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

Eterogeneità del tumore all'ovaio: metodi di classificazione per la comprensione della sua complessità

Relatore: Prof.ssa Chiara Romualdi
Dipartimento di Biologia

Laureando: Sara Cozzolino
Matricola: 1134642

Anno Accademico 2017/2018

Indice

Capitolo 1

Introduzione

Il tumore ovarico è al settimo posto tra i tumori più frequentemente diagnosticati nelle donne. Spesso è diagnosticato quando si presenta in stadi già avanzati, risultando in un basso tasso di sopravvivenza (poco meno del 30%). Negli ultimi anni molti studi sono stati dedicati all'approfondimento di questo tipo di tumore, tuttavia l'eziologià non è ancora del tutto chiara e il tasso di sopravvivenza globale ha visto un incremento molto modesto dal 1995 ad oggi (circa del $2-4\%$) [25].
In questa tesi affronto l'analisi di campioni di tumore ovarico epiteliale e di tessuti sani attraverso lo studio della loro espressione genica.
In questo capitolo fornisco una breve introduzione al tumore ovarico epiteliale e alla tecnologia microarray, utilizzata per quantificare l'espressione genica. Infine presento i principali obiettivi di questa tesi.

1.1 Tumore ovarico

La quasi totalità dei tumori ovarici maligni origina da uno dei seguenti tre tipi di cellule: cellule epiteliali, cellule stromali e cellule germinali. Nei paesi sviluppati, più del 90% dei tumori ovarici maligni è di origine epiteliale e, anche per questo motivo, la maggior parte delle ricerche epidemiologiche si focalizza su questo sottotipo. Il tumore ovarico epiteliale si presenta in cinque principali istotipi: sieroso di alto grado (70%), endometrioide (10%), clear cell (10%), mucinoso (3%) e sieroso di basso grado ($< 5\%$) [25].
In diversi studi si è cercato di approfondire l'origine dei tumori ovarici epiteliali.

Per gli istotipi clear cell, mucinoso ed entrometriode c'è una chiara somiglianza, a livello istologico, rispettivamente con i tessuti epiteliali di rene, colon o stomaco, ed ovaio. L'istotipo sieroso è invece quello più difficile da classificare a livello istologico, ed è quello su cui sono concentrati la maggior parte degli studi. In passato si credeva che i tumori ovarici sierosi originassero dal tessuto dell'ovaio, mentre negli ultimi anni diversi studi propongono le tube di Falloppio come possibile origine [11].
Esistono diversi criteri per valutare la grandezza e il livello di diffusione dei tumori; una delle classificazioni più usate riguarda lo stadio. Il tumore ovarico viene così classificato: stadio I (tumore confinato alle ovaie), stadio II (il tumore coinvolge una o entrambe le ovaie con estensione pelvica o cancro primario peritoneale), stadio III (il tumore coinvolge una o entrambe le ovaie con diffusione al peritoneo fuori le pelvi e/o metastasi ai linfonodi retroperitoneali), stadio IV (Metastasi distanti, escluse le metastasi peritoneali).

1.2 Microarray

I microarray vengono utilizzati per misurare l'insieme di trascritti delle cellule. I trascritti, ovvero porzioni di mRNA, derivano dalla trascrizione di geni contenuti nel DNA che vengono poi tradotti in proteine, ovvero sequenze di aminoacidi. Le proteine svolgono le funzioni biologiche delle cellule e, sebbene subiscano spesso delle modifiche post-traduzionali, studiare i trascritti fornisce la maggior parte delle informazioni sulle proteine sintetizzate e quindi, indirettamente, sui processi che avvengono nelle cellule.
I microarray sono piccoli vetrini sui quali vengono fissati filamenti singoli di DNA, chiamati *probe*, in posizioni specifiche, chiamate *spot*. Ogni microarray può contenere decine di migliaia di spot, e in ogni spot ci sono milioni di copie della stessa sequenza di DNA. I microarray sono disegnati specie-specifici affinché ogni sequenza sia complementare ad un particolare trascritto proveniente da un gene fissato. I microarray possono essere a singolo oppure a doppio canale. In questa tesi tutti i dati utilizzati provengono da microarray a singolo canale. Nel caso di singolo canale le sequenze di mRNA di uno specifico campione da analizzare vengono ibridate sul vetrino, ovvero le sequenze complementari a quelle fissate sul

vetrino si appaiano. Ogni filamento di mRNA è stato in precedenza marcato con un fluoroforo che emette una luce quando sollecitato. Il vetrino viene quindi scansionato, producendo così un'immagine; maggiore è il numero di mRNA identici che si sono appaiati alla relativa sequenza, maggiore è l'intensità di luce prodotta nella scansione per quello specifico spot. L'immagine viene quindi analizzata per quantificare l'intensità di ogni spot. Viene inoltre misurata l'intensità del background di ciascuno spot, infatti, a causa dei fluorofori e della possibile presenza di materiale fuori dallo spot, si produce una luminosità, che è un rumore di fondo. Infine l'intensità di background viene sottratta all'intensità dello spot e si ottiene, per ogni campione analizzato, un vettore di intensità, che rappresenta il livello di espressione dei geni.
Per N campioni e p spot, corrispondenti a p geni, si ottiene una matrice di intensità di espressione del tipo:

$$
\begin{array}{c} \\ p \text{ geni} \end{array}
\begin{array}{c}
\overset{N \text{ campioni}}{} \\
\begin{bmatrix}
y_{11} & y_{12} & \cdots & y_{1i} & \cdots & y_{1N} \\
y_{21} & y_{22} & \cdots & y_{2i} & \cdots & y_{2N} \\
\vdots & \vdots & \ddots & \vdots & & \vdots \\
y_{k1} & y_{k2} & \cdots & y_{ki} & \cdots & y_{kN} \\
\vdots & \vdots & & \vdots & \ddots & \vdots \\
y_{p1} & y_{p2} & \cdots & y_{pi} & \cdots & y_{pN}
\end{bmatrix}
\end{array}
$$

Tale matrice è la base di partenza per tutte le analisi successive.

1.3 Scopo della tesi

In questa tesi affronto diverse questioni riguardanti il tumore ovarico epiteliale, cercando da una parte evidenze a favore di ipotesi formulate negli ultimi anni, dall'altra di cogliere ed approfondire ciò che i dati stessi suggeriscono.
Gli obiettivi di questa tesi possono essere riassunti in tre grandi quesiti, ciascuno affrontato in una sezione della tesi.
Primo quesito. Il primo argomento trattato è l'approfondimento della relazione tra tessuti sani e tessuti tumorali. L'analisi si è sviluppata attorno a due principali obiettivi: approfondire la somiglianza dei campioni sierosi con i tessuti di tube e

di ovaio, e capire se, conoscendo le caratteristiche a livello genico dei tessuti sani, è possibile prevedere la classificazione in istotipi dei campioni tumorali. Per affrontare il primo obiettivo sono partita da un articolo pubblicato su *Nature* [11], nel quale, tra le altre cose, hanno studiato la somiglianza di alcuni campioni di tumore sieroso con i tessuti di tube ed ovaio, che costituiscono possibili siti di origine del tumore. Nell'articolo è stato usato un approccio basato sulla correlazione calcolata su un sottoinsieme di geni. Nelle mie analisi ho seguito questa proposta e ho poi proseguito con dei modelli di classificazione che affrontano entrambi gli obiettivi sopra descritti. In tali modelli, infatti, ho utilizzato i tessuti sani per stimare il modello, con l'obiettivo di utilizzarlo per classificare i tessuti tumorali. Osservando la classificazione ottenuta ho cercato informazioni riguardo le somiglianze dei vari istotipi ai tessuti sani.

Secondo quesito. Il passo successivo è stato l'analisi dei soli campioni tumorali. Ho utilizzato diversi modelli di classificazione adattati agli istotipi degli stadi iniziali e agli stadi avanzati, individuando, per ogni tipologia, il modello ad errore minimo. Oltre a valutare la capacità predittiva di ogni modello nel classificare nuove unità, questo approccio mi ha consentito di individuare dei marcatori di ogni condizione (istotipi e stadi avanzati). Per ogni modello, ovvero, ho estratto i geni che hanno un maggior ruolo nel classificare un'unità in una specifica classe; questo ha portato ad avere, per ogni condizione, un insieme di geni che, per il modello utilizzato, sono i più rilevanti per distinguere quella classe dalle altre.

Terzo quesito. Analizzando i dati nelle fasi precedenti, ho osservato un'elevata eterogeneità tra campioni, anche appartenenti alla stessa condizione. Senza imporre a priori una divisione in classi, ho analizzato i dati con dei metodi non supervisionati, che utilizzano le informazioni sui valori di espressione dei geni, senza specificare una classe di appartenenza. Tali metodi raggruppano le unità, con diversi criteri, in base alla somiglianza dei valori di espressione. Ho quindi osservato i raggruppamenti emersi da tali analisi ed approfondito le caratteristiche dei soggetti appartenenti ai vari gruppi, per capire se tali raggruppamenti coincidono con una divisione data da alcune caratteristiche cliniche dei pazienti. Nei casi più rilevanti ho proseguito le analisi per individuare eventuali differenze tra i gruppi anche a livello genico o in termini di sopravvivenza alla malattia.

Nel capitolo 3 descrivo la preparazione dei dati e le prime analisi esplorative che

hanno fornito una prima idea sulle caratteristiche dei diversi campioni. Il primo quesito, qui descritto, è affrontato nel capitolo 4 di questa tesi, mentre i quesiti 2 e 3 sono affrontati nel capitolo 5, in quanto sono due approcci da un certo punto di vista opposti, ma complementari per l'analisi dei campioni tumorali.

Capitolo 2

Materiale e metodi utilizzati

2.1 Dati biologici

In questa tesi ho utilizzato dati di espressione di microarray di mRNA presi da tre diverse fonti. In tutti e tre i casi l'analisi di espressione genica è stata condotta tramite microarray Agilent a singolo canale.
Complessivamente ho utilizzato 17 campioni di tessuto sano, 83 campioni di tumore ovarico epiteliale di stadio I di diverse istologie, 88 campioni di tumore ovarico sieroso di stadi avanzati.
I campioni di tumore di stadio avanzato e 9 campioni di tessuto epiteliale sano di tube e ovaio provengono da biopsie ottenute presso la Divisione di Ostetricia e Ginecologia, ASST Spedali Civili, Università di Brescia, tra il 2003 e il 2013. Le biopsie sono conservata nella biobanca all'Istituto "A. Nocivelli", ASST Spedali Civili di Brescia. É stata utilizzata la piattaforma Agilent "Agilent-039494 SurePrint G3 Human GE 8x60K Microarray".
Gli 83 campioni di tumore allo stadio I provengono dallo studio [6] pubblicato in *Annals of Oncology*. In questo caso è stata utilizzata la piattaforma "Agilent-028004 SurePrint G3 Human GE 8x60K Microarray" per l'analisi di espressione genica.
Infine alcuni dati sono tratti dallo studio [39], su *Cancer Research*. I dati sono disponibili presso il sito dell'NCBI, nel database GEO, attraverso il codice di accesso "GSE77199". Sono stati ottenuti attraverso la piattaforma Agilent "Agilent-028004 SurePrint G3 Human GE 8x60K Microarray". Per il mio lavoro

ho utilizzato i 4 campioni disponibili di tessuto endoteliale sano del colon e i 4 campioni di tessuto endoteliale sano del rene.

2.2 Metodi statistici

In questa sezione vengono descritti i principali strumenti statistici utilizzati in questo lavoro. La normalizzazione è un'operazione finalizzata a rimuovere parte della variabilità sistematica dei dati dovuta a variazioni sperimentali e/o tecniche e non a cause biologiche. L'analisi dei geni differenzialmente espressi consiste nell'individuare, attraverso test statistici, geni che presentano livelli di espressione significativamente diversi tra due o più condizioni. Per analizzare i dati a disposizione ho impiegato sia tecniche di apprendimento non supervisionato, che supervisionato. Per quanto riguarda l'apprendimento non supervisionato ho utilizzato il cluster gerarchico, il cluster non gerarchico, modelli mistura di gaussiane e infine lo scaling multidimensionale; per l'apprendimento supervisionato sono stati invece utilizzati i seguenti modelli di classificazione: foreste casuali, support vector machines, "nearest shrunken centroids" e metodi di analisi del discriminante. L'analisi di arrichimento e il metodo "Source Set" sono strumenti utili ad interpretare i risultati ottenuti dai metodi precedenti. A partire da un elenco di geni differenzialmente espressi tra due condizioni, l'analisi di arricchimento ha l'obiettivo di individuare categorie biologiche o pathway biologici che siano significativamente associati ai geni. Il metodo Source Set, invece, è pensato per individuare quali geni sono la causa primaria di una disregolazione all'interno di un pathway biologico.

2.2.1 Normalizzazione

La normalizzazione tra array è una fase utile a rendere i diversi campioni tra loro confrontabili. Ci sono infatti diverse cause di variabilità sistematica tra array (per esempio diversa efficacia nella fase di trascrizione o di ibridazione, problemi con i campioni di partenza, diverse condizioni di laboratorio) e, con la normalizzazione, si cerca di ridurla il più possibile per poter cogliere, con le analisi successive, la variabilità biologica di interesse. Per i dati utilizzati in questa tesi ho scelto una normalizzazione Loess ciclica (versione *fast*). Essa, come la maggior parte delle normalizzazioni più usate, si basa su 2 assunzioni [2]:

- Presenza di pochi geni differenzialmente espressi tra le due condizioni, rispetto al totale di geni;
- Simmetria tra geni sovra-espressi e sotto-espressi: i geni differenzialmente espressi devono essere circa ugualmente ripartiti.

Considerati n campioni e p geni, sia y_{ki} il valore di espressione per il gene k e il campione i. Un utile strumento grafico per visualizzare la distribuzione delle intensità di espressione tra campioni è il grafico MA. Dati i valori di espressione per il gene k, per gli esperimenti i e j, il grafico MA rappresenta sulle ascisse la media delle log-intensità, e sulle ordinate la differenza delle log-intensità, che corrispondono alle trasformazioni

$$\begin{cases} A_k = \log_2(y_{ki}/y_{kj}) = \log_2(y_{ki}) - \log_2(y_{kj}) \\ M_k = \frac{1}{2}\log_2(y_{ki} \cdot y_{kj}) = \frac{1}{2}(\log_2(y_{ki}) + \log_2(y_{kj})) \end{cases}$$

Il grafico MA per una normalizzazione dei dati ideale dovrebbe mostrare una nuvola di punti sparsa attorno all'asse $M = 0$ (un esempio in Figura 2.1).

La normalizzazione Loess ciclica *fast* agisce sulle coordinate del grafico MA: per ogni esperimento i viene creato un grafico MA modificato, nel quale il campione i viene confrontato con un campione "medio". La correzione per il campione i-esimo consiste nella traslazione della coordinata M_k di una quantità pari alla curva loess adattata ai punti del grafico; ovvero, dopo la normalizzazione, la curva loess coincide con l'asse $M = 0$.

Siano $\tilde{y}_{ki}$ i valori di espressione trasformati in scala logaritmica (solitamente in base e o in base 2). I passaggi dell'algoritmo, come descritto in [2], sono i seguenti:

1. per ogni gene k, creare il vettore $\hat{y}_{k\cdot} = \frac{1}{n}\sum_{i=1}^{n} \tilde{y}_{ki}$, media tra gli esperimenti della log-intensità del gene k;
2. ciclo per $i \in 1, ..., n$
   (a) disegnare il grafico MA con $M_k = \hat{y}_{k\cdot}$ e $A_k = \tilde{y}_{ki} - \hat{y}_{k\cdot}$;
   (b) adattare una curva loess $f(x)$ ai dati;
   (c) sottrarre $f(x)$ ai valori M_k.

La sequenza di passaggi è da ripetere fino a convergenza, ovvero finché il vettore medio $\hat{y}_{k\cdot}$ rimane costante tra una sequenza di passaggi e la successiva. Generalmente è sufficiente un iterazione, o al più due.

