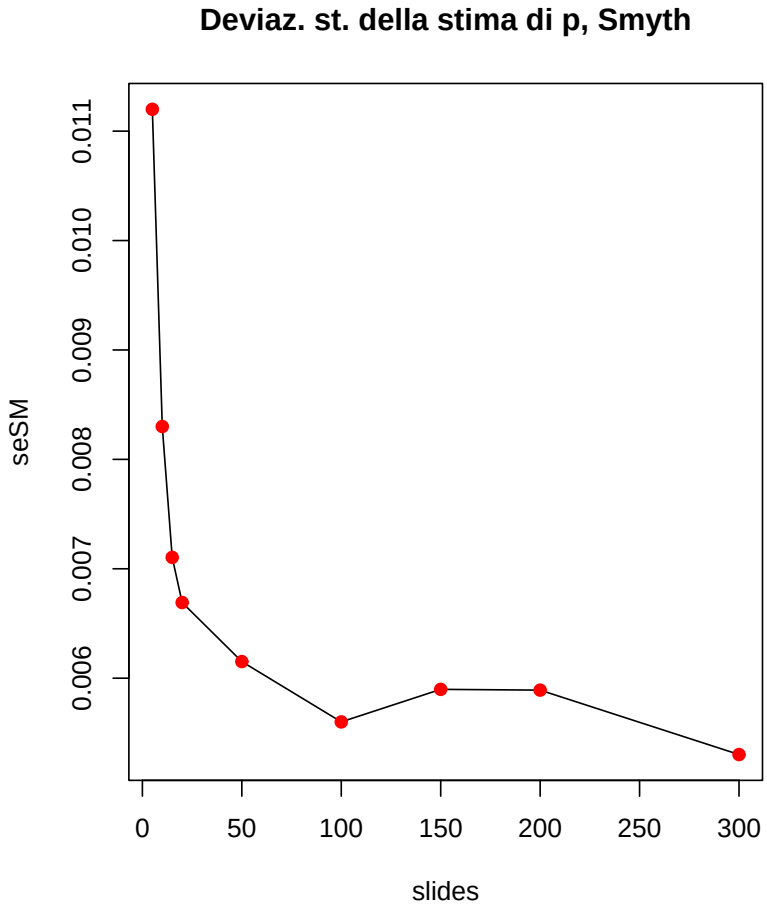

Figura 3.9: Grafico delle deviazioni standard delle stime di p (Smyth) all'aumentare di n.

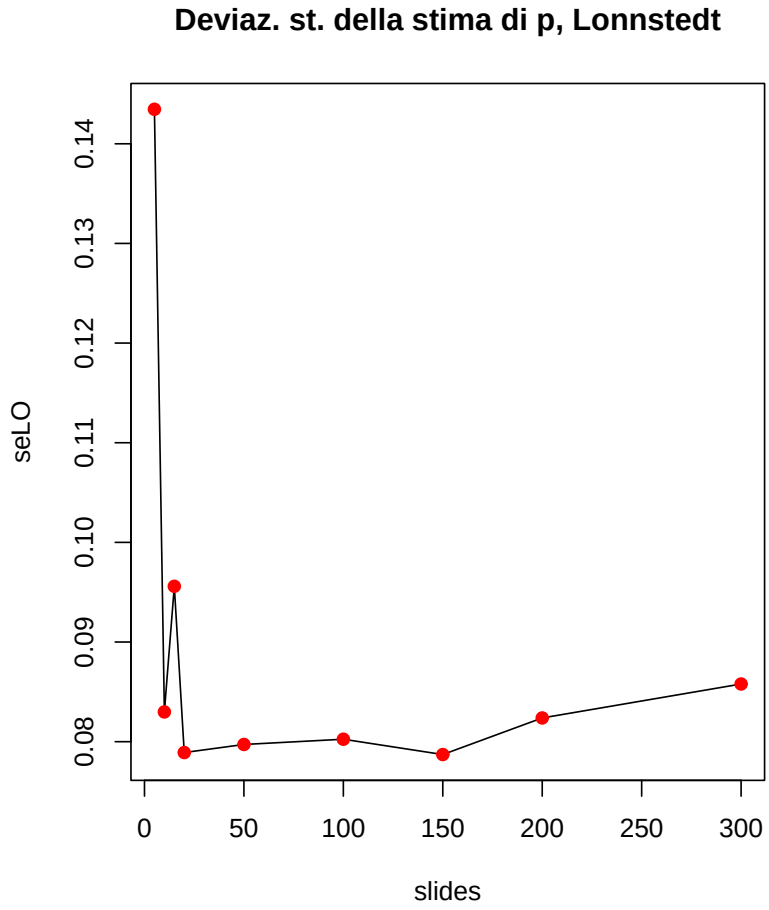

Figura 3.10: Grafico delle deviazioni standard delle stime di p (Lönnstedt) all'aumentare di n.

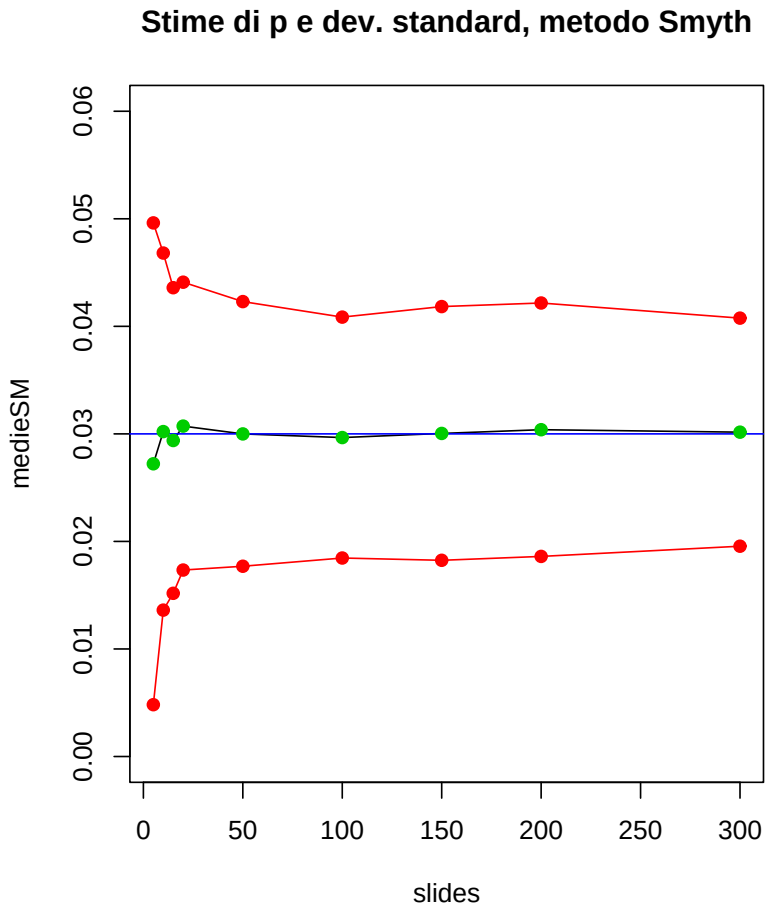

Figura 3.11: Stime di p e deviazioni standard, metodo di Smyth.

con un numero di geni grande. Il progresso non è però molto evidente, e c'è una piccola distorsione; questo può essere dovuto anche all'esigua numerosità campionaria considerata, nonostante nella pratica sia frequente disporre di queste replicazioni. La fig. 3.13 presenta lo stesso grafico con $n = 50$, e si può notare che la precisione delle stime è maggiore. Si ha conferma invece, dalla fig. 3.14, che il metodo di Lönnstedt non sia in grado di individuare il parametro p con accuratezza.

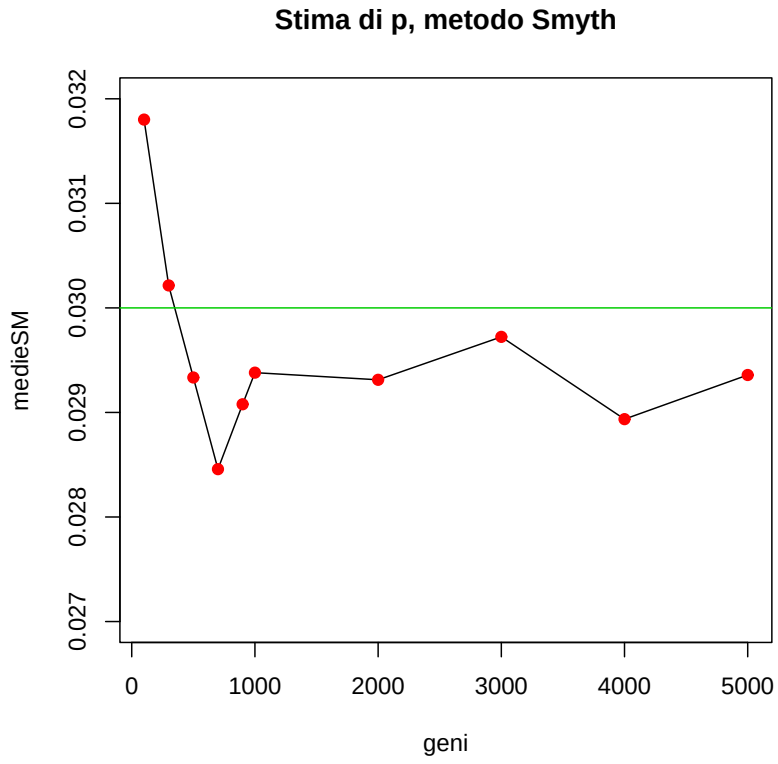

Figura 3.12: Stime di p al variare di G, metodo di Smyth.

Anche in questo caso ci si attende che un valore grande di G contribuisca a ridurre la variabilità delle stime. Le figg. 3.15 e 3.16 raffigurano l'andamento delle deviazioni standard all'aumentare di G. Il risultato è simile a quello

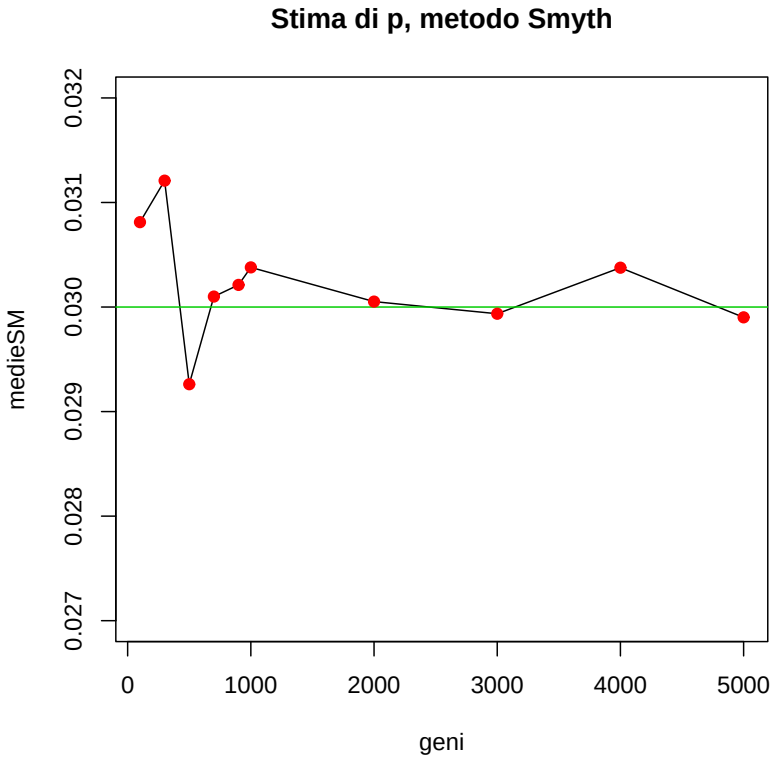

Figura 3.13: Stime di p al variare di G, metodo di Smyth.

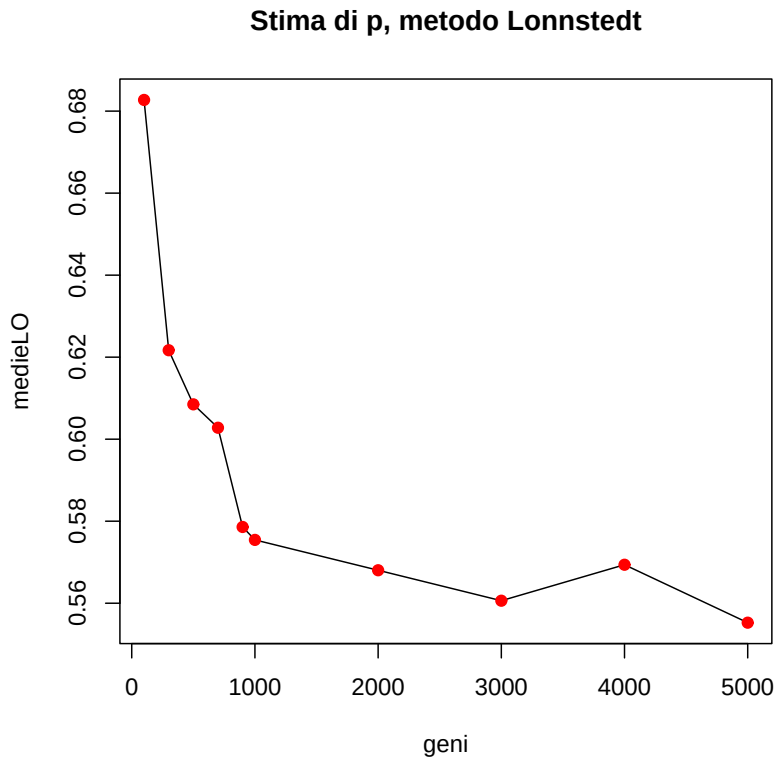

Figura 3.14: Stime di p al variare di G, metodo di Lönnstedt.

ottenuto per varie numerosità campionarie, poiché si osserva una diminuzione dell'incertezza in entrambi gli stimatori; il metodo proposto da Lönnstedt & Britton ha, in ogni caso, variabilità maggiore.

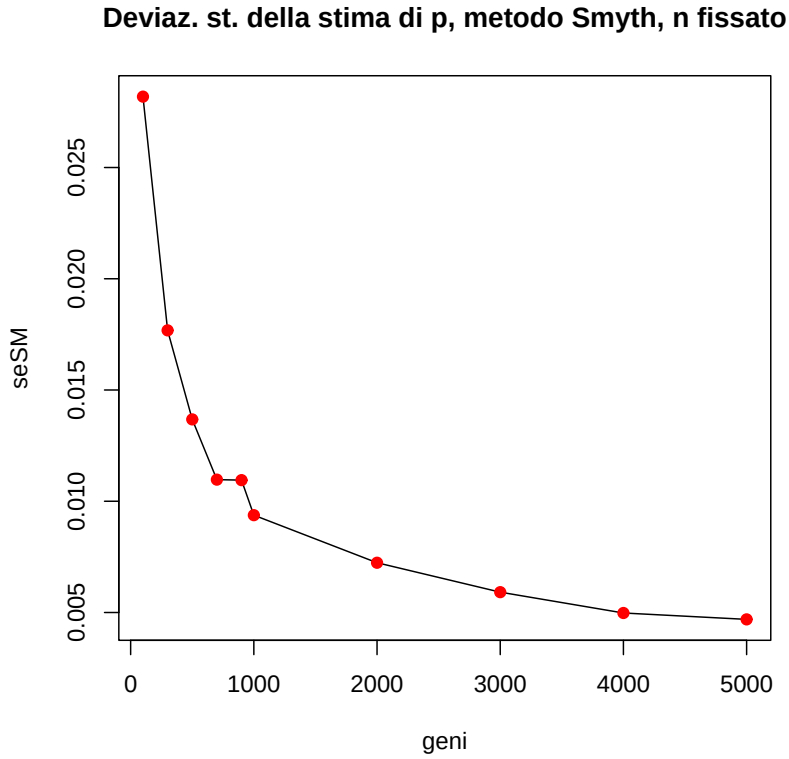

Figura 3.15: Grafico delle deviazioni standard delle stime di p (Smyth) all'aumentare di G.

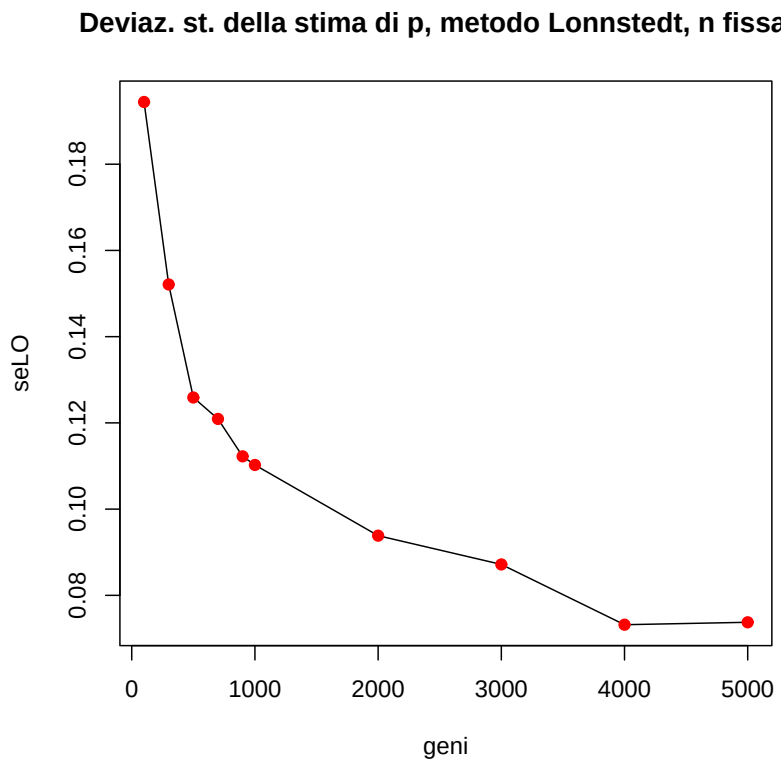

Figura 3.16: Grafico delle deviazioni standard delle stime di p (Lönnstedt) all'aumentare di G.