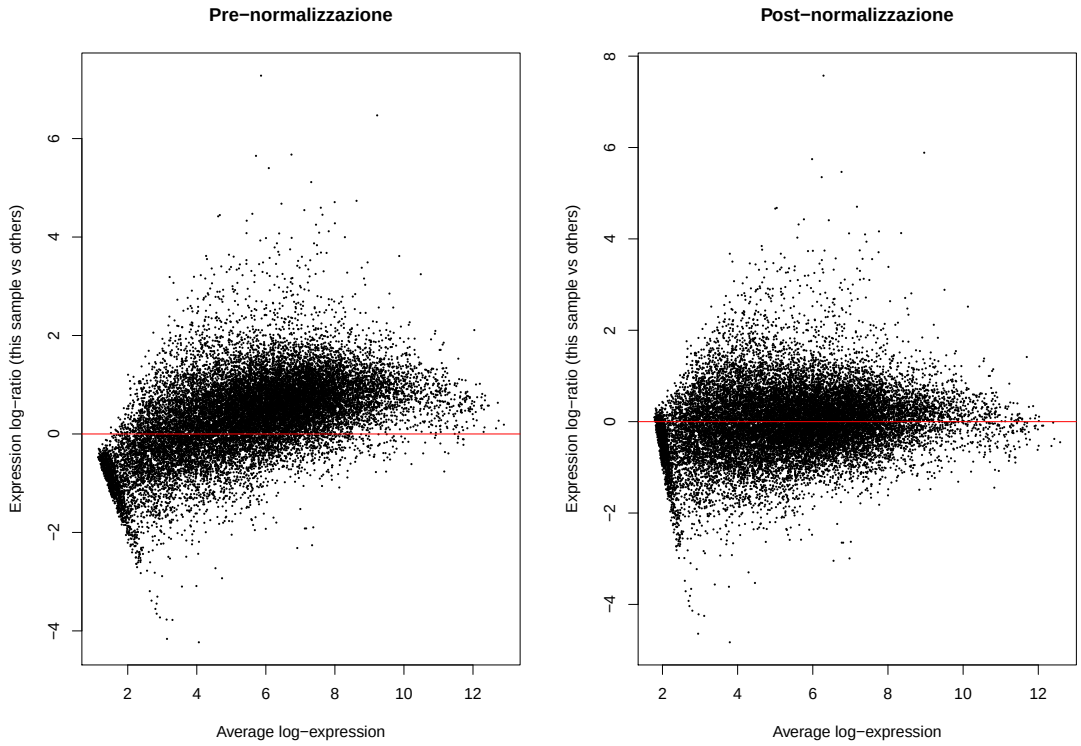


Figura 2.1: Esempio di grafico MA prima e dopo la normalizzazione

2.2.2 Geni differenzialmente espressi

In questa tesi ho utilizzato l'informazione sui geni differenzialmente espressi soprattutto come strumento per selezionare un sottoinsieme di geni su cui stimare i modelli. Spesso, infatti, i modelli adattati sulla totalità di geni a disposizione non hanno una buona capacità di previsione, perché solitamente solo una piccola parte dei geni è effettivamente utile a classificare le osservazioni, mentre i rimanenti geni apportano solo rumore.
Per identificare i geni differenzialmente espressi esistono test statistici specifici, che tengono conto della particolare struttura dei dati, ovvero della grande numerosità dei geni e del basso numero di campioni. Ho utilizzato tre diversi test per individuare i geni differenzialmente espressi, ovvero il test *Empirical Bayes*, il test SAM (Significance Analysis of Microarrays) ed infine il test *Product Rank*.
È inoltre necessario tener conto della presenza di confronti multipli nel valutare la significatività dei test. In questo contesto, infatti, vengono svolti tanti test quanti

i geni a disposizione, e non considerare questo aspetto comporta un considerevole aumento della proporzione di falsi positivi. In questa tesi ho utilizzato la procedura di controllo FDR (False Discovery Rate), proposta da Benjamini e Hochberg [3].
Di seguito la definizione del FDR ed una breve descrizione dei test, focalizzata sui microarray a singolo canale. Da qui in avanti, per tutto il capitolo, quando ci si riferisce ad una matrice di espressione è da intendere già normalizzata.

False Discovery Rate Questo metodo si basa sul controllo della proporzione attesa di falsi positivi. Si consideri il problema di testare simultaneamente m ipotesi, delle quali m_0 sono vere. Sia R il numero di ipotesi rifiutate. In tabella 2.1 gli errori commessi nel testare le ipotesi nulle. Sia $\mathbf{Q} = \mathbf{V}/(\mathbf{V}+\mathbf{S})$ la proporzione

	Dichiarati negativi	Dichiarati positivi	Totale
Veri negativi	**U**	**V**	m_0
Veri positivi	**T**	**S**	$m - m_0$
	$m - \mathbf{R}$	**R**	m

Tabella 2.1: Errori commessi nel testare le ipotesi nulle

di ipotesi nulle erroneamente rifiutate, con $\mathbf{Q} = 0$ quando $\mathbf{V} + \mathbf{S} = 0$. $\mathbf{Q}$ è una variazione casuale non osservata. Si definisce il False Discovery Rate Q_e come il valore atteso di $\mathbf{Q}$ [3]:

$$Q_e = E[\mathbf{Q}] = E[\mathbf{V}/(\mathbf{V}+\mathbf{S})] = E[\mathbf{V}/\mathbf{R}].$$

Empirical Bayes Il metodo *Empirical Bayes*, ovvero bayesiano empirico, viene descritto in [33]. Di seguito una descrizione del metodo tratta da tale articolo.
Si consideri una matrice di espressione con p geni ed n campioni, divisi in condizione A e condizione B. Siano $\mathbf{y}_k^T = (y_{k1}, ..., y_{kN})$, $k = 1, ..., p$, i vettori contenenti i valori di espressione in scala logaritmica per ogni gene. Siano X la matrice del disegno contenente l'informazione sulle due classi A e B, ed $\boldsymbol{\alpha}_k$ un vettore di coefficienti. Si assume valgano le relazioni:

$$\begin{aligned} E(\mathbf{y}_k) &= X\boldsymbol{\alpha}_k \\ var(\mathbf{y}_k) &= W_k\sigma_k^2 \end{aligned}$$

dove W_k è una matrice di pesi nota, non negativa. Siano $\boldsymbol{\beta}_k = C^T\boldsymbol{\alpha}_k$ coefficienti di interesse biologico, per i quali si vuole testare l'ipotesi $\beta_{ki} = 0$, $i = 1, ..., N$.
Si assuma di aver stimato il modello lineare descritto per ogni gene e di aver ottenuto gli stimatori $\hat{\boldsymbol{\alpha}}_k$ e s_k^2 rispettivamente di $\boldsymbol{\alpha}_k$ e σ_k^2, e le matrici stimate di covarianza

$$var(\boldsymbol{\alpha}_k) = V_k s_k^2$$

dove V_k è una matrice definita positiva non dipendente da s_k^2. Gli stimatori dei contrasti sono $\hat{\boldsymbol{\beta}}_k = C^T\hat{\boldsymbol{\alpha}}_k$, con matrice di covarianza stimata

$$var(\hat{\boldsymbol{\beta}}_k) = C^T V_k C s_k^2.$$

Sia v_{ki} l'i-esimo elemento diagonale di $C^T V_k C$. Si assumono le seguenti distribuzioni approssimate:

$$\begin{aligned} \hat{\beta}_{ki} \mid \beta_{ki}, \sigma_k^2 &\sim \mathcal{N}(\beta_{ki}, v_{ki}\sigma_k^2) \\ s_k^2 \mid \sigma_k^2 &\sim \tfrac{\sigma_k^2}{d_k}\chi_{d_k}^2 \end{aligned}$$

dove d_k sono i gradi di libertà per il modello lineare riferito al gene k. Sotto queste ipotesi la statistica t ordinaria

$$t_{ki} = \frac{\hat{\beta}_{ki}}{s_k\sqrt{v_{ki}}}$$

segue una distribuzione approssimata t con d_k gradi di libertà.
Si introduce ora la componente gerarchica per la quale si assume che $\hat{\boldsymbol{\beta}}_k$ e s_k^2 di geni differenti siano indipendenti. L'obiettivo è descrivere come i coefficienti β_{ki} e le varianze σ_k^2 ignoti variano tra i geni. Si suppongano al momento note le quantità d_0 ed s_0^2. Per le varianze si ipotizzano le distribuzioni a priori

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

Per ogni campione i si supponga nota la vera proporzione di geni differenzialmente espressi $p_i = P(\beta_{ki} \neq 0)$. Per i geni differenzialmente espressi si suppone la distribuzione a priori dei coefficienti:

$$\beta_{ki} \mid \sigma_k^2, \beta_{ki} \neq 0 \sim \mathcal{N}(0, v_{0i}\sigma_k^2).$$

Sotto il modello gerarchico specificato, la media a posteriori di σ_k^2 dato s_k^2 è $\tilde{s}_k^{-2}$ con

$$\tilde{s}_k^2 = \frac{d_0 s_0^2 + d_k s_k^2}{d_0 + d_k}.$$

Si definisce dunque la statistica t moderata:

$$\tilde{t}_{ki} = \frac{\hat{\beta}_{ki}}{\tilde{s}_k \sqrt{v_{ki}}}.$$

Questa statistica rappresenta un ibrido tra l'approccio classico e quello bayesiano, infatti nella classica statistica t la varianza campionaria è stata sostituita dalla varianza a posteriori bayesiana. Si dimostra che, sotto l'ipotesi nulla $H_0 : \beta_{ki} = 0$, la statistica $\tilde{t}_{ki}$ segue una distribuzione t con $d_k + d_0$ gradi di libertà.
Ciò che rende questo approccio bayesiano *empirico* è il fatto che in realtà a d_0 e s_0^2 non viene attribuito un valore a priori, ma vengono stimati dai dati. La procedura usata per la loro stima è descritta in [33]. Per quanto riguarda la probabilità p_i, a livello pratico si preferisce utilizzare un valore piccolo a piacere, per esempio $p_i = 0.01$, in quanto la stima effettuata tramite i dati può portare a risultati instabili.

SAM Si consideri una matrice di espressione con p geni ed N campioni, di cui N_A appartenenti alla condizione A, e N_B appartenenti alla condizione B; siano $\bar{y}_{kA}$ e $\bar{y}_{kB}$ le medie di espressione per il gene k nelle condizioni A e B e sia s_k la deviazione standard "pooled" per il gene k:

$$s_k = \sqrt{\left(\frac{1}{N_A} + \frac{1}{N_B}\right) \frac{\sum_{i=1} N_A (y_{ki} - \bar{y}_{kA})^2 + \sum_{i=1} N_B (y_{ki} - \bar{y}_{kB})^2}{N_A + N_B - 2}}.$$

La statistica test standard per due campioni non appaiati ed indipendenti è la statistica t di student:

$$t_k = \frac{\bar{y}_{kB} - \bar{y}_{kA}}{s_k}.$$

Per geni poco espressi, può capitare che anche la varianza abbia valori bassi, ai quali corrispondono alti valori della statistica test t_k e quindi il rifiuto dell'ipotesi nulla. Per evitare che geni con bassi valori di espressione possano dominare l'analisi, nel

metodo SAM viene proposta la statistica t moderata:

$$d_k = \frac{\bar{y}_{kB} - \bar{y}_{kA}}{s_k + s_0}$$

dove s_0 è una piccola costante positiva scelta in modo da minimizzare il coefficiente di variazione di d_k, calcolato in funzione di s_k, per finestre mobili dei dati. In [9] viene descritto l'algoritmo completo. Per calcolare la distribuzione nulla viene utilizzato un approccio permutazionale. La procedura SAM consiste nei seguenti passaggi [37, 22]:

1. calcolare le statistiche ordinate $d_{(1)} \leq d_{(2)} \leq ... \leq d_{(p)}$;
2. considerare B permutazioni delle etichette dei gruppi. Per ogni permutazione b calcolare le statistiche d_k^b e le relative statistiche ordinate $d_{(1)}^b \leq d_{(2)}^b \leq ... \leq d_{(p)}^b$. Calcolare la statistica ordinata attesa $\bar{d}_{(k)} = (1/B)\sum_{b=1}^{B} d_{(k)}^b$, per $k = 1, 2, ..., p$;
3. si consideri il grafico dei valori osservati $d_{(k)}$ contro i valori attesi $\bar{d}_{(k)}$. Per una fissata soglia $\Delta \geq 0$, partendo dall'origine e muovendosi verso destra, si cerca il primo valore k_2 tale che $d_{(k_2)} - \bar{d}_{(k_2)} \geq \Delta$; tutti i geni oltre k_2 sono considerati significativamente sovra-espressi nella condizione A rispetto alla B. Similmente, partendo dall'origine e muovendosi verso sinistra, si cerca il primo valore k_1 tale che $\bar{d}_{(k_1)} - d_{(k_1)} \geq \Delta$; tutti i geni oltre k_1 sono considerati significativamente sotto-espressi;
4. per ogni Δ, sia $t_2(\Delta)$ il più piccolo d_k tra i geni significativamente positivi, e $t_1(\Delta)$ il più grande d_k tra i geni significativamente negativi. Se $t_1(\Delta) > t_2(\Delta)$, allora si pone $t_2(\Delta) = t_1(\Delta) = 0$.

Per ogni fissato Δ viene quindi stimato il valore di FDR.

Product Rank Siano $y_1, ..., y_{N_A}$ vettori contenenti i valori di espressione per la prima condizione e $y_1, ..., y_{N_B}$ per la seconda. Il metodo procede nel seguente modo [5]:

1. per ogni confronto tra un esperimento della condizione A ed uno della condizione B calcolare i rapporti dei livelli di espressione, per un totale di $K = N_A \cdot N_B$ confronti: $y_1/y_1, y_1/y_2, ..., y_{N_A}/y_1, ..., y_{N_A}/y_{N_B}$;

2. all'interno di ogni confronto i, ordinare in modo decrescente i rapporti e considerare, per ogni gene k, il rango: r^A_{ki}, $i \in \{1, ..., K\}$, $k \in \{1, ..., p\}$, dove $r^A_{ki} = 1$ corrisponde al gene maggiormente up-regolato nella condizione A rispetto alla condizione B;
3. determinare il prodotto dei ranghi per ogni gene: $RP^A_k = (\prod_i r^A_{ki})^{1/K}$;
4. per il calcolo della significatività: sia L il numero di permutazioni (per esempio $L = 100$).
   Ciclo per $l \in \{1, ..., L\}$: all'interno di ogni esperimento, permutare indipendemente i valori di espressione e ripetere i passaggi 1-3 per ottenere $RP^{A(l)}_k$;
5. sia x^A_k il numero di volte in cui $RP^{A(l)}_k$ è maggiore o uguale di RP^A_k; calcolare il valore atteso medio $E(RP^A_k) \approx x^A_k/L$;
6. calcolare la percentuale di falsi positivi (PFP) nel caso in cui il gene k (e tutti quelli con RP maggiore) è considerato significativamente differenzialmente espresso: $q^A_k = E(RP^A_k)/\operatorname{rank}(k)$, dove $\operatorname{rank}(k)$ è la posizione del gene g nella lista di geni ordinati per valore decrescente di RP, ovvero il numero di geni considerati differenzialmente espressi.

Ripetere similmente i passaggi 2-6, ordinando in modo decrescente i rapporti, per ottenere quantità analoghe, riferite ad una maggior espressione dei geni nella condizione B rispetto alla A.

2.2.3 Apprendimento non supervisionato

I metodi di apprendimento non supervisionato hanno l'obiettivo di inferire proprietà sulla distribuzione dei dati senza avere informazioni sulla correttezza o meno della risposta o sul grado di errore commesso [14]. Vengono presentati i metodi di cluster gerarchico, scaling multidimensionale, cluster non gerarchico, e modelli mistura di gaussiane. I primi due metodi sono stati utilizzati principalmente per le analisi esplorative, mentre gli ultimi due per cercare di evidenziare dei gruppi di pazienti in base ai livelli di espressione dei geni, con l'obiettivo di confrontare questi risultati con le annotazioni cliniche disponibili.

Cluster gerarchico

Il cluster gerarchico si basa su una misura di dissimilarità tra gruppi di osservazioni, basata sulla dissimilarità tra tutte le possibili coppie di osservazioni. Il metodo fornisce una rappresentazione gerarchica, nella quale il livello più basso è costituito da cluster contenenti ciascuno una sola osservazione, mentre il livello più alto è costituito da un unico gruppo contenente tutte le osservazioni. Partendo dal basso, ad ogni iterazione dell'algoritmo, due gruppi del livello precedente vengono uniti in un unico gruppo; la scelta dei due gruppi da unire si basa sulla minor dissimilarità. Ciò che caratterizza il metodo è la scelta della misura di dissimilarità e la definizione di dissimilarità tra gruppi. [14]
Siano G ed H due gruppi e d_{ij} la dissimilarità tra i punti i e j. I 3 modi più comuni di definire la dissimilarità tra i gruppi G ed H sono i seguenti:

- *Single Linkage*: la minor distanza tra tutte le coppie di punti, uno appartenente a G ed uno ad H

$$d(G,H) = \min_{i \in G, j \in H} d_{ij}.$$

- *Complete Linkage*: la maggior distanza tra tutte le coppie di punti, uno appartenente a G ed uno ad H

$$d(G,H) = \max_{i \in G, j \in H} d_{ij}.$$

- *Average Linkage*: la distanza media tra tutte le coppie di punti, uno appartenente a G ed uno ad H

$$d(G,H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{j \in H} d_{ij}$$

  dove N_G ed N_H sono le numerosità dei gruppi.