3.3.5 Analisi dei falsi positivi e negativi

È importante a questo punto considerare un aspetto fondamentale: gli errori di identificazione. In ogni analisi che si conduca, c'è il rischio che alcuni geni differenzialmente espressi non vengano riconosciuti (falsi negativi) e, dall'altro lato, che alcuni geni vengano dichiarati differenzialmente espressi quando in realtà non lo sono (falsi positivi). Nelle simulazioni svolte sinora, si è constatato come il metodo di Smyth funzioni abbastanza bene per specificare la percentuale di geni da considerarsi espressi, ma, oltre al numero, è importante verificare che i geni identificati come differenzialmente espressi siano effettivamente tali. Si vuole evitare che la stima appaia convincente pur operando in modo errato, ad esempio compensando i due errori. Attraverso le simulazioni è possibile valutare questo aspetto. In particolare, si analizzano gli errori al variare di n e G, che assumono i valori descritti al par. 3.3.4.

È legittimo attendersi che all'aumentare dei geni in un esperimento cresca anche il numero di errori commessi in senso assoluto; le figg. 3.17 e 3.18 ne danno conferma.
All'opposto, parallelamente all'incremento delle replicazioni, ci si attende un decremento degli errori. La fig. 3.19 mostra la percentuale d'errore al crescere delle replicazioni (calcolata come somma dei falsi negativi e positivi diviso il numero totale di geni). Ovviamente, gli errori commessi calano rapidamente con n elevato: più replicazioni sono disponibili per ogni gene, più si riesce a stimare meglio la varianza della y e, di conseguenza, identificare un gene differenzialmente espresso risulta più facile. Non si può applicare lo stesso ragionamento per il comportamento degli errori all'aumentare del numero di geni (fig. 3.20). Intuitivamente, il fatto di aggiungere geni ad un esperimento non dovrebbe avere conseguenze sulla percentuale di errori commessi: è vero, però, che anche in questo caso la conoscenza portata dal numero di geni si riflette sulle varianze delle stime, che quindi sono calcolate con maggior precisione. Una stima dell'errore più precisa, quindi, si trasforma comunque in un vantaggio per la stima di p.

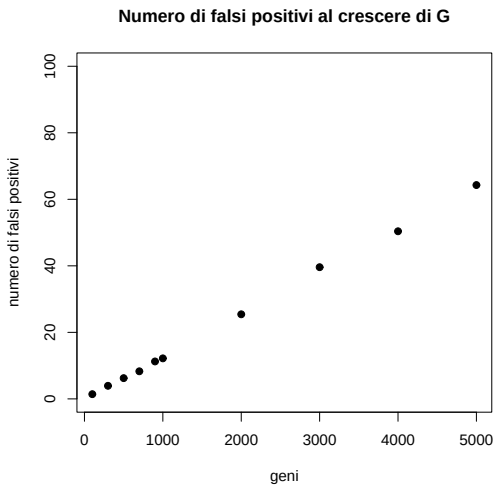

Figura 3.17: Numero dei falsi positivi all'aumentare di G.

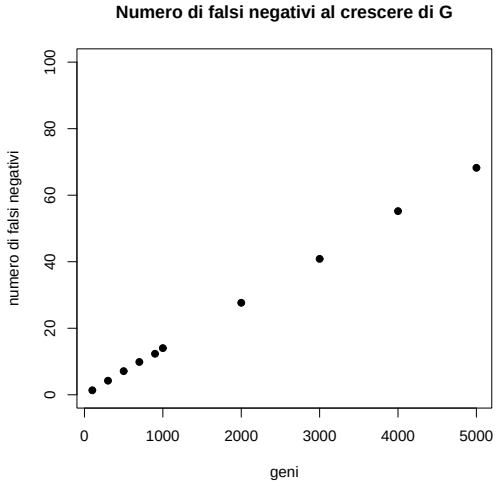

Figura 3.18: Numero dei falsi negativi all'aumentare di G.

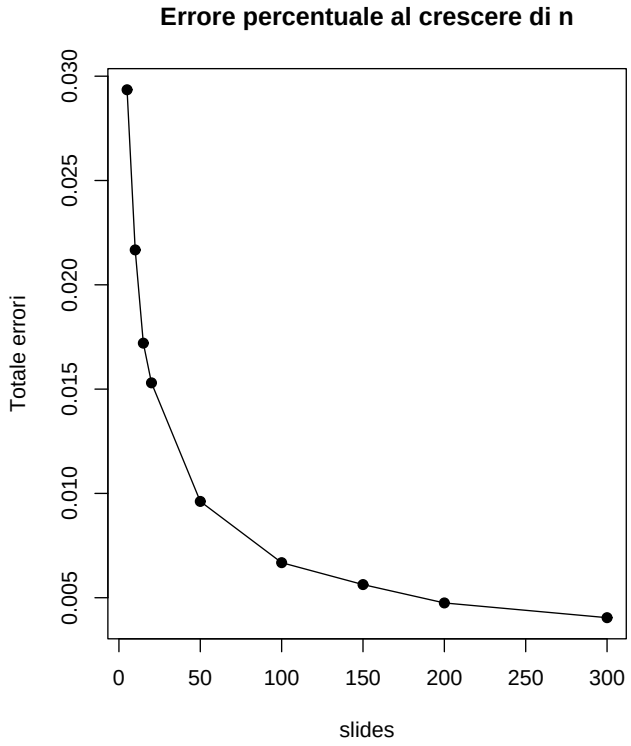

Figura 3.19: Grafico dell'errore percentuale al crescere di n.

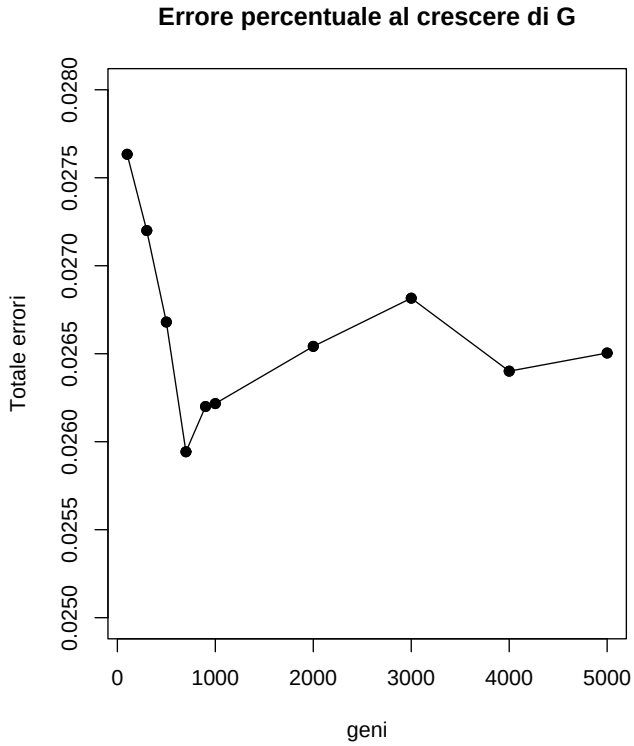

Figura 3.20: Grafico dell'errore percentuale al crescere di G.

3.4 Cenno al *false discovery rate*

Un metodo implicito per la stima di p può essere ottenuto attraverso il *false discovery rate* (Benjamini & Hochberg, 1995). È un criterio di aggiustamento dei valori-p dei test effettuati singolarmente sui geni, che si prefigge di tenere controllato l'errore di primo tipo α. Di conseguenza, una volta fissata una soglia di tolleranza (ad esempio $\alpha = 0.05$ oppure $\alpha = 0.01$), si riesce a risalire al numero di geni differenzialmente espressi identificati. In particolare, in questo elaborato sono stati considerati i valori dei test t-moderati forniti da LIMMA assieme alle statistiche B, con la correzione dell'FDR.
La fig. 3.21 rappresenta l'andamento delle "stime" di p all'aumentare del numero di geni, con $n = 6$ e $\alpha = 0.05$; il parametro p viene nettamente sottostimato (la media delle stime è 0.00958), d'altra parte, questo metodo ha natura conservativa per costruzione. Un comportamento inaspettato delle stime si osserva all'aumentare delle replicazioni (come in precedenza, il numero di geni è fissato $G = 1000$): la fig. 3.22 mostra chiaramente che la stima converge al vero valore $p = 0.03$.