Scaling multidimensionale

L'idea dello scaling multidimensionale è quella di mappare i vettori dello spazio p-dimensionale $y_1, ..., y_N$ in uno spazio di dimensione più piccola, cercando di preservare il più possibile le distanze tra coppie di vettori. Solitamente come misura

di dissimilarità si usa la distanza Euclidea $d_{ij} = \|y_i - y_j\|$, tuttavia può essere usata qualunque altra distanza [14]. I valori $z_1, ..., z_N$ nello spazio di arrivo q-dimensionale, $q < p$, sono scelti in modo da minimizzare la funzione

$$S_M(z_1, ..., z_N) = \sum_{i \neq j} (d_{ij} - \|z_i - z_j\|)^2.$$

Cluster non gerachico

L'algoritmo K-means è un metodo di clusterizzazione non gerarchica. Dato un prefissato numero di gruppi G, il metodo assegna ogni osservazione ad uno specifico gruppo, in modo da minimizzare una qualche funzione di perdita basata sulla dissimilarità tra le osservazioni. Siano $y_1, ..., y_N$ i vettori di osservazioni per N campioni, dove $y_i = (y_{1i}, ..., y_{pi})^T$ sono i valori di espressione di p variabili (geni). Sia C un'assegnazione di cluster, $g = C(i)$, che assegna l'i-esima unità al gruppo g-esimo. Il metodo K-means utilizza come misura di dissimilarità la distanza euclidea al quadrato $d(y_i, y_j) = \|y_i - y_j\|^2$.
Si definisce la funzione di perdita

$$W(C) = \sum_{g=1}^{G} N_k \sum_{C(i)=g} \|y_i - \bar{y}_g\|^2$$

che viene minimizzata attraverso un algoritmo di ottimizzazione in due passi [14]:

1. data l'assegnazione dei gruppi C, si calcolano le medie dei gruppi correnti: $m_g = \text{argmin}_m \sum_{C(i)=g} \|y_i - m\|^2$, $g = 1, ..., G$;
2. dato l'insieme delle medie correnti $\{m_1, ..., m_G\}$, si assegna ogni osservazione alla media del gruppo più vicina: $C(i) = \text{argmin}_{1 \leq g \leq G} \|y_i - m_g\|^2$.

I passi 1 e 2 vengono ripetuti fino a convergenza.

Modelli mistura di gaussiane

I modelli mistura di gaussiane sono presentati in [12]. Di seguito una breve descrizione del metodo proposto.
La clusterizzazione tramite modelli mistura di gaussiane suppone che le osservazioni $y_1, ..., y_N$ siano campioni casuali indipendenti di una mistura di distribuzioni

normali, e ogni componente della mistura rappresenta un cluster. Ovvero ogni campione y_i ha densità di probabilità

$$f(y_i, \boldsymbol{\Psi}) = \sum_{m=1}^{M} \pi_m f_m(y_i; \boldsymbol{\vartheta}_m)$$

dove M è il numero di componenti della mistura e $\boldsymbol{\Psi} = \{\pi_1, ..., \pi_{M-1}, \boldsymbol{\vartheta}_1, ..., \boldsymbol{\vartheta}_M\}$ sono i parametri del modello di mistura. Ogni componente del modello di mistura ha distribuzione gaussiana multivariata, $f_m(\boldsymbol{y}; \boldsymbol{\vartheta}_m) \sim \mathcal{N}(\boldsymbol{\mu}_m, \Sigma_m)$, con $\boldsymbol{y} = \{y_1, ..., y_N\}$, pertanto i cluster sono ellissoidali, centrati nel vettore delle medie μ_m, e con volume, forma ed orientamento determinati dalla matrice di covarianza Σ_m. Le scelte del numero M di componenti della mistura e della struttura della matrice di covarianza possono essere, per esempio, effettuate utilizzando il criterio di informazione BIC. Siano $\mathcal{M}_1, ..., \mathcal{M}_R$ i modelli presi in condiserazione, si ha allora

$$BIC_{\mathcal{M}_r} = 2\ell_{\mathcal{M}_r}(\boldsymbol{y} \mid \hat{\boldsymbol{\Psi}}) - \nu_r \log(N)$$

dove $\ell_{\mathcal{M}_r}$ è la log-verosimiglianza del modello $\mathcal{M}_r$ e ν_r il numero di parametri indipendenti da stimare.
Per ottenere la stima di massima verosimiglianza viene utilizzato l'algoritmo EM (Expectation-Maximization), che alterna un passo di stima, nel quale viene calcolato il valore atteso della log-verosimiglianza condizionatamente ai valori osservati e alla stima corrente dei parametri, e un passo di massimizzazione nel quale vengono determinati i parametri che massimizzano la log-verosimiglianza dell'iterazione corrente.

2.2.4 Apprendimento supervisionato

I metodi di apprendimento supervisionato, a differenza dei precedenti, si basano sulla conoscenza di un insieme di dati che comprende, per ogni campione, sia le informazioni sui predittori (geni) che sulla risposta (classe di appartenenza). É necessario quindi definire fin da subito rispetto a quale caratteristica si vogliono classificare i campioni. Inoltre, parte dei dati deve essere usata per la stima dei modelli, parte per verificare la bontà del modello nel prevedere la classificazione su nuovi dati. Di seguito vengono descritti i modelli utilizzati in questa tesi.

Foreste casuali

Le foreste casuali, o *random forest*, sono un metodo di combinazione di classificatori che utilizza alberi di classificazione come classificatori di base. Il vantaggio di usare il risultato di più alberi di classificazione deriva dal fatto che la stima di un albero dipende molto dai dati utilizzati, e la variazione anche di una piccola percentuale di dati può portare a risultati molto diversi, sebbene con un errore di previsione simile. Combinare quindi le previsioni di più modelli spesso aiuta a migliorare la classificazione [1].
Siano $y_1, ..., y_N$ le osservazioni per N campioni, divisi in due classi, in uno spazio p-dimensionale. Il metodo prevede l'utilizzo di B alberi di classificazione, che insieme costituiscono la "foresta". Per ogni albero, ad ogni nodo, viene selezionato casualmente un piccolo gruppo di variabili, sulle quali viene calcolato il miglior punto di suddivisione per definire il nodo. Per ogni nodo vengono quindi utilizzate solo $q \ll p$ variabili, ed ogni albero viene fatto crescere fino alla massima grandezza.
Solitamente, quando si costruisce una foresta casuale, si associa una procedura "bagging", ovvero ogni albero viene fatto crescere su un diverso campione bootstrap delle osservazioni di partenza, ciascuno della dimensione di partenza. Questa procedura consente di calcolare l'errore di previsione del metodo sulle osservazioni di volta in volta non utilizzate per stimare ogni albero, invece di dover disporre di un insieme di verifica [1].
La previsione finale viene calcolata come voto di maggioranza tra le previsioni fornite da ogni albero, assegnando l'osservazione alla classe con più voti.

Support vector machines

L'idea di base delle *support vector machines* per la classificazione è quella di trovare un iperpiano che riesca a separare al meglio punti appartenenti a due classi distinte. Siano $y_1, ..., y_N$ le osservazioni per N campioni, in uno spazio p-dimensionale (p variabili), divise in due classi indicate con $z = 1$ e $z = -1$. Il metodo è reso più flessibile considerando una trasformazione delle variabili di partenza in uno spazio q-dimensionale $h(y) = (h_1(y), ..., h_q(y))^T, y \in \mathbb{R}^p$, dove q può essere minore, uguale o maggiore di p [1].
La specificazione della trasformazione $h(y)$ può avvenire attraverso una *funzio-*

ne kernel, che calcola il prodotto interno nello spazio delle variabili trasformate: $K(y, y') = \langle h(y), h(y') \rangle$.
L'obiettivo è quindi cercare una curva del tipo $f(y) = \beta_0 + h(y)^T\beta = 0$, che separi al meglio le due classi, ovvero che dia luogo al maggior margine tra i punti appartenenti alla classe 1 e quelli appartenenti alla classe -1. Sia M la distanza, su ciascun semispazio, tra la curva $f(y)$ e il primo punto della rispettiva classe. Si ammette, in realtà, che alcuni punti possano trovarsi sul lato sbagliato del margine per la loro classe; per consentire ciò, si definiscono le variabili ausiliarie non negative $\xi_1, ..., \xi_N$, che rappresentano di quanto i punti si trovano sul lato sbagliato del semispazio definito dal margine, rispetto alla loro classe. Senza perdere di generalità si può porre $\|\beta\| = 1/M$. Il problema di ottimizzazione è così definito [14]:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \qquad \text{soggetto a:} \quad \begin{cases} z_i(h(y_i)^T\beta + \beta_0) \geq 1 - \xi_i, & i = 1, ..., N \\ \xi_i \geq 0 \end{cases}$$

dove il parametro C rapprenta il "costo" di violare il margine.
Si dimostra che la soluzione per i coefficienti β è della forma

$$\hat{\beta} = \sum_{i=1}^{N} \hat{a}_i z_i h(y_i)$$

dove solo alcuni coefficienti $\hat{a}_i$ sono non nulli. Le osservazioni corrispondenti sono chiamate "vettori di supporto", in quanto i coefficienti $\hat{\beta}$ sono rappresentati solo in funzione di essi. $\hat{\beta}_0$ viene stimato dalle osservazioni che si trovano esattamente sui bordi del margine ($\hat{\xi}_i = 0$) [14]. La curva stimata è dunque

$$\hat{f}(y) = h(y)^T\hat{\beta} + \hat{\beta}_0 = \sum_{i=1}^{N} \hat{a}_i z_i \langle h(y), h(y_i) \rangle$$

da cui si osserva che non è necessario specificare direttamente la trasformazione $h(y)$, ma è sufficiente conoscere la funzione kernel $K(y, y') = \langle h(y), h(y') \rangle$.
Le funzioni kernel più usate sono le seguenti:

- *lineare*: $\langle y, y' \rangle$
- *polinomiale*: $(\langle y, y' \rangle + r)^d$

- *radiale*: $\exp(-\gamma\|y - y'\|^2)$, $\gamma > 0$
- *sigmoide*: $\tanh(\gamma\langle y, y'\rangle + r)$.

In tutti i modelli che ho adattato in questa tesi ho utilizzato $r = 0$.
La regola di classificazione è data da

$$\hat{G}(y) = sgn\left[\hat{f}(y)\right].$$

Esistono diverse possibilità per generalizzare il metodo al caso di più classi. In questa tesi ho usato la strategia "one-vs-one", che prevede di stimare $G(G-1)/2$ modelli, dove G è il numero di classi, confrontando due classi alla volta, in tutte le combinazioni possibili. L'osservazione viene infine assegnata alla classe che ha ricevuto il maggior numero di previsioni [38].

Predictive Analysis of Microarray (PAM)

Il metodo PAM viene proposto in [36]. Si basa sulla contrazione dei centroidi di ogni classe, ed infatti è noto anche come "Nearest Shrunken Centroids".
Siano y_{ki} i valori di espressione per il gene k ed il campione i, $k = 1, ..., p$, $i = 1, ..., N$; siano C_g gli indici degli N_g campioni nella classe g, per $g = 1, ..., G$. La k-esima componente del centroide per la classe g è data dalla media di espressione nella classe g per il gene k: $\bar{y}_{kg} = \sum_{i \in C_g} y_{ki}/N_g$; la k-esima componente del centroide complessivo è $\bar{y}_k = \sum_{i=1}^{N} y_{ki}/N$.
Il metodo consiste nel comprimere i centroidi delle classi verso il centroide complessivo, dopo aver standardizzato, per ogni gene, per la deviazione standard calcolata all'interno di ogni classe. Sia

$$d_{kg} = \frac{\bar{y}_{kg} - \bar{y}_k}{m_g(s_k + s_0)} \tag{2.1}$$

una statistica t per il gene k-esimo, dove s_i è la deviazione standard per il gene k

$$s_k^2 = \frac{1}{N - G} \sum_g \sum_{j \in C_g} (y_k j - \bar{y}_k g)^2$$

e $m_g = \sqrt{1/N_g + 1/N}$. Il valore s_0 è una costante positiva con funzione simile alla costante s_0 nel test SAM: limitare la possibilità di valori alti in d_{kg} a causa di gene

con bassi livelli di espressione. s_0 è posto uguale al valore medio di s_k tra tutti i geni.
Riscrivendo la 2.1 si ottiene

$$\bar{y}_{kg} = \bar{y}_k + m_g(s_k + s_0)d_{kg}.$$

La quantità d_{kg} viene contratta verso lo zero nel seguente modo :

$$d'_{kg} = sgn(d_{kg})(|d_{kg}|-\Delta)_+$$

dove + indica la parte positiva. La quantità di compressione Δ viene scelta tramite convalida incrociata. Si ottengono quindi i centroidi compressi

$$\bar{y}'_{kg} = \bar{y}_k + m_g(s_i + s_0)d'_{kg}.$$

Un nuovo campione $y^* = (y_1^*, ..., y_p^*)$ viene classificato nel centroide più vicino, sempre standardizzando per $s_k + s_0$, ed inoltre correggendo per il numero relativo di campioni in ogni classe. Viene definito lo score per la classe g

$$\delta_g(y^*) = \sum_{k=1}^{p} \frac{(y_k^* - \bar{y}_{kg})^2}{(s_k + s_0)^2} - 2\log(\pi_g)$$

dove i π_g rappresentano la probabilità a priori di appartenere alla classe g. La regola di classificazione è dunque:

$$C(y^*) = g^* \quad \text{dove} \quad \delta_{g^*} = \min_g \delta_g(y^*).$$

Analisi discriminante lineare

I metodi di analisi del discriminante sono metodi ideati specificatamente per la classificazione e si basano sulle probabilità condizionate date dal teorema di Bayes. Sia Y una variabile casuale p-dimensionale continua e Z una variabile categoriale che rappresenta la classe di appartenza dei soggetti. Sia G il numero totale di classi, aventi densità di probabilità $p_0(y), ..., p_{G-1}(y)$ per la distribuzione condizionata di Y, e pesi $\pi_0, ..., \pi_{G-1}$, con somma ad 1. La densità marginale dell'intera

popolazione risulta essere allora

$$p(y) = \sum_{g=0}^{G-1} \pi_g p_g(y).$$

Sia y_0 il valore osservato per Y e π_g rappresenta la probabilità a priori che un soggetto appartenga alla classe g; la probabilità a posteriori che il soggetto appartenga alla classe g è

$$\mathbb{P}\{Z = g \mid Y = y_0\} = \frac{\pi_g p_g(y_0)}{p(y_0)}.$$

Il confronto tra due classi può avvenire attraverso il log-rapporto

$$\log \frac{\mathbb{P}\{Z = g \mid Y = y_0\}}{\mathbb{P}\{Z = h \mid Y = y_0\}} = \log \frac{\pi_g}{\pi_h} + \log \frac{p_g(y_0)}{p_h(y_0)}$$

o, equivalentemente, attraverso la *funzione discriminante* [1]

$$d_g(y_0) = \log \pi_g + \log p_g(y_0).$$

Si suppone ora che ogni densità $p_g(y)$ segua una distribuzione normale multivariata $\mathcal{N}_p(\mu_g, \Sigma)$; la funzione discriminante diventa allora

$$d_g(y_0) = \log \pi_g - \frac{1}{2}\mu_g^T \Sigma^{-1} \mu_g + y^T \Sigma^{-1} \mu_g.$$

Le probabilità a priori e i vettori delle medie vengono stimati con $\hat{\pi}_g = N_g/N$ e $\hat{\mu}_g = 1/N_g \sum_{i:z_i=g} y_i$, dove N_g sono le numerosità dei diversi gruppi [1].
La stima di Σ caratterizza il metodo. L'approccio classico, nel quale si utilizza la matrice di covarianza campionaria, non è applicabile per i dati utilizzati in questa tesi, per i quali $p \gg n$. Consideriamo dunque due varianti: l'analisi del discriminante lineare diagonale e l'analisi del discriminante lineare con *shrinkage*. Il primo metodo è il più semplice e prevede semplicemente di utilizzare come stima di Σ la matrice diagonale delle varianze campionarie. Il secondo metodo utilizza, invece, una matrice di covarianza con *shrinkage* ed esistono diverse possibilità per stimarla. Qui consideriamo la proposta di [18]: si tratta di una matrice della forma $\Sigma^* = \varrho_1 I + \varrho_2 S$, dove I è la matrice identità ed S è la matrice di covarianza campionaria. I coefficienti ϱ_1 e ϱ_2 vengono stimati in modo da minimizzare il

valore atteso $\mathbb{E}\left[\|\Sigma^* - \Sigma\|_F^2\right]$, dove Σ rappresenta la matrice di covarianza e $\|\cdot\|_F$ rappresenta la norma di Frobenius. Per i dettagli sulla stima si veda [18].

2.2.5 Importanza delle variabili

A seconda del modello considerato, esistono diversi metodi per valutare quali variabili hanno maggior influenza nel determinare la classificazione. Soprattutto per dati di genomica, dove si hanno migliaia di geni, è ragionevole pensare che solo parte di essi hanno un ruolo importante nella stima del modello e quindi nell'attribuzione delle osservazioni ad una classe piuttosto che ad un'altra. L'obiettivo di questi metodi è appunto individuare i geni più importanti nel prevedere ogni classe, per poter avere una miglior comprensione del fenomeno a livello biologico ed, in particolare, individuare dei marcatori delle condizioni studiate.
In questa tesi ho calcolato l'informazione sull'importanza delle variabili nei modelli *random forest*, *support vector machines* e PAM. Per ciascuno una breve descrizione di come si è ottenuta tale misura.

Foreste casuali Per le foreste casuali si utilizzano i campioni "OOB" (Out Of Bag), ovvero quelli esclusi dal campione bootstrap, per valutare l'errore del modello. Per ogni albero b della foresta, viene salvato l'errore di previsione sui campioni OBB e poi i valori del gene k-esimo vengono permutati in modo casuale e viene ricalcolato l'errore sul campione OBB. Per il gene k-esimo viene calcolata la media tra tutti gli alberi della diminuzione di accuratezza dovuta all'operazione di permutazione e questa viene considerata come misura di importanza del gene k-esimo [1]. Ho poi selezionato, per ogni classe, i geni con misura di importanza maggiore (in valore assoluto).

Support vector machines Per il modello di classificazione *support vector machines* con funzione kernel lineare, un metodo per valutare l'importanza delle variabili consiste nel valutare la grandezza, in valore assoluto, dei coefficienti β: maggiore è la quantità $|\beta_k|$, maggiore è l'importanza del gene k nella funzione di decisione [8].
Nel confronto tra G classi, avendo utilizzato una strategia "one-vs-one", si hanno

$G(G-1)/2$ modelli, e ogni classe compare in $G-1$ modelli. Per valutare i geni più importanti nel predire la classe g-esima ho deciso di utilizzare l'unione dei geni risultati più importanti nei singoli modelli contenenti la classe g.

PAM Per il modello PAM ho selezionato, per ogni classe g, i geni che presentano i valori maggiori, in valore assoluto, dei centroidi standardizzati, ovvero di $\frac{\bar{y}_{kg}-\bar{y}_g}{s_k}$.

2.2.6 Analisi di arrichimento

Le analisi di arricchimento sono analisi che, a partire da un insieme di geni differenzialmente espressi, cercano insiemi di geni (categorie di Gene Ontology, oppure geni appartenenti ad un pathway biologico) significativamente associati alla lista di geni di partenza. Questa fase di analisi può essere utile per l'interpretazione: oltre ad interpretare il ruolo dei singoli geni, si può avere una visione più generale sul processo biologico coinvolto.
Per valutare la significatività degli insiemi di geni considerati ho utilizzato test basati sulla distribuzione ipergeometrica. Siano N il numero totale di geni, dei quali m nella lista di geni differenzialmente espressi, e G un insieme di geni; tra i geni differenzialmente espressi, sia m_G il numero di geni appartenente all'insieme G, e, tra i geni totali, siano N_G quelli appartenenti all'insieme G (Tabella 2.2).

Si vuole valutare se il campionamento di m_G geni da G sia casuale oppure

	DEG	non DEG	TOT
G	m_G	$N_G - m_G$	N_G
non G	m_{G^C}	$N_{G^C} - m_{G^C}$	N_{G^C}
TOT	m	$N-m$	N

Tabella 2.2: Distribuzione dei geni

significativo. I p-value vengono calcolati nel seguente modo [4]:

$$p = \sum_{i=m_G}^{m} \frac{\binom{N_G}{i} \cdot \binom{N-N_G}{m-i}}{\binom{N}{m}}.$$

2.2.7 Source Set

Il metodo *SourceSet* [30] ha l'obiettivo di individuare, nel contesto dei modelli grafici, l'insieme di geni che sono l'origine delle differenze tra due condizioni, che costituiscono cioè la *disregolazione primaria*. Tali geni sono la reale fonte della perturbazione e vengono distinti dalla *disregolazione secondaria*, ovvero l'insieme di geni che risulta diversamente regolato nelle due condizioni a causa della propagazione della disregolazione primaria.
La definizione di *source set*, ovvero dell'insieme di geni causa della disregolazione primaria, è la seguente [10, 29].

2.1 Definizione. Sia V l'insieme di geni sotto studio e siano $X_V^{(1)}$ e $X_V^{(2)}$ i loro valori di espressione in due condizioni. L'insieme $D \subseteq V$ è detto *source set* se:

1. la distribuzione di $X_D^{(1)}$ differisce da quella di $X_D^{(2)}$;

2. le distribuzioni condizionate $X_{\bar{D}}^{(1)} \mid X_D^{(1)}$ e $X_{\bar{D}}^{(2)} \mid X_D^{(2)}$ coincidono, dove $\bar{D} = V \setminus D$.

Inoltre D è detto *source set minimale* se nessun suo sottoinsieme proprio è a sua volta un source set.

Il metodo si basa sull'analisi di pathway biologici che devono essere preventivamente trasformati in grafi, e successivamente in grafi cordali, ovvero grafi nei quali ogni ciclo di quattro o più vertici ha una corda. Quest'ultimo passaggio si ottiene attraverso le procedure di moralizzazione e triangolazione.
Si assuma di modellare lo stesso pathway, in due diverse condizioni sperimentali, come la realizzazione di due modelli grafici Gaussiani che condividono lo stesso grafo cordale G. $G = (V, E)$ è ottenuto dalla conversione del pathway in grafo, dove V ed E rappresentano, rispettivamente, i geni e le reazioni biochimiche. Siano $X^{(i)}$, $i = 1, 2$, le variabili che rappresentano i valori di espressione dei geni V in due condizioni. Si assume:

$$X^{(i)} \sim \mathcal{N}_p(\mu^{(i)}, \Sigma^{(i)}), \quad (\Sigma^{(i)})^{-1} \in \mathcal{S}^+(G), \quad i = 1, 2$$

dove p è il numero di geni e $\mathcal{S}^+(G)$ è lo spazio delle matrici simmetriche definite positive, con elementi nulli in corrispondenza degli archi mancanti in G.

I grafi cordali ammettono una scomposizione in sottografi più piccoli mediante separatori di cricche. Siano C_i, $i = 1, ..., k$ le cricche massimali del grafo G, ovvero i sottografi completi di G che non possono essere ulteriormente estesi, e S_i, $i = 2, ..., k$ i separatori minimali del grafo G. I separatori sono insiemi di nodi la cui rimozione dal grafo G induce una disconnessione nel grafo risultante. Sono minimali i separatori che non contengono sottoinsiemi propri che a loro volta sono separatori. Le cricche massimali possono essere ordinate in modo da soddisfare la *running intersection property*[1] ed esistono k tali possibili ordinamenti, uno per ogni scelta della cricca di partenza. Sia $C_{i,1}, ..., C_{i,k}$ l'ordinamento i-esimo, e sia $S_{i,1}, ..., S_{i,k}$ un opportuno riordinamento dei separatori, dove $S_{i,1} = \{\emptyset\}$ e tale che $S_{i,j} = C_{i,j} \cap \cup_{l=1}^{j-1} C_{i,l} = C_{i,j} \cap C_{i,\bar{j}}$, per un opportuno $\bar{j} < j$.
Si consideri l'ipotesi di uguaglianza della distribuzione nelle due condizioni: $H : \Sigma^{(1)} = \Sigma^{(2)} \quad \text{e} \quad \mu^{(1)} = \mu^{(2)}$. Dato l'ordinamento i-esimo delle cricche, tale ipotesi si scompone in un insieme di k ipotesi indipendenti $H_{i,j}$, $j = 1, ..., k$, di uguaglianza delle distribuzioni condizionate di $X_{C_{i,j} \setminus S_{i,j}} \mid X_{S_{i,j}}$, $j = 1, ..., k$. Tali ipotesi possono essere testate attraverso una scomposizione del test rapporto di verosimiglianza, specifica per ogni ordinamento i-esimo: $\lambda(V) = \sum_{j=1}^{k} [\lambda(C_{i,j}) - \lambda(S_{i,j})]$, dove $\lambda(C_{i,j}) - \lambda(S_{i,j})$ corrisponde al rapporto di verosimiglianza per testare l'ipotesi $H_{i,j}$.

2.1 Proposizione. *Sia $d_i^* = (d_{i,1}^*, ..., d_{i,k}^*)$ il vettore delle decisioni corrette (vere) per l'ipotesi $H_{i,j}$ ($d_{i,j}^* = 1$ quando $H_{i,j}$ è vera). L'insieme D_G, definito come:*

$$D_G = \bigcap_{i=1}^{k} D_{G,i}, \quad dove \quad D_{G,i} = \bigcup_{\{j : d_{i,j}^* = 1\}} C_{i,j}$$

è un source set.

In generale D_G non coincide con il source set minimale, tuttavia è il minor source set identificabile attraverso la decomposizione del grafo di partenza in cricche e separatori. L'insieme $\mathbb{D}_G = \cup_{i=1}^{k} D_{G,i}$ contiene tutti i geni affetti dalla perturbazione, mentre l'insieme $D_G \subseteq \mathbb{D}_G$ contiene i geni responsabili della disregolazione.

[1] Una sequenza di sottoinsiemi $F_1, ..., F_m$ di un insieme finito X soddisfa la *running intersection property* se

$$\forall k > 1 \quad \exists i < k \quad \text{tale che} \quad F_k \cap \bigcup_{j<k} F_j \subset F_i.$$

D_G viene chiamato source set (primario) e $\mathbb{D}_G \setminus D_G$ set secondario.
Di seguito descrivo la stima delle quantità teoriche descritte, come proposta in [10]. Sia ϕ_i il vettore corrispondente a d_i^*, contenente i risultati dei test di ipotesi $H_{i,j}$, $i, j = 1, ..., k$. La stima di D_G è data da:

$$\hat{D}_G = \bigcap_{i=1}^{k} \hat{D}_{G,i}, \quad \text{dove} \quad \hat{D}_{G,i} = \bigcup_{\{j:\phi_{i,j}=1\}} C_{i,j}.$$

Per il test rapporto di verosimiglianza è sufficiente calcolare le quantità:

$$\lambda(A) = \sum_{l=1}^{2} n_l \log \frac{|\hat{\Sigma}_A|}{|\hat{\Sigma}_A^{(l)}|}$$

per $A \in \{C_1, ..., C_k, S_1, ..., S_k\}$, dove $\hat{\Sigma}_A$ indica la sottomatrice di Σ corrispondente ai nodi in A, $|\hat{\Sigma}|$ è il determinante della stima di massima verosimiglianza di Σ sotto H, $\hat{\Sigma}^{(l)}$ la stima di massima verosimiglianza di $\Sigma^{(l)}$ sotto l'ipotesi alternativa, e n_l le numerosità campionarie nelle due condizioni.
Il test rapporto di verosimiglianza descritto è ben definito quando il numero di campioni del gruppo più piccolo è maggiore della cardinalità della cricca più grande. Poiché tale relazione spesso non è soddisfatta nell'ambito degli esperimenti di genomica, in [29] viene proposto l'utilizzo di uno stimatore Ridge, aggiungendo una piccola quantità alle diagonali delle matrici di covarianza stimate. Utilizzando tali stimatori la distribuzione nulla dei test non è più nota, pertanto ci si affida ad un metodo permutazionale.
Infine per la presenza di test multipli viene proposta una correzione, basata anch'essa su un metodo permutazionale per il calcolo della distribuzione dei p-value. Per i dettagli a riguardo si veda [29].

Capitolo 3

Preparazione dei dati ed analisi esplorative

In questo capitolo descrivo le fasi iniziali necessarie a preparare i dati per l'analisi, tra le quali il filtraggio dei geni, la normalizzazione tra campioni e l'eliminazione dei geni non annotati. La seconda parte del capitolo contiene invece alcune analisi esplorative che costituiscono un primo approccio ai dati e forniscono un'idea generale delle caratteristiche dei vari gruppi di soggetti.

3.1 Operazioni preliminari sui dati

I dati utilizzati provengono, come descritto nella sezione 2.1, da diversi esperimenti. Per procedere con le analisi è quindi necessario creare un unico dataset contenente tutte le osservazioni. Le piattaforme di microarray utilizzate sono però due diverse, pertanto l'unione dei dati ha comportato l'eliminazione dei *probe* non in comune. Entrambe le piattaforme, quella Agilent 028004, utilizzata per i campioni di tumore ovarico di stadio iniziale e per i tessuti sani di colon e rene, e quella Agilent 039494, utilizzata per i dati di tumore di stadi avanzati e per i tessuti sani di tube ed ovaio, consistono di 62976 probe. Ho innanzitutto eliminato i probe di controllo, ovvero probe disegnati appositamente per avere sempre un livello di espressione o molto basso o molto alto. L'intersezione dei probe rimanenti delle due piattaforme ha portato ad un totale di 51961 probe. All'interno di ogni piattaforma alcuni probe, tuttavia, avevano lo stesso nome e rappresentano lo stesso gene,

pertanto, per ogni identificativo di probe, ho tenuto un solo vettore di espressione, quello con maggior coefficiente di variazione, dato da $\text{cv}_k = \text{sd}_i(y_{ki})/\,\text{media}_i(y_k i)$, dove y_{ki} è l'espressione per il gene k nel campione i, e sd indica la deviazione standard. Oltre all'identificativo di probe, nei dati acquisiti viene riportato un identificativo, chiamato *RefSeq*, utilizzato nel database di sequenze di DNA ed RNA dell'NCBI (National Center for Biotechnology Information). Alcuni probe, sebbene condividessero lo stesso identificativo di probe tra le due piattaforme, non corrispondevano in realtà allo stesso identificativo RefSeq, pertanto in realtà si trattava di trascritti diversi. Queste osservazioni sono quindi state eliminate. A seguito di queste operazioni, i dataset risultanti contengono 28531 probe.

Filtraggio dei geni

L' operazione successiva sui dati è stata il filtraggio dei geni. Le piattaforme Agilent forniscono dei *flag* di qualità che indicano, per ogni probe, se il valore di espressione osservato è più o meno affidabile. In particolare i flag "IsPosAndSignif" e "IsWellAboveBG" indicano se il valore di segnale del probe è significativamente superiore al segnale del background dello spot (rumore di fondo). Ho filtrato i probe a disposizione mantenendo quelli che avevano flag positivo in almeno il 60% dei campioni, in almeno una classe. Come classi ho considerato: campioni tumorali di stadio I, campioni sani di tube ed ovaio, campioni tumorali di stadi avanzati ed infine campioni sani di rene e colon. Ho utilizzato, di volta in volta, il flag più restrittivo dei due. Quest'operazione ha portato ad eliminare 4199 probe il cui livello di espressione era molto basso e non significativo. A seguire, ho svolto un ultimo controllo per assicurare una corrispondenza tra i probe delle due diverse piattaforme. È risultato che alcuni probe, all'interno di ogni piattaforma, condividevano lo stesso identificativo RefSeq, pertanto, probe diversi rappresentavano in realtà lo stesso trascritto; per questi probe è stata fatta la medie dei valori di espressione, ottenendo così una matrice di espressione contenente 20399 probe corrispondenti ad altrettanti trascritti.

Normalizzazione

La normalizzazione è un passaggio solitamente fondamentale per l'analisi dei dati, in quanto serve ad uniformare i campioni, correggendo parte della variabilità

dovuta a cause tecniche e non a fenomeni biologici. Per valutare la bontà della normalizzazione scelta ho utilizzato un grafico RLE (Relative Log Expression), per il quale viene creato un boxplot per ogni campione, nel seguente modo:

1. sia y_{ki} il valore di espressione in scala logaritmica per il gene k e il campione i;
2. per ogni gene k si calcoli la mediana tra tutti i campioni $\text{Med}_i\, y_{ki}$ e la deviazione dalla mediana $y_{ki} - \text{Med}_i\, y_{ki}$;
3. per ogni campione si generi un boxplot di tutte le deviazioni.