3.5 Considerazioni conclusive

Attraverso queste analisi si è potuto studiare il comportamento di due stimatori del parametro p. Alla luce dei risultati ottenuti, è naturale constatare che la proposta di Smyth stima il parametro con successo, nella maggior parte delle simulazioni effettuate. È chiaro che nella pratica il numero di replicazioni per ogni gene è quasi sempre esiguo (spesso inferiore a una decina), e che questo va ad influenzare il livello di precisione. Di conseguenza, assume ancora più importanza l'informazione che si riesce a ricavare dall'insieme di geni, anche perché negli esperimenti di *microarray* se ne possono analizzare molte migliaia contemporaneamente. Va sottolineato anche che il modello stimato con il valore di p calcolato riesce a stimare molto bene gli altri iperparametri d_0, s_0^2 e v_0. Lo stimatore di Lönnstedt & Britton, invece, si dimostra inefficace sin dall'inizio, in quanto, almeno empiricamente, non converge al vero p.

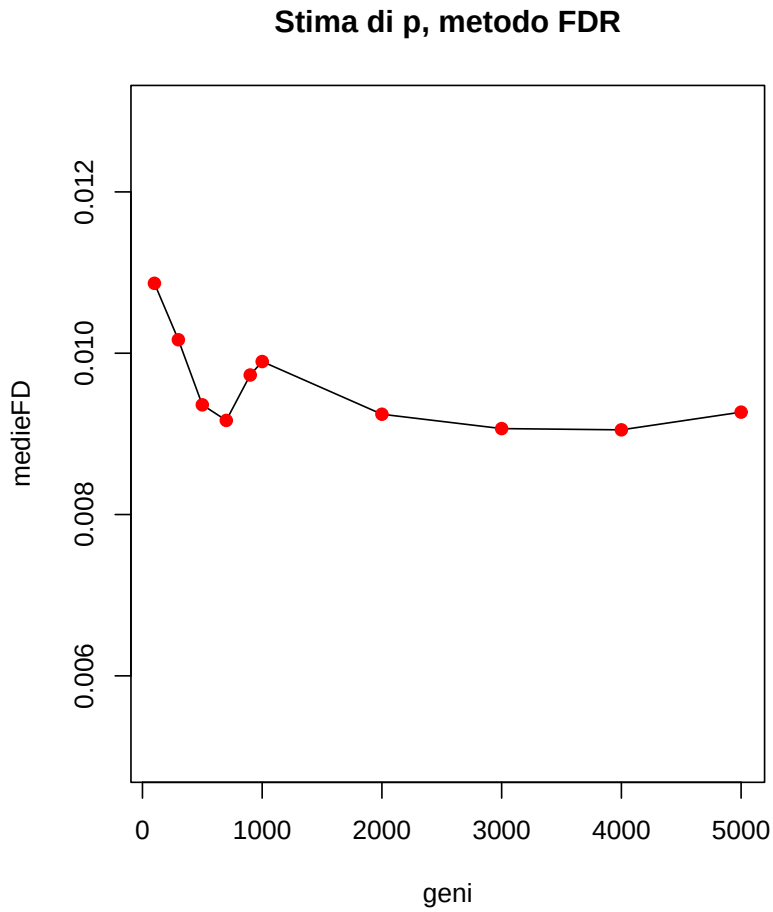

Figura 3.21: Grafico delle stime di p (FDR) all'aumentare di G.

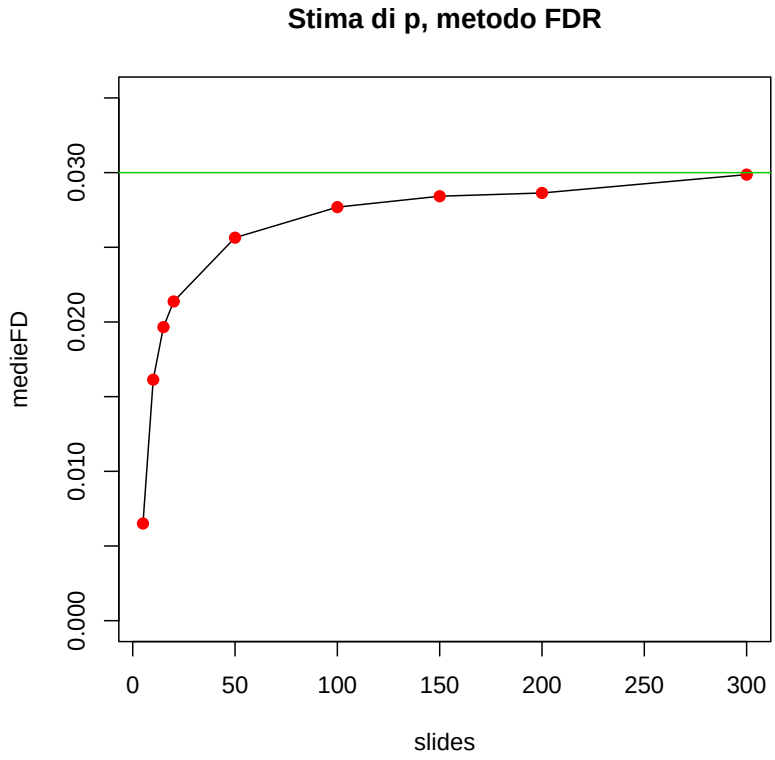

Figura 3.22: Grafico delle stime di p (FDR) all'aumentare di n.

Capitolo 4

Un'applicazione a dati reali

In questo capitolo si propone un'applicazione a dati reali del metodo di Smyth implementato nel capitolo precedente. Nel par. 4.1 viene descritta la patologia a cui si riferiscono i dati analizzati, il cancro, mentre nel par. 4.2 si presentano i dati utilizzati. Nel par. 4.3 viene esposta l'analisi svolta attraverso LIMMA, unita alla ricerca del valore di p.

4.1 Introduzione

Una patologia che può trarre sicuri benefici dallo studio approfondito del genoma è il cancro. È una malattia genetica, in sostanza, poiché deriva da variazioni patologiche dell'informazione portata dal DNA. Differisce dalle altre malattie genetiche perché è dovuto a mutazioni che colpiscono singole cellule sparse in un organismo, e non a mutazioni della linea germinale, che si trasmettono con le cellule germinali, da cui potrebbe svilupparsi un intero organismo pluricellulare. Nello sviluppo della malattia, viene violata una condizione fondamentale per il mantenimento dell'ordine nel corpo, ossia che ogni singola cellula armonizzi il suo comportamento alle esigenze dell'organismo nel suo insieme. Una cellula deve dividersi quando sono necessarie nuove cellule di quel tipo, e astenersi dal farlo quando non lo sono; deve mantenere le sue caratteristiche e occupare il posto che le compete. Naturalmente, in un organismo grande, non fa gran

danno una cellula che occasionalmente si comporta in modo anomalo. Se, invece, una cellula subisce un'alterazione genetica che le consente di sopravvivere e moltiplicarsi, mentre non dovrebbe, generando altre cellule che manifestano lo stesso comportamento deviante, l'organizzazione del tessuto può risultare compromessa: questa è la catastrofe che si verifica nel cancro. Le cellule cancerose si definiscono in base a due caratteri ereditabili: si riproducono incuranti delle limitazioni previste normalmente e invadono "territori" solitamente riservati ad altre cellule. Tramite esperimenti di *microarray* è possibile valutare la diversità di tali cellule in termini di espressione di geni, questione che risulta fondamentale per la messa a punto di terapie geniche.

4.2 I dati

I dati analizzati riguardano il tumore al seno. È importante sottolineare che provengono da un esperimento effettuato su linee cellulari, e di conseguenza si riferiscono a replicazioni di cellule "tipo" che contengono variabilità dovuta alle condizioni sperimentali e a casualità, ma non risentono della variabilità tra soggetti che si riscontra normalmente nelle popolazioni. Per questo i risultati delle analisi non possono essere automaticamente estesi alla popolazione.

L'obiettivo dello studio è confrontare cellule malate non trattate con cellule malate trattate con un ormone, al fine di valutare le diverse risposte dei geni al trattamento. L'esperimento è stato effettuato su 960 geni e 6 *microarrays*; ogni vetrino è stato preparato con due replicazioni per ogni gene ed alcuni *spot* di controllo. Alle tre coppie di *slides* è stata applicata la tecnica del *dye-swap*. Le immagini sono state ottenute con il software `GenePix`, che fornisce i seguenti files:

- un file per ogni *array* analizzato, contenente varie informazioni sui geni e alcune statistiche;

- un file contenente informazioni sugli *spot* di controllo (*gal file*);

- un file contenente i tipi di *spot*, ognuno associato ad un colore per le rappresentazioni grafiche.

Una volta ottenute le intensità, sono stati "rimossi" i geni saturati e gli spot di controllo, assegnando loro peso 0.1. Infine, i valori sono stati normalizzati, sia tra *array* che all'interno degli stessi, utilizzando il metodo *loess* (che è uno dei più usati dai biologi). Le figg. 4.1, 4.2 e 4.3 rappresentano gli *MA-plot* normalizzati dei 6 *slides*.