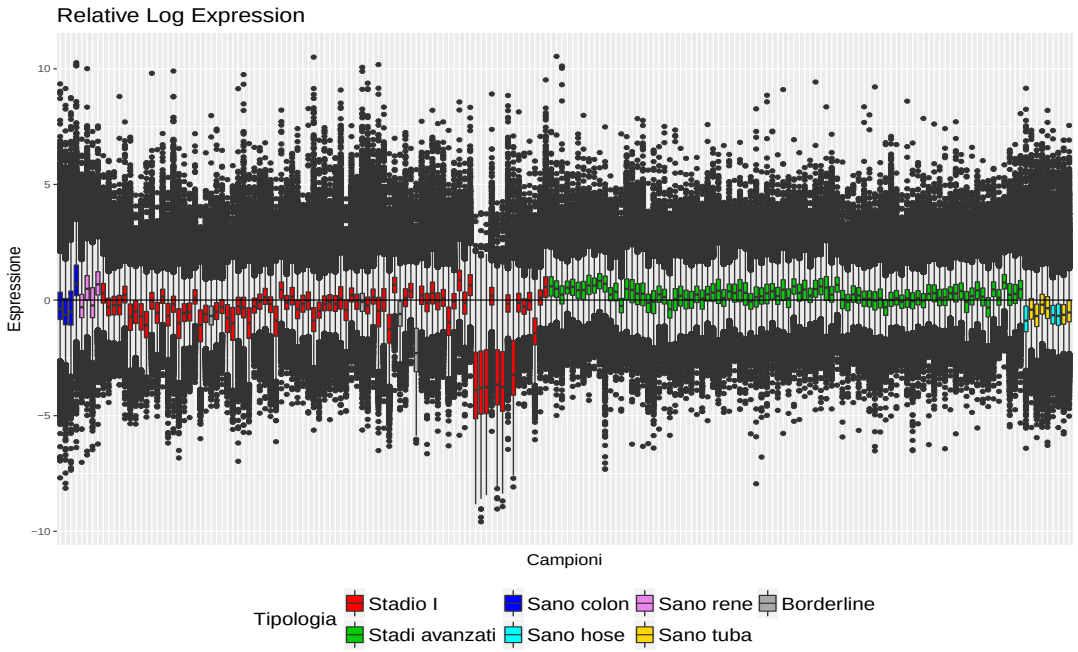
Date le assunzioni richieste dalla normalizzazione (si veda 2.2.1), il grafico RLE sui dati normalizzati dovrebbe idealmente mostrare tutti i boxplot all'incirca centrati sull'asse $M = 0$, con un intervallo di valori non troppo grande [13].
Ho applicato ai dati una normalizzazione loess ciclica fast utilizzando la funzione *normalizeCyclicLoess* del pacchetto *limma* [27, 34]. In figura 3.1b è riportato il grafico RLE sui dati precedenti alla normalizzazione e sui dati normalizzati; si osserva che la normalizzazione fornisce un buon risultato.

Eliminazione geni non annotati

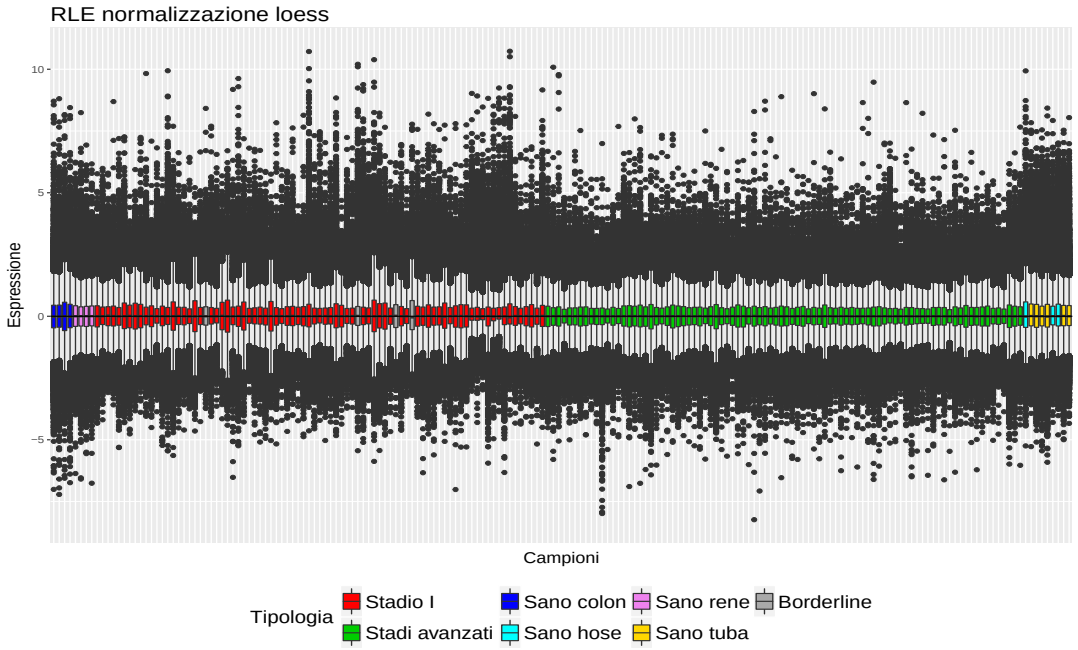
L'ultima operazione che ho effettuato sui dati, prima di procedere con le analisi, è stata l'eliminazione dei probe che non hanno più un'annotazione, ovvero ai quali non corrisponde più un gene. Ciò è dovuto al fatto che, negli anni, le piattaforme di microarray vengono aggiornate, come vengono anche aggiornate le informazioni sul genoma umano. I probe che non hanno un annotazione sono quindi vecchi probe che, al giorno d'oggi, non corrispondono più a dei geni; per questo motivo li ho eliminati dalla matrice di dati.
Quest'operazione ha portato ad avere una matrice finale di dati contenente 18191 probe e 188 campioni.

3.2 Analisi esplorative

I dati analizzati in questa tesi comprendono sia tessuti sani che tessuti malati, come già introdotto in precedenza. In totale si hanno 17 campioni di tessuti sani (4 del rene, 4 del colon, 3 dell'ovaio, 6 delle tube), 4 campioni di tumore borderline,



(a) grafico RLE pre normalizzazione



(b) grafico RLE post normalizzazione

Figura 3.1: Grafici RLE

79 campioni di tumore allo stadio iniziale, 88 campioni di tumore a stadi avanzati. I tessuti sani dell'ovaio sono indicati anche con "hose" (Human Ovarian Surface Epitelium). In tabella 3.1 sono riassunte le distribuzioni delle principali caratteristiche per i campioni tumorali, rilevate all'ultimo follow-up delle pazienti. Il grado indica il livello di differenziazione delle cellule. Le cellule tumorali, infatti, tendono a perdere la differenziazione acquisita in quanto cellule mature e assomigliano sempre meno alle cellule del tessuto di partenza. Il grado 1 (G1) indica che le cellule sono ben differenziate e quindi molto simili al tessuto di partenza; il grado 2 (G2) indica cellule moderatamente differenziate; il grado 3 (G3) indica cellule scarsamente differenziate o indifferenziate. Lo stato descrive la condizione della paziente al momento del follow-up. La classificazione FIGO include l'informazione sullo stadio (qui I, III oppure IV) ed un ulteriore suddivisione indicata con A, B e C per gli stadi I e III[1]. Per gli stadi iniziali viene riportata l'informazione se, al momento del follow-up, si è già verificato un episodio di recidiva. L'indicazione "tumore residuo", per gli stadi avanzati, indica la dimensione del tumore rimasto dopo l'operazione. Nessun tumore residuo è classificato come 0, un tumore residuo inferiore ad 1 cm è classificato come 1, ed infine uno superiore al centimetro è classificato come 2. "Platinum Status" indica, per gli stadi avanzati, la (parziale) sensibilità o resistenza alla chemioterapia al platino, definite in termini di mesi dalla fine del trattamento fino ad una progressione della malattia: resistente (1-6 mesi), parzialmente sensibile (6-12 mesi), sensibile (> 12 mesi).

In questa prima fase di analisi ho iniziato ad esplorare i dati con dei metodi non supervisionati per vedere se emergono evidenti raggruppamenti dei campioni in base alla loro tipologia. In particolare si cerca di vedere come si dispongono i tessuti sani rispetto a quelli malati, la relazione tra i diversi istotipi e gli stadi avanzati.

[1]IA: tumore limitato ad un ovaio, capsula intatta, assenza di tumore sulla superficie dell'ovaio, assenza di ascite e lavaggi peritoneali negativi.
IB: tumore limitato ad entrambe le ovaie, con caratteristiche come IA.
IC: tumore limitato ad una o entrambe le ovaie, con una delle seguenti: rottura della capsula, tumore sulla superficie dell'ovaio, ascite con cellule maligne o lavaggio peritoneale positivo.
IIIA: linfonodi retroperitoneali positivi e/o micrometastasi peritoneali extrapelviche.
IIIB: metastasi macroscopiche peritoneali extrapelviche (≤ 2 cm), con o senza linfonodi retroperitoneali positivi.
IIIC: metastasi macroscopiche peritoneali extrapelviche (> 2 cm), con o senza linfonodi retroperitoneali positivi. Include l'estensione alla capsula fegato/milza.
IV: Metastasi a distanza, oltre alle metastasi peritoneali.

Stadio Iniziale		
Istotipo[1]	Cc	17
	End	23
	Muc	16
	Sier	23
Grado	G1	19
	G2	24
	G3	36
Stato[2]	AWD	2
	DOC	2
	DOD	19
	NED	56
FIGO	IA	23
	IB	5
	IC	51
Recidiva	Si	27
	No	52

Stadi Avanzati		
Tumore Residuo	0	18
	1	19
	2	51
Stato	AWD	9
	DOC	4
	DOD	63
	NED	12
FIGO	IIIA	2
	IIIB	1
	IIIC	63
	IV	22
Platinum Status[3]	No CT	1
	PS	17
	S	34
	R	36

Tabella 3.1: Distribuzioni delle caratteristiche per i campioni tumorali

[1] Cc: Clear Cell, End: Endometrioide, Muc: Mucinoso, Sier: Sieroso;
[2] AWD: Vivo con la malattia (Alive With Disease), DOC: Morto per altra causa (Dead of Other Cause), DOD: Morto per malattia (Dead Of Disease), NED: Nessuna evidenza di malattia (No Evidence of Disease);
[3] No CT: Nessuna chemioterapia, PS: Parzialmente sensibile, S: Sensibile, R: Resistente.

Ho confrontato i risultati delle analisi esplorative svolte su due diversi insiemi di dati: l'intero campione a disposizione e un sottoinsieme di 10000 geni, individuati con l'obiettivo di ridurre il numero di geni che apportano rumore. La selezione è stata fatta scegliendo i geni con maggior coefficiente di variazione.

Cluster gerarchico Come prima cosa ho utilizzato il cluster gerarchico con *linkage* completo, implementato nella funzione *hclust* del pacchetto *stats* [24], e rappresentato graficamente come un albero, utilizzando il pacchetto *ape* [21].
In riferimento alle figure 3.2 e 3.3, si ha che, partendo dall'esterno, i campioni vengono considerati tra loro simili tanto più rapidamente vengono raggruppati sotto uno stesso nodo. Ogni campione è stato etichettato con il nome della classe di

appartenenza, e gli istotipi dello stadio iniziale sono stati colorati allo stesso modo dei tessuti sani di riferimento; il tumore sieroso è stato associato al tessuto delle tube.

Nella figura 3.2 si osserva che i campioni di stadi avanzati, tranne qualche ec-

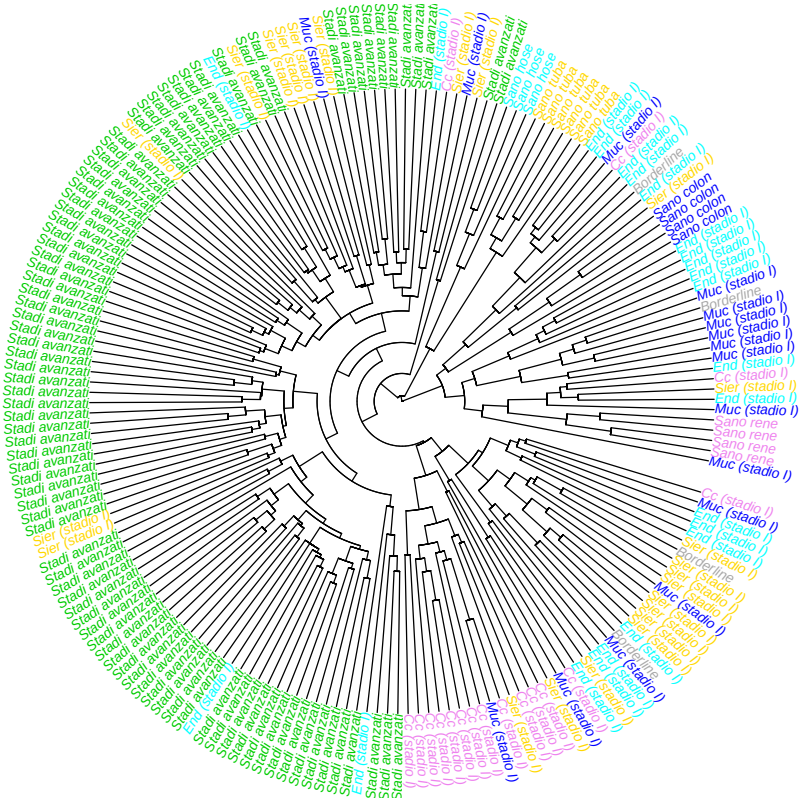


Figura 3.2: Cluster gerarchico su tutti i geni

cezione, vengono raggruppati tra loro prima di essere accorpati ad altri gruppi: anche se si formano dei sotto gruppi, i campioni di stadi avanzati sembrano ben divisi dagli altri campioni. Per quanto riguarda i tessuti sani, si osserva che ciascun tessuto viene subito raggruppato agli altri della propria tipologia; inoltre i campioni di tube ed hose sembrano essere molto simili tra loro in quanto, dove

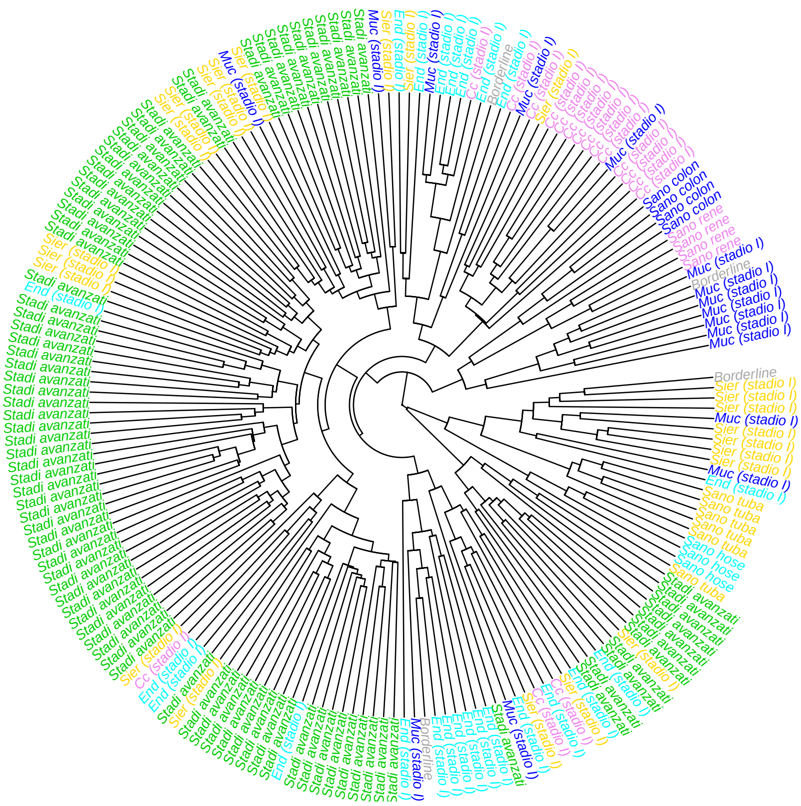


Figura 3.3: Cluster gerarchico sulla selezione di geni

aver formato i rispettivi gruppi, vengono uniti, al passaggio successivo, sotto un unico nodo. I diversi istotipi dei campioni tumorali di stadio iniziale sono quelli meno omogenei: sebbene alcuni campioni dello stesso istotipo vengano raggruppati assieme abbastanza velocemente, molti altri rimangono separati fino ai livelli più alti dell'albero. Quelli meglio raggruppati sono i campioni clear cell. Alcuni campioni mucinosi si avvicinano ai tessuti sani del colon, mentre gli altri campioni sani non mostrano un'evidente vicinanza ad un istotipo piuttosto che ad un altro. I campioni borderline sono distribuiti senza mostrare un particolare raggruppamento.

Nel cluster gerarchico effettuato sulla selezione di geni sopra descritta (figura 3.3) si osserva un andamento complessivamente simile, ma con qualche differenza. Innanzitutto alcuni campioni di stadio avanzato (in basso a destra) non vengono accorpati a tutti gli altri, ma formano un gruppo a sé. Inoltre i tessuti sani delle tube mostrano una somiglianza con alcuni campioni sierosi di stadio iniziale, mantenendo comunque, in secondo piano, il raggruppamento con i tessuti sani dell'ovaio. I tessuti sani del rene vengono raggruppati assieme a quelli del colon, che mantengono comunque la somiglianza con alcuni dei tessuti di tumore mucinoso.

Scaling multidimensionale Come secondo approccio ho deciso di utilizzare lo scaling multidimensionale, che fornisce una rappresentazione grafica dei dati in uno spazio di dimensione molto più piccola di quella di partenza. Il metodo preserva al meglio le distanze tra le osservazioni, pertanto ritengo sia un buon strumento per visualizzare i campioni ed avere un'idea delle loro similarità.
Ho implementato il metodo con la funzione *cmdscale* del pacchetto *stats* [24], con uno spazio di arrivo di dimensione 3. I risultati dello scaling su tutti i dati e sul sottoinsieme di 10000 sono molto simili, con conclusioni analoghe. Riporto qui i grafici relativi all'intero campione di dati, raffiguranti le osservazioni nelle nuove componenti, date dallo scaling. Si osservi il primo grafico in figura 3.4. I campioni di stadi avanzati sono ben raggruppati tra loro e distinti dagli altri campioni. I tessuti sani dello stesso tipo risultano sempre vicini tra loro; tra questi, i tessuti di tube sono i più lontani dal resto dei dati. Per quanto riguarda lo stadio iniziale, si può notare il raggruppamento di alcune unità all'interno di ogni istotipo, tuttavia molte si mescolano tra loro. Qualche campione di quelli sierosi di stadio iniziale si dispone all'interno del gruppo degli stadi avanzati. I campioni borderline sono sparsi e non formano un gruppo a sé stante. Il secondo grafico in figura, rispetto al primo, suggerisce che anche i tessuti sani delle hose siano in realtà lontani dal resto dei dati e simili ai tessuti delle tube. Il terzo grafico, infine, conferma quanto detto finora per i tessuti sani e gli stadi iniziali, mentre gli stadi avanzati non vengono ben distinti dagli stadi iniziali.

Le analisi presentate finora suggeriscono la presenza di una distinzione abbastanza netta tra i campioni di stadio iniziale e quelli di stadi avanzati, con qualche

eccezione, soprattutto per alcuni campioni sierosi di stadio iniziale. Inoltre, il cluster gerarchico potrebbe suggerire una suddivisione degli stadi avanzati in più gruppi. Gli istotipi dello stadio iniziale non sembrano ben distinti l'uno dall'altro. I tessuti sani sono ben raggruppati per tipologia, ma i vari gruppi hanno comportamenti diversi: tube ed hose sembrano distanti dal resto delle osservazioni, ma simili tra loro; colon e rene sono più simili al resto dei dati, in particolare i tessuti del rene che spesso si collocano in mezzo ad altre osservazioni di tessuti tumorali. Nei prossimi capitoli cercherò di rispondere ai tre quesiti formulati nell'introduzione della tesi, e molti degli aspetti emersi in queste analisi esplorative compariranno anche nelle analisi successive.

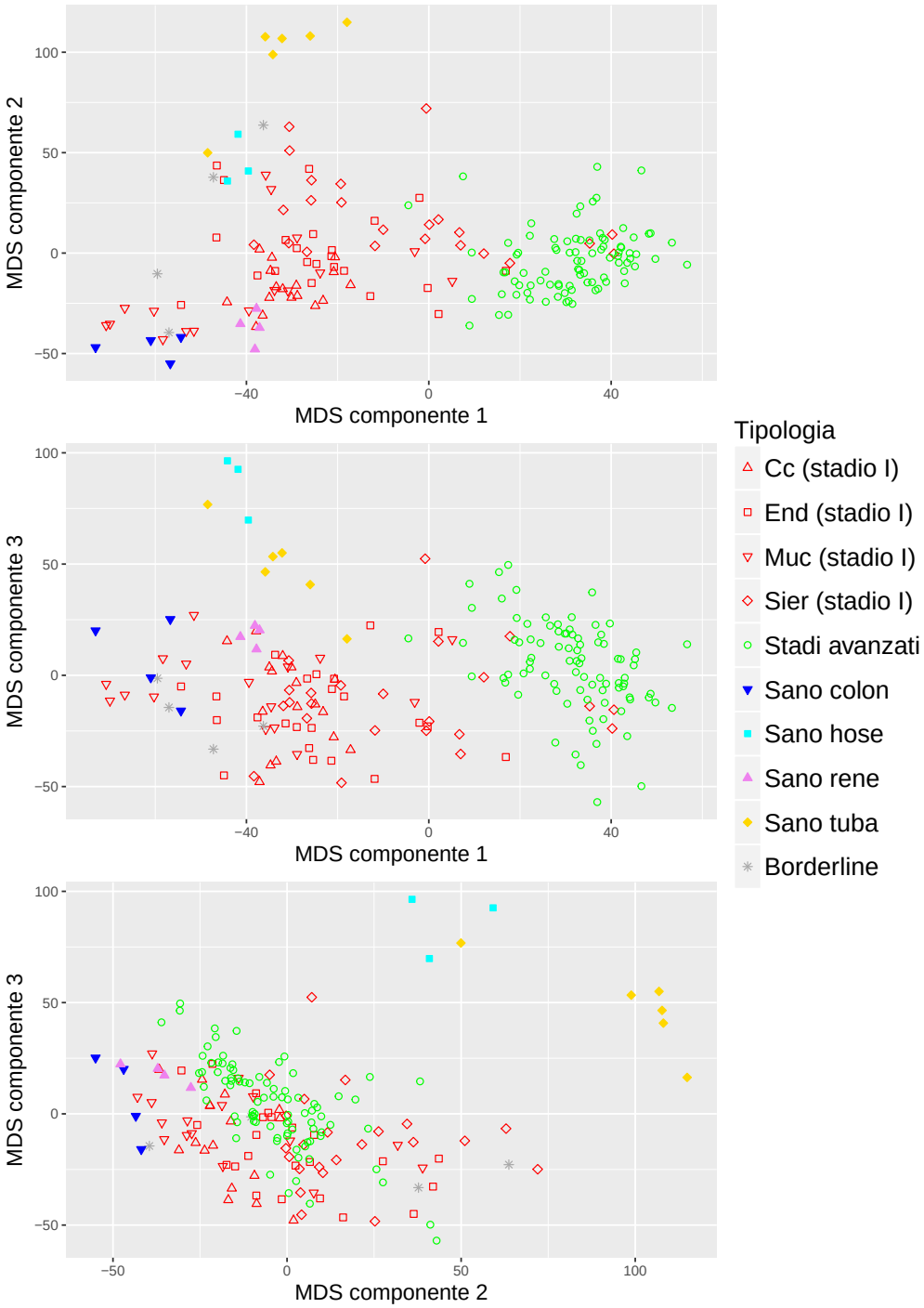


Figura 3.4: Scaling multidimensionale

Capitolo 4

Confronti tra tessuti sani e malati

In questo capitolo affronto il primo obiettivo di questa tesi, ovvero valutare se i campioni tumorali sierosi sono più simili ai tessuti sani delle tube o dell'ovaio, per avere un'indicazione sul possibile sito di origine del tumore. Nella prima sezione cerco di ripercorrere, con i miei dati, la strategia proposta in un articolo su *Nature Communications* [11], utilizzando quindi un approccio di correlazione per capire se i campioni sierosi possano essere ritenuti più simili al tessuto sano dell'ovaio oppure delle tube. Nella seconda sezione affronto lo stesso problema con modelli di classificazione, introducendo anche nuove componenti. Procedo, infatti, adattando dei modelli su tutti i tessuti sani e, attraverso questi, classifico i campioni tumorali: ogni campione tumorale sarà quindi etichettato come uno dei tessuti sani. Utilizzo l'informazione sulla classificazione dei campioni sierosi come criterio per stabilire una maggior somiglianza con un tessuto piuttosto che un altro, mentre utilizzo gli altri campioni per valutare la bontà del modello, osservando come si distribuiscono rispetto ai loro tessuti di riferimento.

4.1 Approccio tramite correlazione

Una parte del lavoro di Ducie et al. (2017) intitolato *Molecular analysis of high-grade serous ovarian carcinoma with and without associated serous tubal intra-epithelial carcinoma*, è dedicata al confronto tra campioni tumorali sierosi con tre tessuti sani di riferimento: tessuto sano dell'ovaio, delle tube e peritoneo. Per questa fase dell'analisi nell'articolo sono stati utilizzati 85 campioni di tumore sieroso,

e dieci profili di tessuti sani, ciascuno costituito dall'aggregazione di più materiale biologico.
I dati sono stati ottenuti con la tecnologia RNAseq, che consiste nel sequenziamento diretto dei trascritti di mRNA, senza avvalersi di sequenze fissate, come accade invece per i microarray.
Per definire a quale tessuto sano assomigliava maggiormente ogni campione tumorale, sono stati considerati 50 geni marcatori di ogni tessuto sano, ottenuti valutando i geni maggiormente differenzialmente espressi tra un tessuto sano fissato e i due rimanenti. È stata quindi calcolata la correlazione di Spearman tra i campioni tumorali e i diversi tessuti sani, sui geni selezionati. La correlazione di Spearman è data dalla correlazione classica, calcolata però sui ranghi dei valori di partenza, invece che sui valori stessi. È risultato che 75 campioni su 85 avevano maggior correlazione col tessuto sano delle tube, rispetto agli altri, suggerendo le tube come sito di origine più verosimile. Nell'ultimo decennio anche in molti altri studi è stato ipotizzato che l'origine di molti tumori ovarici sierosi non sia l'ovaio, ma le tube di Falloppio. Ho deciso quindi di indagare quest'aspetto anche nella mia tesi, utilizzando come punto di partenza il metodo della correlazione appena descritto. Ho svolto quest'analisi separatamente per i campioni sierosi di stadio I e per quelli di stadi avanzati, per poter cogliere eventuali differenze tra i due gruppi. Ho considerato tre test per valutare la differenziale espressione tra i tessuti di tube e quelli di ovaio: test SAM, test bayesiano empirico e test *product rank*. Per selezionare i geni ho utilizzato la significatività del test bayesiano empirico, considerando però solo i geni che risultavano significativi al 5% in tutti e tre i test. Ho così selezionato i 50 e i 100 geni più significativi tra i differenzialmente espressi. Per ogni campione di tumore sieroso ho quindi calcolato la correlazione di Spearman su questi geni, utilizzando, per i tessuti sani, il valore medio, e inoltre ho considerato un intervallo di confidenza bootstrap al 90% ottenuto tramite la funzione *cor.ci* del pacchetto *psych* [26].
I risultati nei due casi, con 50 e 100 geni, portano alle stesse conclusioni, tuttavia nel caso di 100 geni si ottengono intervalli di confidenza più stretti che, per qualche campione, consentono di evitare la sovrapposizione tra l'intervallo di confidenza riferito alle tube e quello riferito all'ovaio. Riporto quindi i risultati di quest'ultimo caso. In figura 4.1 gli intervalli di confidenza per le correlazioni di tube ed ovaio, per tutti i campioni.

Classificando i campioni in "hose-like" e "tube-like" in base alla correlazione più

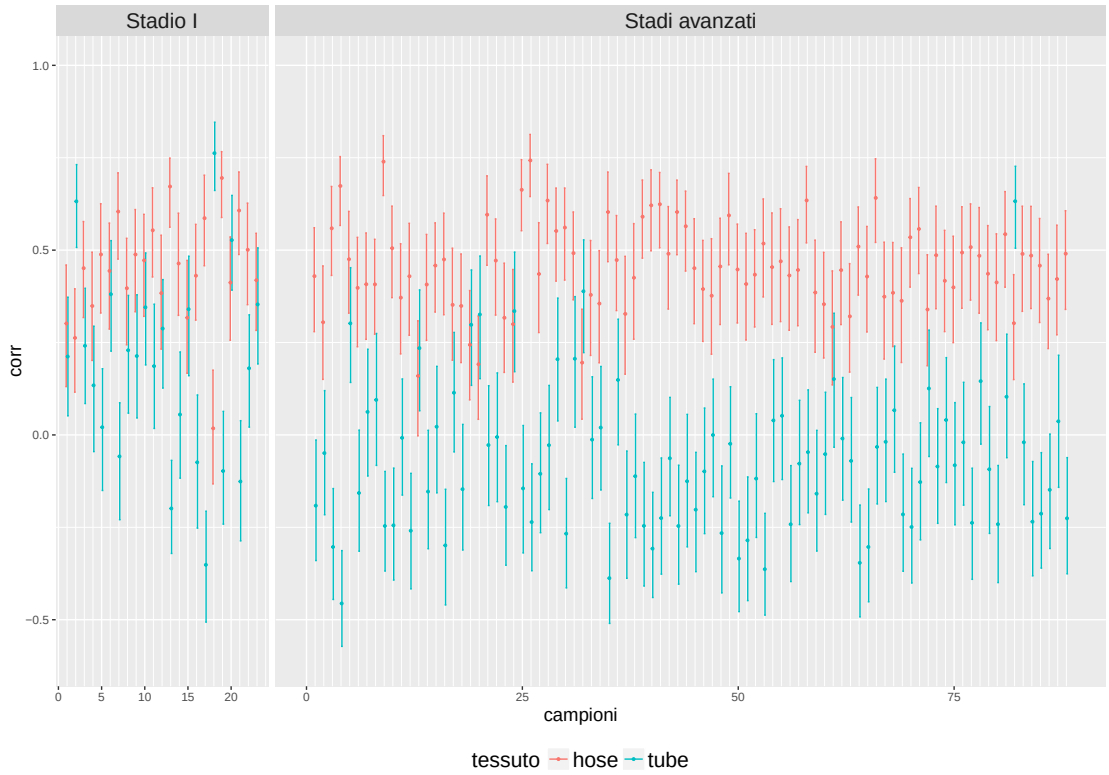


Figura 4.1: Intervalli di confidenza per le correlazioni

alta, e lasciando non classificati quelli per i quali gli intervalli di confidenza relativi a tube e a hose si sovrappongono, si ottengono i risultati riportati nella prima parte della tabella 4.1 (indicati con "100 DEG"). Si osserva che, esclusi i campioni

		Tube-like	Hose-like	Non classificati	Tot
100 DEG	Stadio I	2	10	11	23
	Stadi avanzati	1	75	12	88
EB-2000	Stadio I	4	11	8	23
	Stadi avanzati	3	77	8	88
RP-2000	Stadio I	4	10	9	23
	Stadi avanzati	3	73	12	88

Tabella 4.1: Classificazione "Tube-like" ed "Hose-like"

non classificati, la quasi totalità delle osservazioni è classificata come "hose-like",

sia per lo stadio iniziale, che per gli stadi avanzati.
Ho così deciso di ripetere le analisi con altri due insiemi di geni molto più numerosi, per valutare la stabilità dei risultati. Ho usato due insiemi di 2000 geni ciascuno, formati dai 1000 geni maggiormente up-regolati e dai 1000 geni maggiormente down-regolati in un tessuto rispetto all'altro, ottenuti una volta attraverso il test bayesiano empirico ed una volta attraverso il test product rank.
I valori di correlazione e gli intervalli ottenuti sono diversi dai precedenti, tuttavia la relazione tra le correlazioni dei due tessuti è simile. In figura 4.2 i valori di correlazione con gli intervalli di confidenza al 90% ("EB-2000" indica il test bayesiano empirico, "RP-2000" il rank product). Le classificazioni in "hose-like"

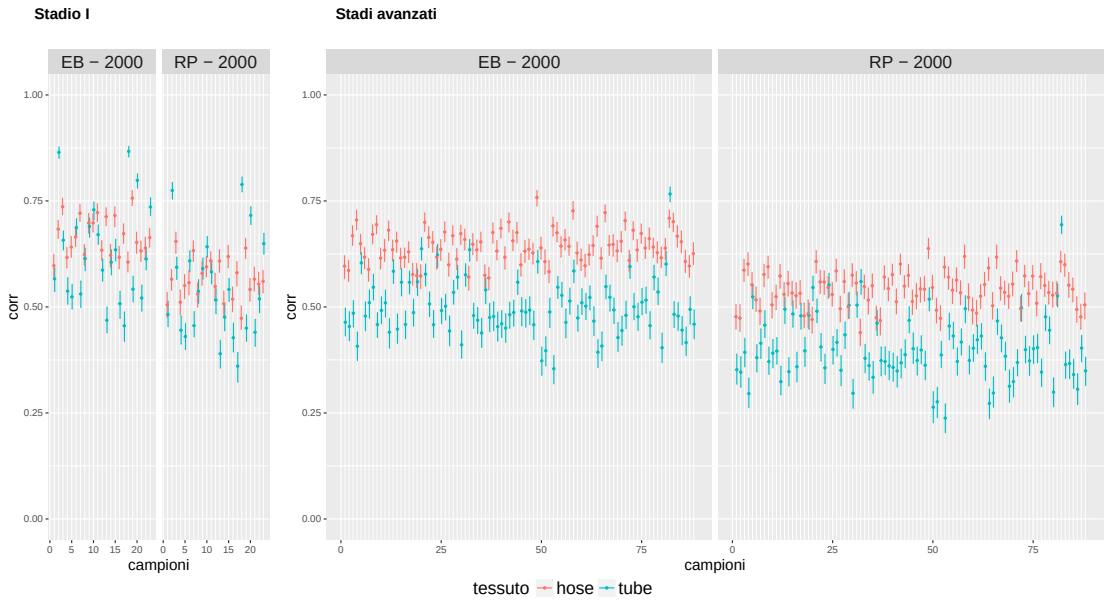


Figura 4.2: Intervalli di confidenza per le correlazioni

e "tube-like" sono simili a quella ottenuta con i 100 geni differenzialmente espressi, anche se, complessivamente, viene classificata qualche unità in più rispetto a prima. Nella seconda parte della tabella 4.1, è riportata la classificazione ottenuta utilizzando rispettivamente i geni selezionati con il test bayesiano empirico (EB-2000) e il test rank product (RP-2000). Nelle tabelle 4.2 e 4.3 le due nuove classificazioni sono confrontate con la precedente ("NC" per "Non classificati"), e si può vedere che sono molto simili.