4.3 L'analisi dell'espressione genica

A questo punto è stato possibile stimare un modello lineare per ogni gene attraverso la funzione `lmFit`. Considerando l'applicazione del *dye-swap* e rispettando l'ordine di caricamento dei file, la matrice del disegno X è stata fissata a

$$X = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 1 \\ -1 \end{pmatrix}.$$

Ottenuti i modelli, sono state calcolate le statistiche B e gli iperparametri (escluso p) tramite la funzione `eBayes`, fissando la proporzione di geni differenzialmente espressi $p = 0.01$. Le stime ricavate sono presentate nella tabella 4.1. Quella che sembra meno affidabile è la stima del parametro v_0, dato che

$d_0 = 1.5013$
$s_0^2 = 0.0144$
$v_0 = 37.9572$

Tabella 4.1: Stime degli iperparametri con $p = 0.01$

assume un valore abbastanza elevato; non bisogna dimenticare, d'altra parte,

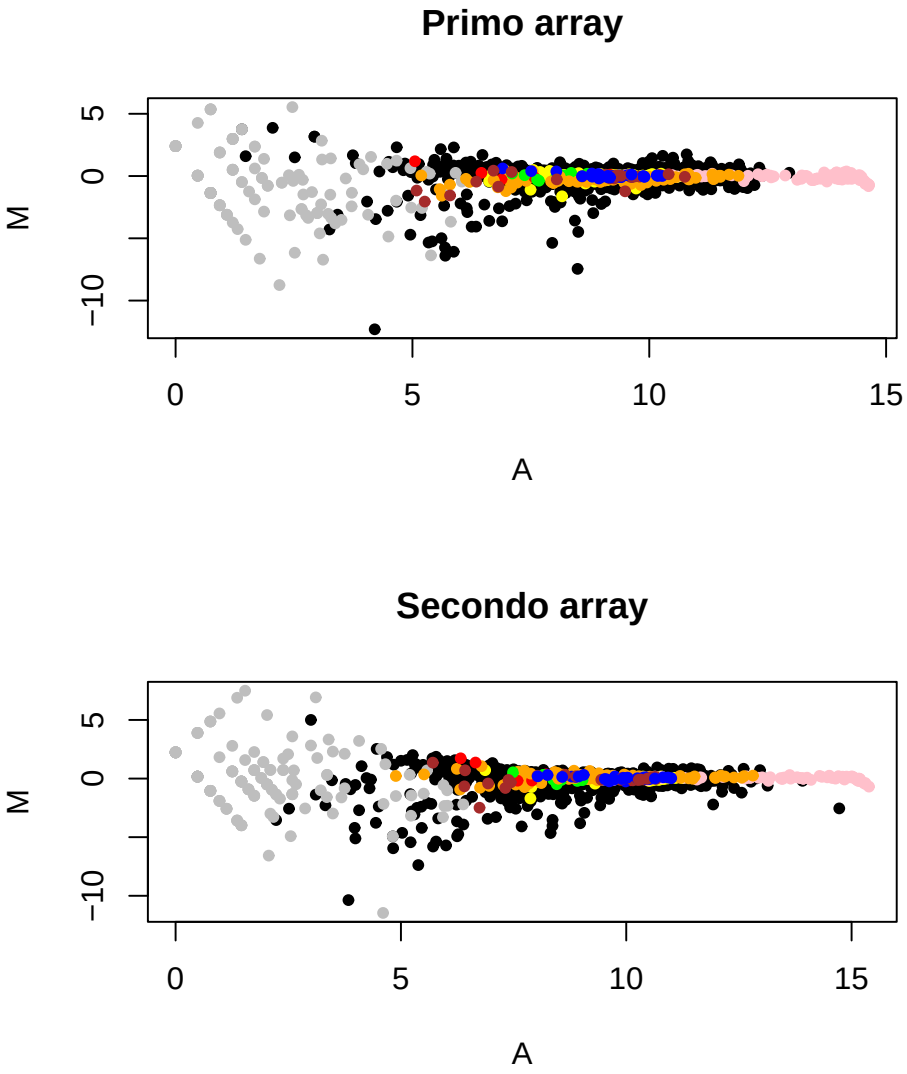

Figura 4.1: Grafici MA dei primi due *arrays*.

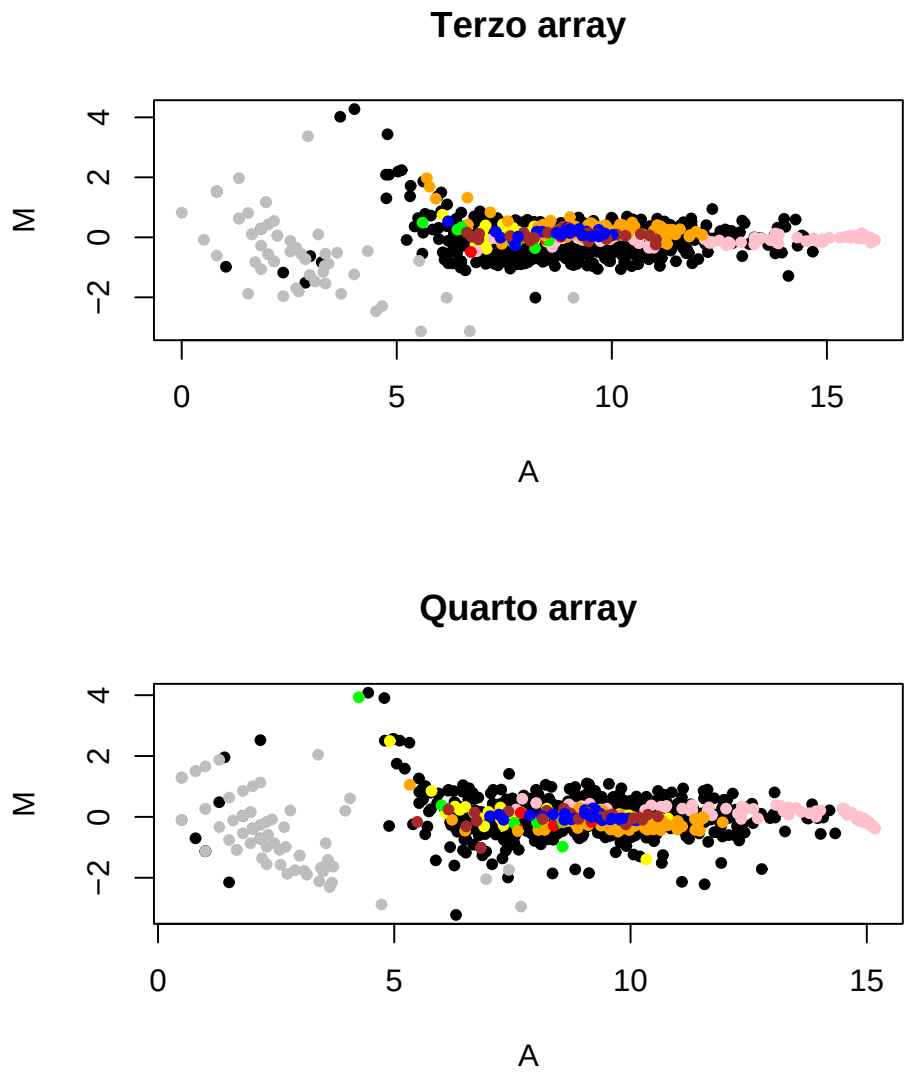

Figura 4.2: Grafici MA del terzo e quarto *array*.

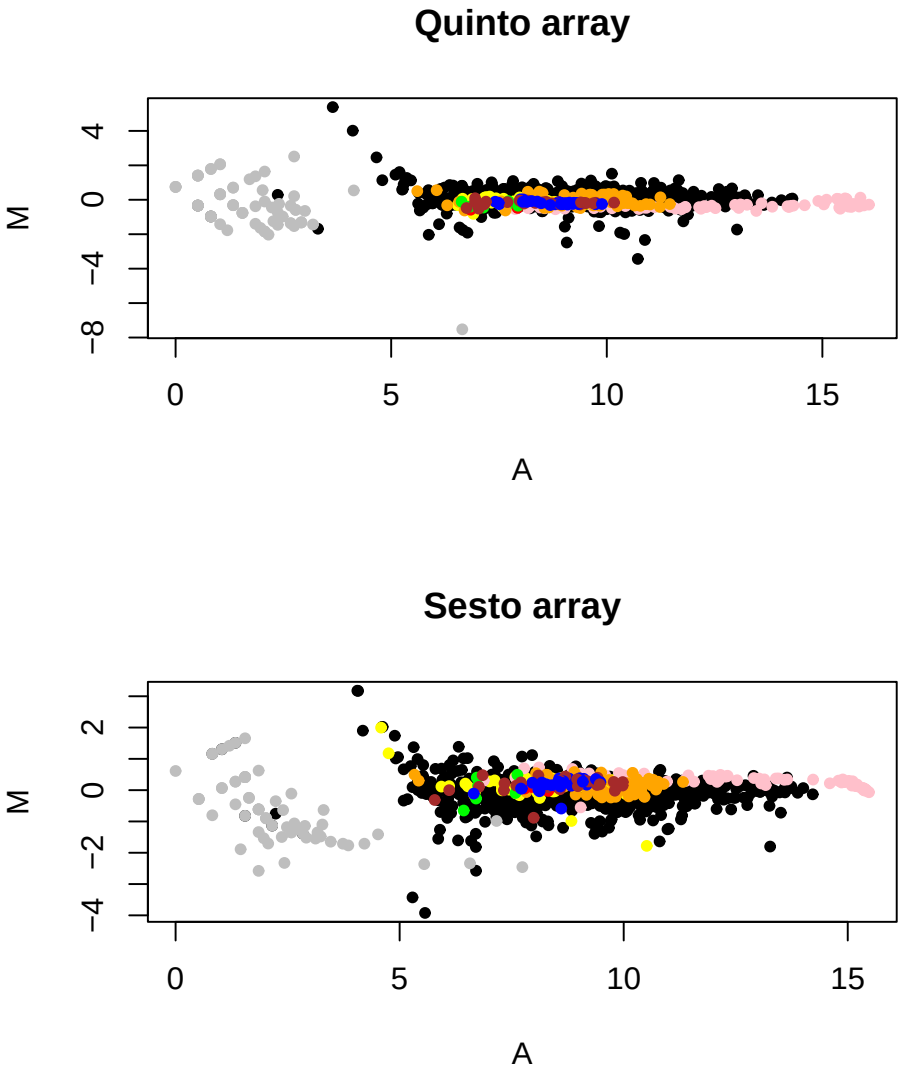

Figura 4.3: Grafici MA del quinto e sesto *array*.

che LIMMA pone dei limiti alla stima della quantità $\sqrt{v_0 s_0^2}$, e di conseguenza i due parametri v_0 e s_0^2 sono in qualche modo legati. In fig. 4.4 è rappresentato l'istogramma dei valori di B ottenuti con questo modello, mentre la fig. 4.5 mostra il cosiddetto "vulcano *plot*". Ci si aspetta che i valori più grandi indichino geni differenzialmente espressi.

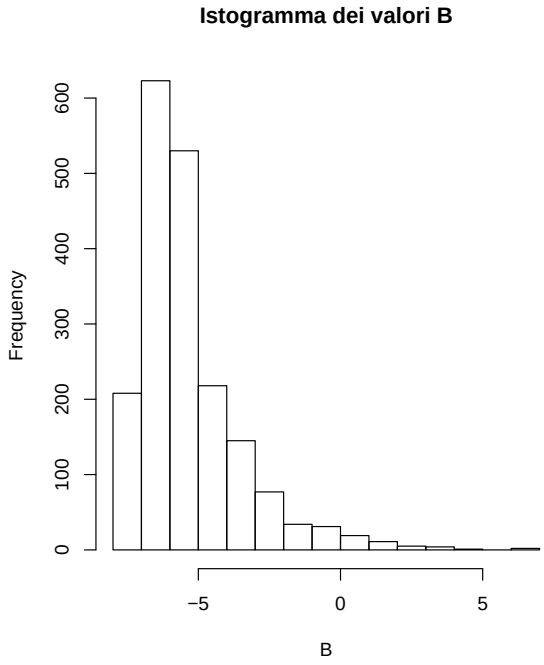

Figura 4.4: Istogramma dei valori di B stimati con $p = 0.01$.

A conclusione dell'analisi, solitamente, si decide arbitrariamente quanti geni considerare differenzialmente espressi, prendendo ad esempio i primi 50, o i primi 100 ordinati secondo il valore di B. La tabella 4.4 presenta le statistiche t, B e i valori-p dei primi 20 *spot* identificati da questo modello (i nomi dei geni sono di fantasia per garantire la riservatezza delle informazioni).

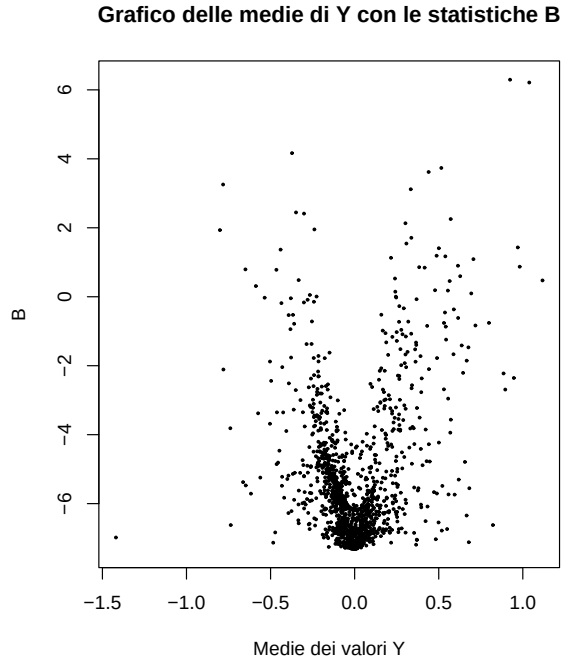

Figura 4.5: Grafico delle medie dei valori di Y con le statistiche B stimate (vulcano *plot*).

Index	Name	t	P.Value	B
813	AAA2	17.0944	0.0013	6.2952
333	AAA1	16.7179	0.0013	6.2130
278	BBB1	-12.0713	0.0069	4.1649
1514	EEE1	11.2081	0.0073	3.7354
344	FFF1	11.0028	0.0073	3.6183
1610	CCC1	-10.5036	0.0078	3.2546
1698	DDD1	10.2866	0.0078	3.1178
753	MMM1	10.7733	0.0116	2.5294
758	BBB2	-9.1554	0.0116	2.4455
1484	HHH1	-9.1072	0.0116	2.4117
524	III1	8.8842	0.0116	2.2529
1419	GGG1	8.8522	0.0116	2.1316
1306	LLL1	-8.5475	0.0123	1.9524
1130	CCC2	-8.5878	0.0123	1.9331
526	TTT1	8.0833	0.0158	1.7093
1756	UUU1	7.7923	0.0158	1.5427
641	OOO1	7.8841	0.0158	1.4305
525	PPP1	7.8557	0.0158	1.4073
251	NNN1	-7.8725	0.0158	1.3664
1034	EEE2	7.5198	0.0161	1.1891

Tabella 4.2: Tabella dei primi 20 *spots* identificati con $p = 0.01$

Implementazione del metodo iterativo

Partendo dal valore $p = 0.01$ e applicando il metodo di Smyth, sono state effettuate 100 iterazioni sul modello per ottenere la stima del parametro: il valore raggiunto è $p = 0.3573$, ossia il 35.73% degli *spots* analizzati risulta differenzialmente espresso (per ottenere il numero di geni differenzialmente espressi è sufficiente dividere questo risultato per due, essendo ogni gene considerato due volte negli *array*). In fig. 4.6 è rappresentata la successione dei valori. Rispetto

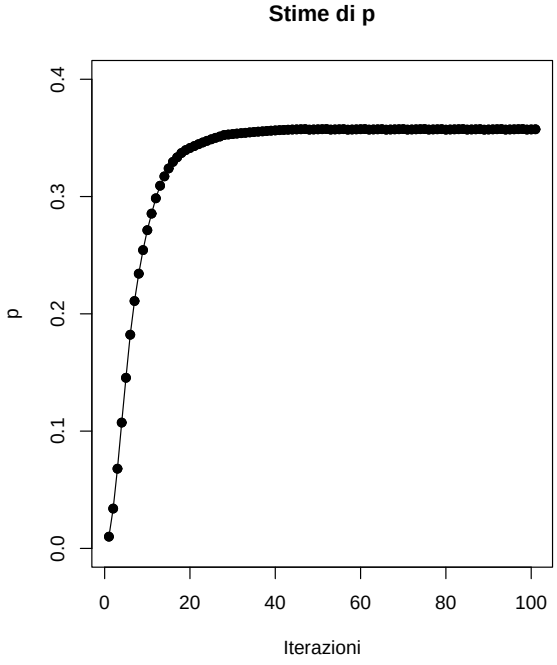

Figura 4.6: Grafico dei valori di p stimati nelle 100 iterazioni.

alle simulazioni effettuate nel capitolo precedente, e alle aspettative che si hanno a priori, il valore di p sembra elevato. Due sono le possibili spiegazioni:

- i geni differenzialmente espressi sono effettivamente il 35.73% del totale, che potrebbe essere ragionevole, dato che l'esperimento è stato effettuato considerando geni selezionati in base a conoscenze precedenti;

- il modello utilizzato per determinare p non riesce a spiegare questi dati, che non può essere assolutamente escluso, data la peculiarità di questo tipo di esperimenti.