		Classif. RP		
		NC	hose-like	tube-like
Classif. 100 DEG	Classif. EB			
NC	NC	6	0	0
	hose-like	0	3	0
	tube-like	0	0	2
hose-like	NC	2	0	0
	hose-like	1	7	0
	tube-like	0	0	0
tube-like	NC	0	0	0
	hose-like	0	0	0
	tube-like	0	0	2

Tabella 4.2: Classificazione "Tube-like" ed "Hose-like", stadio I

		Classif. RP		
		NC	hose-like	tube-like
Classif. 100 DEG	Classif. EB			
NC	NC	5	0	0
	hose-like	3	2	0
	tube-like	0	0	2
hose-like	NC	3	0	0
	hose-like	1	71	0
	tube-like	0	0	0
tube-like	NC	0	0	0
	hose-like	0	0	0
	tube-like	0	0	1

Tabella 4.3: Classificazione "Tube-like" ed "Hose-like", stadi avanzati

Al di là delle piccole differenze nei risultati in base ai geni utilizzati, escludendo i campioni non classificati, questo approccio basato sulla correlazione sembra suggerire che, più del 70% dei campioni di stadio I, e più del 95% di quelli di stadio avanzato assomigliano maggiormente al tessuto sano dell'ovaio piuttosto che delle tube.
Nella prossima sezione affronto lo stesso problema utilizzando modelli di classificazione, nei quali includo anche i rimanenti istotipi con i relativi tessuti sani.
L'approccio basato sulla correlazione è un metodo semplice per attribuire un campione ad un tessuto piuttosto che ad un altro. Se per alcuni aspetti la sua semplicità

rappresenta un pregio, d'altra parte il metodo potrebbe non essere sufficientemente flessibile per cogliere le caratteristiche d'interesse. Non avendo informazioni sulla "verità" riguardo la classificazione dei campioni, alla fine del capitolo confronterò i risultati suggeriti dall'approccio tramite correlazione con quelli ottenuti dai modelli di classificazione, per vedere se sono concordi.

4.2 Approccio di classificazione

In questa sezione utilizzo dei modelli di classificazione per cogliere le somiglianze tra istotipi tumorali e tessuti sani di riferimento. I modelli vengono stimati sull'insieme di tessuti sani e poi usati per prevedere la classe di appartenenza dei campioni tumorali. Per gli istotipi clear cell, endometrioide e mucinoso, la corretta classificazione è data dai rispettivi tessuti di riferimento. Per i campioni sierosi si vuole invece osservare la classificazione data dai vari modelli e, se possibile, individuare un'indicazione complessiva di maggior somiglianza ad uno dei due tessuti (hose oppure tube).
Ho utilizzato cinque diverse tipologie di modelli, ciascuna adattata su due sottoinsiemi di geni. I modelli utilizzati sono: *random forest*, *support vector machines*, PAM, analisi del discriminante lineare con matrice di covarianza diagonale (Dlda) ed analisi del discriminante lineare con matrice di covarianza con *shrinkage* (Slda). Per stimare tali modelli ho usato, rispettivamente, le funzioni *randomForest*, *svm*, *pamr.train*, *Dlda*, *Slda*, dei pacchetti *randomForest* [19], *e1071* [20], *pamr* [15], *HiDimDA* [23]. Il primo sottoinsieme di geni ("Selezione con cv") è dato dai 2000 geni con maggior coefficiente di variazione calcolato sui quattro tessuti sani; il coefficiente di variazione è dato da: $\text{cv}_k = \text{sd}_i(y_{ki})/\,\text{media}_i(y_k i)$, dove y_{ki} è l'espressione per il gene k nel campione i, e sd indica la deviazione standard. Il secondo sottoinsieme di geni ("Selezione con DEG") è costituito da 1400 geni, dati dall'unione di 350 geni per ogni tessuto, individuati come più significativi nella differenziale espressione del tessuto considerato, rispetto ai rimanenti. Nelle situazioni in cui ho utilizzato la convalida incrociata, ho creato le divisioni dei campioni in *fold*, bilanciati per classe, attraverso la funzione *createFolds*, del pacchetto *caret* [17].
Tutti i modelli sono adattati sui tessuti sani, pertanto su 17 osservazioni. Di seguito i dettagli sui parametri di ogni modello.

Random Forest: foresta con 5000 alberi, $\sqrt{\tilde{p}}$ variabili usate in ogni nodo (44 per la selezione con cv, 37 per la selezione con DEG), dove $\tilde{p}$ è il numero di geni utilizzati, e soglia per l'attribuzione della classe "vincente" proporzionale alla numerosità delle classi.
Support vector machines: ho valutato il parametro C, il parametro gamma (necessario per le funzioni kernel, eccetto quella lineare) e il tipo di funzione kernel attraverso convalida incrociata, separatamente per ciascun insieme di geni. A parità di errore di convalida incrociata ho fissato gamma al valore di *default* $1/\tilde{p}$ (dove $\tilde{p}$ è il numero di geni utilizzati), mentre ho scelto C di volta in volta, in base al modello. Ho mantenuto invece tutte le funzioni kernel con errore minimo, adattando un modello per ciascuna. Infine, avendo classi asimmetriche, ho impostato un peso per le classi, inversamente proporzionale alla loro numerosità, il quale, moltiplicato per C, rappresenta il costo associato alla violazione dei margini.
PAM: Come valori di soglia per la selezione dei geni vengono scelti internamente dalla funzione 30 possibili valori, per i quali vengono riportati gli errori di convalida incrociata commessi. Ho quindi provato ad utilizzare alcuni dei valori proposti, scegliendoli nel seguente modo: fissato il numero di errori commessi (0,1,2,3), ho selezionato la soglia maggiore che consentiva di rimanere entro quel numero di errori. In ogni contesto ho poi valutato il modello migliore. Le probabilità a priori di appartenenza ad una classe sono impostate di *default* alle proporzioni delle varie classi nell'insieme di stima.
Dlda e Slda: Le probabilità a priori vengono anche qui impostate di *default* alle proporzioni delle classi. Per il metodo con *shrinkage* ho utilizzato come matrice "target" una matrice diagonale costante (si veda par. 2.2.4).

In tabella 4.4 ho riportato le tabelle di classificazione, per i due insiemi di geni utilizzati, per i modelli definitivi scelti. Nessun modello mostra una buona previsione sui nuovi dati di campioni tumorali, infatti in molti modelli la maggior parte delle osservazioni viene classificata come tessuto sano del rene e, nei modelli restanti, le unità tendono comunque ad essere classificate prevalentemente in un'unica classe, creando così una previsione sbilanciata. Riporto qui due modelli per ogni tipologia, uno per ogni selezione dei geni. In particolare i modelli *support vector machines* sono con: kernel polinomiale di grado 3, costo $C = 5$ e gamma con valore di default (selezione con cv); kernel lineare, costo $C = 1$ e gamma con

valore di default (selezione con DEG). I modelli PAM, uno per la selezione con DEG e uno per la selezione con cv, sono con soglie 1.984 e 3.177, che producono, sull'insieme di stima, rispettivamente 0 e 1 errori di convalida incrociata. Per le *support vector machines*, i risultati sono molto instabili, tuttavia la conclusione è sempre simile: le unità non vengono ben distribuite nelle diverse classi, ma si concentrano prevalentemente in una classe o due. Una possibile spiegazione della previsione non soddisfacente, emersa già nelle analisi esplorative, è che il tessuto del rene sembra molto simile alla maggioranza dei campioni tumorali, mentre gli altri tessuti sani sembrano mantenere una certa distanza. Potrebbe essere questo il motivo per cui la maggior parte dei modelli fornisce una classificazione molto sbilanciata e non riesce, invece, a classificare i campioni tumorali distribuendoli tra i vari tessuti. La classificazione di maggior interesse è, tuttavia, quella dei campioni sierosi, e ci si aspetta una maggior somiglianza con i tessuti delle tube oppure delle hose. Per questo motivo, per cercare di superare i problemi riscontrati nei modelli proposti, ho provato ad eliminare i tessuti sani del rene e i campioni di istotipo clear cell, per vedere se la classificazione migliora. Ho dunque riadattato i modelli già descritti, ricalcolando gli insiemi di geni per la stima dei modelli solo sui tre tessuti sani rimanenti. Per la selezione con coefficiente di variazione ho utilizzato sempre 2000 geni, mentre per la selezione sui geni differenzialmente espressi ne ho scelti 1050 (sempre 350 per ogni tessuto). Per la selezione dei parametri ho utilizzato gli stessi criteri descritti per i modelli con tutti i tessuti. Per valutare la bontà del modello ho osservato la classificazione dei campioni mucinosi, che hanno un chiaro legame con i tessuti sani del colon, e dei campioni endometrioidi, rispetto al tessuto delle hose. Le tabelle di classificazione dei campioni tumorali sono riportate nella tabella 4.5.

I modelli *random forest*, sia per la selezione con cv, che per la selezione con DEG, mostrano uno sbilanciamento verso il tessuto sano delle hose: per il primo insieme di geni la maggioranza dei campioni mucinosi viene infatti classificata in questo tessuto, ed anche per il secondo insieme di geni si osserva un comportamento simile, leggermente meno accentuato. Pertanto, anche se la totalità dei campioni di stadi avanzati e la quasi totalità di quelli sierosi di stadio I vengono classificati come tessuti sani delle hose, non si può ritenere questo risultato affidabile.

Per le *support vector machines* i risultati sono di nuovo molto instabili e la maggior parte delle unità si concentra in un'unica classe. Per la selezione con coefficiente di

variazione ho riportato il modello con funzione kernel radiale, costo $C = 5$ e gamma con valore di default, che è uno di quelli ad errore 0 sulla convalida incrociata. Sempre con lo stesso criterio ho utilizzato kernel lineare, costo $C = 1$ e gamma al valore di default nel modello sulla selezione con DEG.
Il modello PAM mostra una discreta previsione dei campioni tumorali. Per la selezione con cv, ho scelto una soglia pari a 0, in quanto per gli altri valori utilizzati si osservava uno sbilanciamento verso il tessuto sano del colon. Per la selezione con DEG, invece, i modelli con soglie 3.863 e 5.151 (corrispondenti rispettivamente a 0 e 1 errore di convalida incrociata sull'insieme di stima) mostrano entrambi una buona classificazione delle nuove unità. Riporto i risultati del modello con soglia 3.863. In entrambi i modelli presentati, la maggior parte dei campioni mucinosi viene classificata come tessuto del colon, così come la maggioranza dei campioni endometrioidi viene classificata come tessuto delle hose. Tale comportamento ovviamente non garantisce una buona capacità predittiva del modello nei confronti dei campioni sierosi, tuttavia è un'indicazione che i modelli riescono a separare almeno quelle due classi, associando ciascuna al proprio tessuto di riferimento. I campioni sierosi, sia di stadio I che di stadi avanzati, vengono classificati da entrambi i modelli come hose, ad eccezione di qualche unità.
Per quanto riguarda i modelli di analisi del discriminante lineare (lineare e con *shrinkage*), le previsioni più attendibili sono quelle dei modelli adattati sui geni selezionati con coefficiente di variazione. Il modello Dlda classifica tutti i campioni sierosi di stadi avanzati, e quasi tutti quelli sierosi di stadio iniziale, come tessuto sano delle hose. Il modello Slda classifica la maggior parte dei campioni sierosi di stadi avanzati come hose, e i restanti come tube. I campioni sierosi di stadio iniziale sono invece divisi, circa a metà, tra hose e tube. I modelli adattati sulla selezione con DEG forniscono risultati simili, tuttavia i campioni mucinosi sono maggiormente ripartiti tra tutte e tre le classi di previsione, motivo per cui si ritiene che i primi modelli descrivano meglio la classificazione dei campioni.
I motivi per i quali molti dei modelli adattati non forniscono una buona classificazione dei campioni tumorali sono verosimilmente due. Sicuramente i modelli di classificazione risentono della scarsa numerosità dei tessuti sani, sui quali sono stati adattati. Inoltre i campioni sani e i campioni tumorali sono, per natura, molto diversi tra loro, pertanto è possibile che, anche con una numerosità maggiore, molti modelli non sarebbero riusciti comunque a separare i campioni nelle diverse

classi, tendendo invece ad aggregarli assieme.
Nonostante queste difficoltà, le classificazioni fornite dai vari modelli sembrano suggerire complessivamente una maggior somiglianza dei tumori sierosi con il tessuto epiteliale dell'ovaio. Per i campioni di stadio avanzato questa somiglianza è ancora più frequente, rispetto ai campioni di stadio I.
I risultati qui ottenuti sono in linea con quanto osservato nella sezione precedente, nella quale avevo utilizzato un approccio basato sulla correlazione, sempre con l'obiettivo di individuare a quale tessuto sano (ovaio o tube) sono più simili i campioni di tumore sieroso analizzati in questa tesi. Entrambi i metodi suggeriscono che, con i dati a disposizione, i campioni sierosi sono in larga maggioranza più simili alle hose piuttosto che alle tube, con l'eccezione di qualche campione, perlopiù di stadio iniziale, che sembra invece più simile alle tube.

Selezione con cv

Random Forest

	Campioni da classificare				
	Muc	End	Cc	Sier	Avanz.
Previsioni					
sano colon	2	0	0	0	0
sano hose	4	10	4	15	57
sano rene	9	12	13	5	31
sano tube	1	1	0	3	0

Support vector machines

	Campioni da classificare				
	Muc	End	Cc	Sier	Avanz.
Previsioni					
sano colon	0	0	0	0	0
sano hose	7	1	2	0	0
sano rene	0	0	0	0	0
sano tube	9	22	15	23	88

PAM

	Campioni da classificare				
	Muc	End	Cc	Sier	Avanz.
Previsioni					
sano colon	7	0	0	0	0
sano hose	0	4	0	7	2
sano rene	7	17	17	13	86
sano tube	2	2	0	3	0

Dlda

	Campioni da classificare				
	Muc	End	Cc	Sier	Avanz.
Previsioni					
sano colon	5	0	0	0	0
sano hose	0	4	0	7	14
sano rene	9	17	17	13	74
sano tube	2	2	0	3	0

Slda

	Campioni da classificare				
	Muc	End	Cc	Sier	Avanz.
Previsioni					
sano colon	6	0	0	0	0
sano hose	0	2	0	4	23
sano rene	8	18	17	13	59
sano tube	2	3	0	6	6

Selezione con DEG

Random Forest

	Campioni da classificare				
	Muc	End	Cc	Sier	Avanz.
Previsioni					
sano colon	5	0	2	1	0
sano hose	3	13	5	19	73
sano rene	6	8	10	0	15
sano tube	2	2	0	3	0

Support vector machines

	Campioni da classificare				
	Muc	End	Cc	Sier	Avanz.
Previsioni					
sano colon	1	3	6	0	0
sano hose	0	1	0	0	0
sano rene	6	6	4	1	11
sano tube	10	13	6	22	77

PAM

	Campioni da classificare				
	Muc	End	Cc	Sier	Avanz.
Previsioni					
sano colon	6	0	1	0	0
sano hose	3	11	2	14	16
sano rene	5	10	14	6	71
sano tube	2	2	0	3	1

Dlda

	Campioni da classificare				
	Muc	End	Cc	Sier	Avanz.
Previsioni					
sano colon	6	2	3	1	0
sano hose	1	6	2	12	28
sano rene	7	12	12	4	57
sano tube	2	3	0	6	3

Slda

	Campioni da classificare				
	Muc	End	Cc	Sier	Avanz.
Previsioni					
sano colon	6	0	1	0	0
sano hose	0	2	0	5	35
sano rene	7	15	14	5	43
sano tube	3	6	2	13	10

Tabella 4.4: Tabelle di classificazione per tutti i campioni

Selezione con cv

Random Forest

	Campioni da classificare			
	Muc	End	Sier	Avanz.
Previsioni				
sano colon	7	1	0	0
sano hose	9	22	20	88
sano tube	0	0	3	0