Con questa nuova proporzione p, gli altri iperparametri risultano (tabella 4.3):

$d_0 = 1.5013$
$s_0^2 = 0.01442$
$v_0 = 4.1979$

Tabella 4.3: Stime degli iperparametri con $p = 0.3573$

Le stime dei parametri d_0 ed s_0^2 non sono cambiate, mentre v_0 ha assunto un valore più ragionevole. Naturalmente, anche i valori di B vengono modificati; le figg. 4.7 e 4.8 rappresentano rispettivamente l'istogramma e il vulcano *plot* dei valori di B stimati con questo nuovo modello.

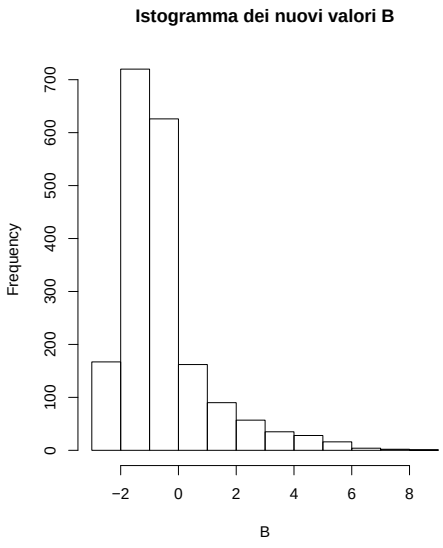

Figura 4.7: Istogramma dei valori di B stimati con $p = 0.3573$.

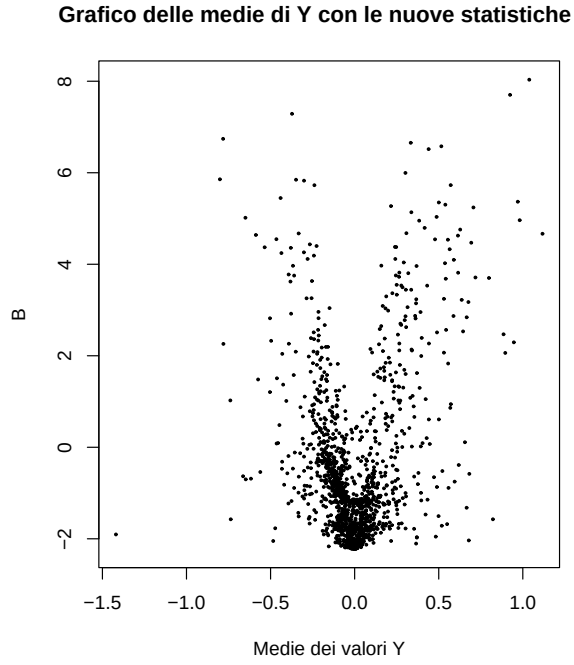

Figura 4.8: Grafico delle medie dei valori di Y con le nuove statistiche B stimate con $p = 0.3573$ (vulcano *plot*).

Nella tabella 4.2, invece, sono riportati i primi 20 geni selezionati da questo secondo modello: è evidente che i valori delle statistiche sono diversi dai precedenti, ed anche l'ordinamento dei geni presenta delle piccole variazioni con il nuovo parametro p; si può comunque presumere che siano dovute ad arrotondamenti. La maggior parte di essi, tuttavia, viene identificata in entrambi i modelli.

Index	**Name**	**t**	**P.Value**	**B**
333	AAA1	16.7179	0.0013	8.0336
813	AAA2	17.0944	0.0013	7.7027
278	BBB1	-12.0713	0.0068	7.2882
1610	CCC1	-10.5036	0.0079	6.7413
1698	DDD1	10.2869	0.0079	6.6545
1514	EEE1	11.2081	0.0073	6.5780
344	FFF1	11.0028	0.0073	6.5162
1419	GGG1	8.8522	0.0120	5.9960
1130	CCC2	-8.5878	0.0123	5.8567
758	BBB2	-9.1554	0.0117	5.8476
1484	HHH1	-9.1072	0.0117	5.8270
524	III1	8.8842	0.0117	5.7294
1306	LLL1	-8.5474	0.0123	5.7275
753	MMM1	10.7732	0.0117	5.5233
251	NNN1	-7.8725	0.0158	5.4471
641	OOO1	7.8841	0.0158	5.3666
525	PPP1	7.8557	0.0158	5.3501
1803	QQQ1	7.6388	0.0162	5.3021
1771	RRR1	7.5892	0.0162	5.2706
35	SSS1	7.5437	0.0162	5.2414

Tabella 4.4: Tabella dei primi 20 *spots* identificati con $p = 0.3573$

Capitolo 5

Conclusioni

Nel corso di questa tesi si è cercato di comprendere le caratteristiche delle procedure di stima del parametro p all'interno dell'analisi di esperimenti di *microarray*, con particolare riferimento al modello bayesiano empirico sviluppato in Lönnstedt & Speed (2002) e ripreso in Smyth (2004). Sono stati considerati due metodi di stima proposti in letteratura: uno da Smyth (2004) e uno da Lönnstedt & Britton (2005).

Poiché non esiste uno studio delle proprietà formali di queste stime, in questa tesi si è effettuato uno studio empirico di tali proprietà, attraverso diverse simulazioni. Con riferimento ad un'esemplificazione del modello di Smyth (2004), è stato costruito un grafo (fig. 2.2), sulla base del quale è stato possibile simulare un ipotetico insieme di dati; si sono fissati arbitrariamente i quattro iperparametri p, v_0, d_0 ed s_0^2, e successivamente sono state generate le rimanenti variabili. Si è studiato il comportamento degli stimatori di p al variare di n, la numerosità campionaria, e G, il numero di geni, apprendendo non solo che è utile all'analisi disporre di un gran numero di replicazioni per ogni gene, ma che anche il numero di geni stesso può influire sulla bontà delle stime dei parametri.

Si è visto che i due stimatori considerati si comportano in modo molto diverso tra loro: quello di Lönnstedt & Britton ha prodotto stime molto distorte in tutte le simulazioni effettuate, mentre quello di Smyth ha rivelato buone capacità di stima. I risultati che sono stati ottenuti incoraggiano, quindi, l'uso dello

stimatore di Smyth che, non solo è nettamente superiore all'altro proposto in termini di distorsione, ma sembra essere anche molto efficiente.
È importante ricordare che i risultati sono stati raggiunti a partire da dati simulati, che, per costruzione, soddisfano gli assunti del modello. Per questo motivo, il passaggio ai dati reali non è immediato, in particolare è da attendersi che non tutte le assunzioni siano soddisfatte. Inoltre, riportandosi nei casi reali in cui spesso si dispone di scarsa informazione, ci si rende immediatamente conto di quanto poco si conosca questo tipo di dati. Per ottenerli, di fatto, occorrono diversi passaggi, ciascuno soggetto ad errori e a valutazioni soggettive; le scelte che si compiono, riguardo ad esempio all'analisi delle immagini o alle tecniche di normalizzazione, fanno parte esse stesse dell'esperimento ed in molti casi non si conosce l'effetto che abbiano sui dati. Nei modelli che si propongono diventa difficile quantificare tutte queste informazioni, e di conseguenza ottenere dati in qualche modo "depurati".
Successivamente, si è applicata la procedura di Smyth su dati provenienti da un esperimento reale, effettuato su linee cellulari e riguardante geni presumibilmente coinvolti nel tumore al seno. Si è ottenuta una stima della proporzione di geni differenzialmente espressi in qualche modo inattesa, in relazione all'ordine di grandezza; questo risultato ha sollevato alcuni dubbi sull'adattamento del modello utilizzato ai dati.

I metodi bayesiani e bayesiani empirici sembrano adeguati all'analisi di questi esperimenti. In particolare, la flessibilità che concedono può essere molto importante per includere nei modelli le conoscenze a priori che si riescono a raccogliere, soprattutto per quel che riguarda la variabilità delle misurazioni. Proseguire su questa strada, quindi, potrà portare a miglioramenti nella precisione delle analisi svolte.

Uno studio della robustezza del metodo di stima di Smyth potrebbe essere un naturale proseguimento di questo lavoro, poiché arricchirebbe sicuramente le conoscenze sia su questo parametro, ovvero la proporzione di geni differenzialmente espressi, sia, conseguentemente, sull'effettiva espressione dei geni analizzati.

Appendice A

Codice R utilizzato nelle simulazioni

Simulazioni con un numero di geni costante

```
seSM<-c()
seLO<-c()
seFD<-c()
medieSM<-c()
medieLO<-c()
medieFD<-c()
totfalsi<-c()
totfalsi2<-c()
totqualiSM<-c()
totneg<-c()
totpos<-c()
slides<-c(5,10,15,20,50,100,150,200,300)
L<-300  #numero di simulazioni
G<-1000 #numero di geni, fissato
d0<-3
s2_0<-1
v0<-4
```

```
p<-0.03  #proporzione di geni diff. espressi

#####################################################################

for (n in slides){
qualiSM<-c()
lista<-c()  #lista che conterrà i valori di p stimati (Smyth)
listap<-c()
listalonn<-c() #lista che conterrà i valori di p stimati (Lönnst.)
listafdr<-c()  #lista che conterrà i valori di p stimati (fdr)
par<-matrix(0,nrow=L,ncol=3,byrow=T)
tot<-c()
totSM<-c()
listaazz<-c()
fpos<-c()
fneg<-c()
 for (j in 1:L){
 cont<-0
 m<-matrix(0,nrow=G,ncol=n,byrow=T)
 quali<-c()    #tiene memoria dei geni realmente diff. espressi
       for (i in 1:G)
       {
         sigma2<-rchisq(1,d0)/(s2_0*d0)
         sigma2<-1/sigma2
         d<-rbinom(1,1,p)
         if (d==0) mu<-0  else
         {
            cont<-cont+1
            quali<-c(quali,i)
            mu<-rnorm(1,0,sqrt(sigma2*v0))
          }
```

```
       y<-rnorm(n,mu,sqrt(sigma2))
       m[i,]<-y
      }
tot<-c(tot,quali)
fitm<-lmFit(m)  #stimo un modello lineare per ogni gene
p2<-0.01
    for (k in 1:15){ #ciclo per la determinazione di p (Smyth)
       #calcolo l'odds a posteriori per ogni gene (e i parametri)
      fit<-eBayes(fitm,proportion=p2)
      listap<-c(listap,p2)
       #calcolo p secondo il metodo iterativo di Smyth
      p2new<-sum(exp(fit$lods)/(1+exp(fit$lods)))/G
      if (k==15) break
      p2<-p2new
    }
ord<-order(fit$lods,decreasing=T)
#indici dei geni che il metodo Smyth dichiara diff. espressi
qualiSM<-ord[1:(p2*G)]
totSM<-c(totSM,qualiSM[order(qualiSM)])

###########ciclo per i falsi positivi e negativi#########

azz<-0
if (length(quali)==0){
  if (length(qualiSM)==0) ("Nessun gene diff. espresso")
   else {
           fpos<-c(fpos,length(qualiSM))
           fneg<-c(fneg,0)}
   }
 else {
 if (length(qualiSM)==0) {
```

```
              fneg<-c(fneg,length(quali))
              fpos<-c(fpos,0)}
 else{
   for (i in 1:(length(qualiSM))){
    mm=1
    while (mm<=length(quali)){
      if (qualiSM[i]==quali[mm]) {
        azz<-azz+1
        break}
      else mm=mm+1
    }
   }
listaazz<-c(listaazz,azz)
fneg<-c(fneg,(length(quali)-azz))
fpos<-c(fpos,(length(qualiSM)-azz))
}
}
#####################
lista<-c(lista,p2)
pfdr<-0   #calcolo p con la stat. t-moderata, aggiustando con fdr
adj<-topTable(fit,number=G,adjust="fdr")
          for (w in 1:G){
            if (adj$P.Value[w]<0.05) pfdr<-pfdr+1
          }
listafdr<-c(listafdr,pfdr/G)
k1<-(n-3)/(n-1)*mean(m^2)-1/n
k2<-((n-2)*sqrt(pi)*gamma((n-1)/2)*sqrt(n))/(2*sqrt(n-1)*
      gamma(n/2))* mean(abs(m))-1
plonn<-k2^2/(n*k1-2*k2)
listalonn<-c(listalonn,plonn)
}
```

```
medieSM<-c(medieSM,mean(lista))
medieLO<-c(medieLO,mean(listalonn))
medieFD<-c(medieFD,mean(listafdr))
seSM<-c(seSM,sqrt(var(lista)))
seLO<-c(seLO,sqrt(var(listalonn)))
seFD<-c(seFD,sqrt(var(listafdr)))
totqualiSM<-c(totqualiSM,totSM)
totfalsi<-c(totfalsi,mean((fpos+fneg)/G))
totfalsi2<-c(totfalsi2,mean((fpos+fneg)/length(quali)))
totneg<-c(totneg,mean(fneg))
totpos<-c(totpos,mean(fpos))
}
```

Semplicemente sostituendo alla variabile `slides` la variabile `geni`, contenente i valori di G per i quali si vuole svolgere l'analisi, si ottiene il listato delle simulazioni effettuate con un numero di replicazioni costante (ossia n fissato).

Bibliografia

[1] Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2005). *L'essenziale di biologia molecolare della cellula*, Zanichelli, seconda edizione.

[2] Baldi, P. & Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17, 509-519.

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

[4] Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple hypothesis testing under dependency. *Annals of Statistics.*

[5] Broët, P., Richardson, S. & Radvanyi, F. (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology*, 9, 671-683.

[6] Buck, M. J. & Lieb, J.D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83, 349-360.

[7] Cobb, G. W. (1998). *Introduction to Design and Analysis of Experiments.* Springer, New York.

[8] Draghici, S., Kuklin A., Hoff B. & Shams S. (2001). Experimental design, analysis of variance and slide quality assessment in gene expression arrays. *Current Opinion in Drug Discovery & Development*, 4 (3), 332-337.

[9] Dudoit, S., Yang, Y., Callow, M. & Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12** (1), 111-139.

[10] Dudoit, S., van der Laan, M. J. & Pollard, K. S. (2004a). Multiple testing. Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology*, **3**, No. 1, Article 13.

[11] Dudoit, S., van der Laan, M. J. & Birkner, M. D. (2004b). Multiple testing procedures for controlling tail probability error rates. (Technical report # 166, Division of Biostatistics, UC Berkeley).

[12] Efron, B., Tibshirani, R., Storey, J.D. & Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association,* 96(456), 1151-1160.

[13] Glonek, G. F. & Solomon, P. J. (2004). Factorial and time course designs for cDNA microarray experiments. *Biostatistics*, 5, 89-111.

[14] Gottardo, R., Pannucci, J. A., Kuske, C. R. & Brettin, T. (2003). Statistical analysis of microarray data: a Bayesian approach. *Biostatistics*, 4, 597-620.

[15] Huber, W., v Heydebreck, A., Sultmann, H., Poutska, A. & Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 1 (1), 1-9.

[16] Ideker, T., Thorsson, V., Siehel, A.F. & Hood, L.E. (2000). Testing for differentially-expressed genes by maximum likelihood analysis of microarray data. *Journal of Computational Biology,* 7, 819-837.

[17] Ishwaran, H. & Rao, J. S. (2003). Detecting differentially expressing genes in microarray using Bayesian model selection. *JASA*, 98, 438-455.

[18] Kendziorski, C. M., Newton, M. A., Lan, H. & Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in medicine* 22, 3899-3914.

[19] Kerr, M. K. & Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics*, 2, 183-201.

[20] Kerr, M. K., Martin, M. & Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7, 819-837.

[21] Long, A. D., Mangalam, H. J., Chan, B. Y. P., Trolleri, L., Hatfield, G. W. & Baldi, P. (2001). Global gene expression profiling in *Escherichia coli* K 12: improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. *Journal of Biol. Chem.*, 276, 19937-19944.

[22] Lönnstedt, I. & Britton, T. (2005). Two hierarchical Bayes models for cDNA microarray gene expression. *Biostatistics*, **1**, 1, 1-23.

[23] Lönnstedt, I. & Speed, T. P. (2002). Replicated microarray data. *Statistica sinica*, 12, 31-46.

[24] Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. & Tsui, K.W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8, 37-52.

[25] Parmigiani, G., Garrett, E. S., Irizarry, R. A. & Zeger, S. L. (2003). *The Analysis of Gene Expression Data: Methods and Software*, New York, Springer.

[26] Qiu, X., Klebanov, L. & Yakovlev, A. (2005). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology*, 4 (1), article 34, `http://www.bepress.com/sagmb/vol4/iss1/art34`.

[27] Smyth, G. K. (2004). Linear Models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3 (1), article 3, `http://www.bepress.com/sagmb/vol3/iss1/art3`.

[28] Tusher, V.G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA,* 98, 5116-5121.

[29] van der Laan, M. J., Dudoit, S. & Pollard, K. S. (2004a). Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology*, 3, No. 1, Article 14.

[30] van der Laan, M. J., Dudoit, S. & Pollard, K. S. (2004b). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3, No. 1, Article 15.

[31] Westfall, P. & Young, S. (1993). *Resampling-Based Multiple Testing.* Wiley, New York.

[32] Wit, E. & McClure, J. (2004). *Statistics for Microarrays. Design, analysis and inference.* Wiley.

[33] Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. & Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8, 625-637.

[34] Yang, Y.H. & Thorne, N.P. (2003). Normalization of two-channel cDNA microarray data. *In D R Goldstein, editor, Science and Statistics: A Festschrift for Terry Speed,* **Volume 40 of LMS Lecture Notes - Monograph Series**, 403-418.

[35] Yang, Y.H., Buckley, M. J., Dudoit, S. & Speed, T.P. (2002a).Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational Biology*, 11, 108-136.

[36] Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. & Speed, T.P. (2002b). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research,* 30 (4), e15.