Support vector machines

	Campioni da classificare			
	Muc	End	Sier	Avanz.
Previsioni				
sano colon	7	0	0	0
sano hose	0	0	0	0
sano tube	9	23	23	88

PAM

	Campioni da classificare			
	Muc	End	Sier	Avanz.
Previsioni				
sano colon	12	3	1	1
sano hose	2	18	19	84
sano tube	2	2	3	3

Dlda

	Campioni da classificare			
	Muc	End	Sier	Avanz.
Previsioni				
sano colon	10	1	1	0
sano hose	4	20	19	88
sano tube	2	2	3	0

Slda

	Campioni da classificare			
	Muc	End	Sier	Avanz.
Previsioni				
sano colon	9	0	0	0
sano hose	5	16	11	78
sano tube	2	7	12	10

Selezione con DEG

Random Forest

	Campioni da classificare			
	Muc	End	Sier	Avanz.
Previsioni				
sano colon	8	0	1	0
sano hose	6	21	19	88
sano tube	2	2	3	0

Support vector machines

	Campioni da classificare			
	Muc	End	Sier	Avanz.
Previsioni				
sano colon	10	4	0	0
sano hose	0	1	0	0
sano tube	6	18	23	88

PAM

	Campioni da classificare			
	Muc	End	Sier	Avanz.
Previsioni				
sano colon	10	3	1	0
sano hose	4	18	19	88
sano tube	2	2	3	0

Dlda

	Campioni da classificare			
	Muc	End	Sier	Avanz.
Previsioni				
sano colon	8	1	1	0
sano hose	6	18	17	85
sano tube	2	4	5	3

Slda

	Campioni da classificare			
	Muc	End	Sier	Avanz.
Previsioni				
sano colon	7	2	0	0
sano hose	4	11	8	72
sano tube	5	10	15	16

Tabella 4.5: Tabelle di classificazione (3 classi)

Capitolo 5

Analisi dei campioni tumorali

In questo capitolo mi concentro sull'analisi dei campioni di tumore, affrontando il secondo ed il terzo obiettivo di questa tesi. Nella prima sezione ho utilizzato dei modelli di classificazione, adattati ai diversi istotipi e agli stadi avanzati. Per ogni tipologia di modello ho selezionato quello con minor errore di previsione e per ciascuno ho estratto i geni più importanti nella previsione di ogni classe, identificando così dei marcatori di ogni condizione. Avendo osservato, attraverso le analisi svolte finora, molta eterogeneità nei dati, ho deciso di proseguire e concludere questo studio con dei metodi non supervisionati. Senza considerare le informazioni riguardanti l'appartenenza alle diverse classi, ho utilizzato dei metodi di raggruppamento basati esclusivamente sulla conoscenza delle variabili esplicative, ovvero, in questo caso, dei valori di espressione dei geni. Ho quindi analizzato i raggruppamenti emersi confrontandoli con le caratteristiche cliniche note delle pazienti, per lo stadio iniziale e per gli stadi avanzati. Per i casi più interessanti ho proseguito con le analisi di arricchimento e con il metodo Source Set, che fornisce una rappresentazione tramite grafi dei pathway biologici coinvolti, evidenziando la fonte primaria di disregolazione tra due condizioni.

5.1 Modelli di classificazione

In questa sezione utilizzo cinque diverse tipologie di modelli di classificazione per studiare il comportamento di ogni istotipo e degli stadi avanzati. I modelli utilizzati sono quelli introdotti nel capitolo 4, ovvero: *random forest*, *support vec-*

tor machines, PAM, analisi del discriminante lineare (con matrice di covarianza diagonale e con *shrinkage*). In tutti questi modelli ho utilizzato, come variabili esplicative, un sottoinsieme dei geni di partenza, dato dai 2000 geni con il maggior coefficiente di variazione, calcolato sui campioni malati.
Per valutare la bontà dei modelli ho utilizzato una procedura di convalida incrociata: a rotazione ho fissato il 75% dei dati per la stima dei modelli e il 25% rimanente per la verifica, ripetendo quindi la procedura quattro volte. Per le *support vector machines* ho adottato un'ulteriore procedura di convalida incrociata per la scelta dei parametri: all'interno del 75% dei dati fissato per la stima, ho utilizzato a rotazione 2/3 dei dati per la stima e 1/3 per il calcolo dell'errore di previsione. I modelli PAM forniscono in automatico l'errore di convalida incrociata al variare della soglia, pertanto ho utilizzato quest'indicazione per selezionare, di volta in volta, il modello con soglia più grande tra quelli ad errore minimo.
Per ogni modello ho calcolato una tabella di classificazione per ogni iterazione della convalida incrociata. Per riassumere l'informazione e ottenere, per ogni classe, un'unica indicazione sulla capacità predittiva del modello, ho creato una tabella riassuntiva che contiene la media e il range della proporzione di unità classificate correttamente nelle quattro iterazioni. Di seguito presento i vari modelli utilizzati specificando i parametri scelti.

Random forest: Ho utilizzato una foresta casuale con 1000 alberi e 44 variabili usate in ogni nodo ($\sqrt{2000}$). La soglia per l'attribuzione della classe "vincente" è proporzionale alla numerosità delle classi. La capacità predittiva di tale modello si riassume nella seguente tabella:

Random Forest

Classi	% media di corrett. classif.	range % di corrett. classif.
Muc	0.5	$[0.25, 0.75]$
End	0.608	$[0.5, 0.667]$
Cc	0.762	$[0.75, 0.8]$
Sier	0.817	$[0.6, 1]$
Stadi avanz.	0.977	$[0.955, 1]$

Support vector machines: Per le *support vector machines* ho fissato gamma al valore di *default* e pesi inversamente proporzionali alla numerosità di ogni classe.

Ho valutato la costante C e il tipo di nucleo tramite convalida incrociata. Ho scelto il modello con kernel lineare e costo $C = 1$. Nonostante il modello ad errore minimo risultasse essere quello con kernel sigmoide e costo pari a 10, la differenza di errore percentuale tra i due modelli è molto piccola (pari a 0.02). Per semplicità di interpretazione ho quindi scelto quello con kernel lineare. Di seguito la tabella corrispondente:

Support vector machines

Classi	% media di corrett. classif.	range % di corrett. classif.
Muc	0.5	$[0.25, 0.75]$
End	0.7	$[0.667, 0.833]$
Cc	0.762	$[0.75, 0.8]$
Sier	0.683	$[0.4, 1]$
Stadi avanz.	0.989	$[0.955, 1]$

PAM: Per il modello PAM ho utilizzato, ad ogni iterazione della convalida incrociata, una soglia leggermente diversa, in modo da utilizzare la soglia più grande che minimizzasse l'errore di convalida incrociata associato all'insieme di stima corrente. La numerosità di ogni classe, in proporzione al totale di osservazioni, viene usata di *default* come probabilità a priori di appartenere ad una classe. Si ottiene la seguente tabella riassuntiva:

PAM

Classi	% media di corrett. classif.	range % di corrett. classif.
Muc	0.5	$[0.25, 0.75]$
End	0.825	$[0.667, 1]$
Cc	0.825	$[0.75, 1]$
Sier	0.558	$[0.4, 0.833]$
Stadi avanz.	0.966	$[0.909, 1]$

Dlda e Slda: Le probabilità a priori vengono anche qui impostate di *default* alle proporzioni delle classi rispetto al totale di osservazioni. Per il metodo con *shrinkage* (Slda) utilizzo come matrice "target" una matrice diagonale costante (si veda par. 2.2.4). La classificazione fornisce i seguenti risultati:

Dlda

Classi	% media di corrett. classif.	range % di corrett. classif.
Muc	0.5	$[0.25, 0.75]$
End	0.817	$[0.6, 1]$
Cc	0.825	$[0.75, 1]$
Sier	0.692	$[0.6, 0.833]$
Stadi avanz.	0.977	$[0.909, 1]$

Slda

Classi	% media di corrett. classif.	range % di corrett. classif.
Muc	0.5	$[0.25, 0.75]$
End	0.833	$[0.667, 1]$
Cc	0.712	$[0.6, 0.75]$
Sier	0.725	$[0.4, 1]$
Stadi avanz.	1	–

In tutti i modelli proposti si può osservare che gli stadi avanzati sono sempre quelli meglio classificati, mentre i campioni mucinosi sempre quelli peggio classificati. Gli altri campioni sono classificati discretamente bene, e ogni modello tende a classificarne meglio uno piuttosto che l'altro.

5.1.1 Identificazione dei marcatori

Per l'identificazione dei geni più importanti nel predire ogni classe, ovvero dei geni marcatori di una certa condizione, ho utilizzato i modelli sopra descritti, ristimati su tutti i dati a disposizione. In particolare, per valutare l'importanza delle variabili ho utilizzato i modelli *random forest*, *support vector machines* e PAM. Per ciascuno ho adottato un metodo specifico, come descritto in 2.2.5. Ho selezionato, per ogni modello e per ogni classe, gli 80 geni più importanti per la previsione, e ho poi considerato, classe per classe, l'unione dei geni individuati dai tre modelli. Ho ottenuto così un totale di 185 marcatori del tumore mucinoso, 174 di quello endometrioide, 169 di quello clear cell, 184 del tumore sieroso di stadio

iniziale e infine 132 del tumore sieroso di stadi avanzati.
Nelle tabelle dalla 5.1 alla 5.5 ho elencato, classe per classe, due marcatori per ogni modello, quelli con maggior indice di importanza. In particolare ho riportato la descrizione di essi e la media di espressione nelle varie classi, con intervallo di confidenza bootstrap al 95%. I valori di espressione di tutti i marcatori individuati sono rappresentati nella *heatmap* in figura 5.1. Per molti geni si riesce ad osservare una diversa espressione dei marcatori nella classe considerata, rispetto alle altre; per altri si osserva, invece, una differenza tra la classe considerata ed una sola delle restanti: questo è in parte dovuto al fatto che, per le *support vector machines*, la valutazione dell'importanza dei geni viene fatta confrontando una coppia di classi alla volta. Sopra alla heatmap ho riportato l'informazione riguardo la recidiva di ogni paziente; non si osserva tuttavia una distinzione dei valori di espressione dei geni tra i soggetti che presentano recidiva e quelli che non la presentano.

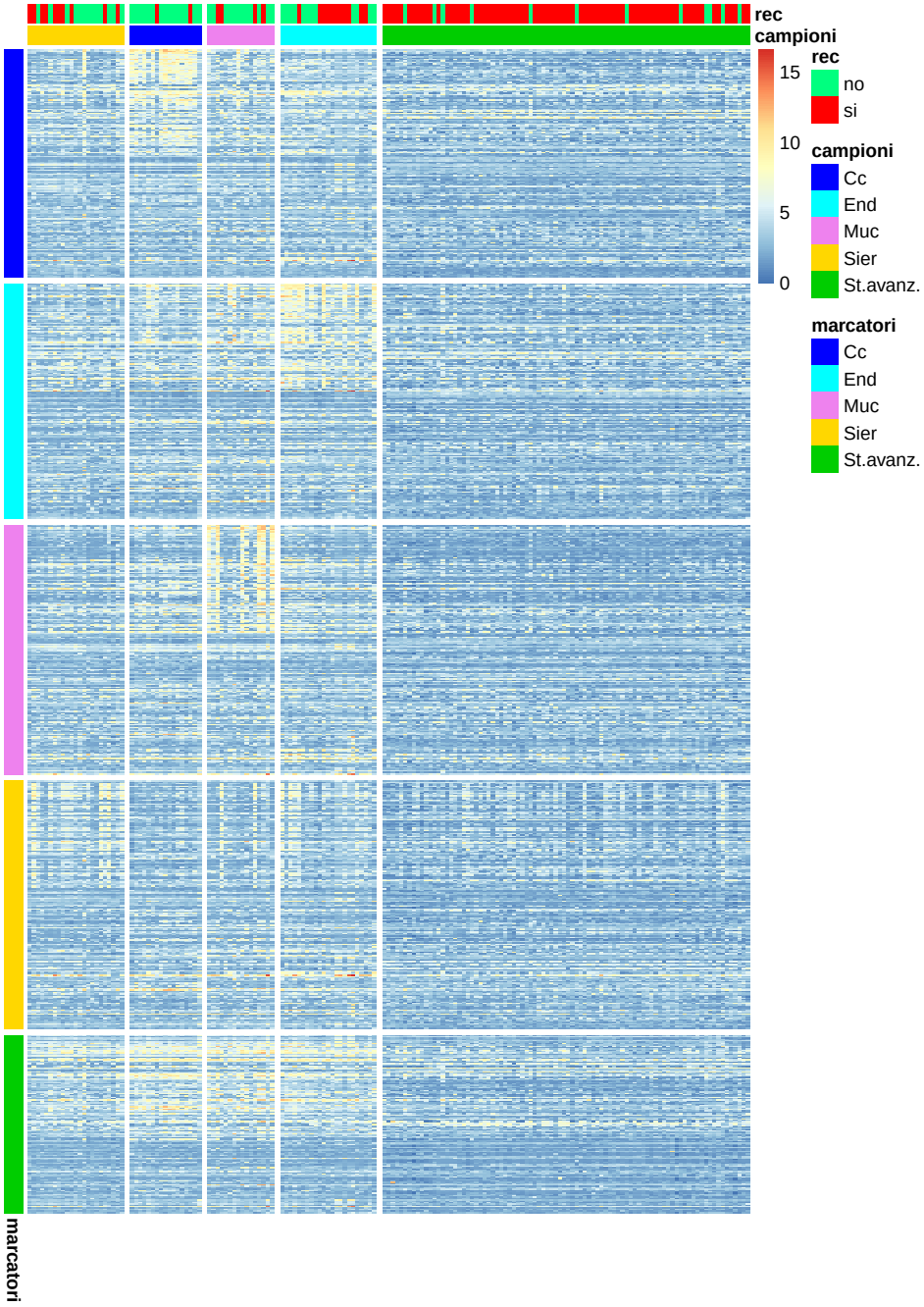


Figura 5.1: Heatmap dei marcatori

Clear cell

Modello	Gene	Descrizione
SVM	SLC22A7	solute carrier family 22 member 7
	OR2B2	olfactory receptor family 2 subfamily B member 2
RF	HAVCR1	hepatitis A virus cellular receptor 1
	FXYD2	FXYD domain containing ion transport regulator 2
PAM	FXYD2	FXYD domain containing ion transport regulator 2
	LEFTY1	left-right determination factor 1

	Cc	End	Muc	Sier	St. Avanz.
OR2B2	2.9	3.54	3.01	2.87	2.37
	[2.52,3.26]	[3.07,3.99]	[2.63,3.42]	[2.56,3.16]	[2.23,2.51]
LEFTY1	9.34	5.88	5.5	5.43	4.27
	[8.38,10.3]	[5.34,6.49]	[4.83,6.2]	[4.64,6.29]	[4.01,4.56]
HAVCR1	5.15	2.25	2.51	2.26	1.72
	[4.26,6.01]	[1.99,2.57]	[2.05,3.1]	[1.95,2.74]	[1.6,1.86]
SLC22A7	3.95	4.48	4.05	3.58	3.29
	[3.55,4.49]	[3.69,5.39]	[3.6,4.75]	[3.43,3.72]	[3.22,3.36]
FXYD2	6.14	3.26	3.67	3.25	3.12
	[5.42,6.87]	[2.92,3.61]	[3.21,4.29]	[2.87,3.76]	[3.01,3.24]

Tabella 5.1: Marcatori clear cell

Endometrioide

Modello	Gene	Descrizione
SVM	SERPINA9	serpin family A member 9
	XAGE3	X antigen family member 3
RF	CKLF	chemokine like factor
	ZNF43	zinc finger protein 43
PAM	SLC6A2	solute carrier family 6 member 2
	DLX5	distal-less homeobox 5

	Cc	End	Muc	Sier	St. Avanz.
XAGE3	5.25	3.15	3.14	3.44	3.9
	[4.35,6.12]	[2.85,3.44]	[2.68,3.75]	[3.09,3.89]	[3.68,4.13]
SERPINA9	1.93	2.34	2.17	2.4	1.97
	[1.66,2.3]	[2.04,2.66]	[1.93,2.46]	[2.08,2.73]	[1.84,2.1]
CKLF	2.18	2.35	2.25	2.37	4.53
	[1.86,2.5]	[2.06,2.66]	[2.03,2.48]	[2.15,2.61]	[4.43,4.65]
DLX5	7.81	7.83	6.45	4.47	3.93
	[7.3,8.38]	[7.01,8.59]	[5.72,7.27]	[3.6,5.45]	[3.66,4.24]
ZNF43	1.8	2.01	2.06	1.92	3.49
	[1.65,1.97]	[1.85,2.17]	[1.8,2.4]	[1.81,2.03]	[3.37,3.62]
SLC6A2	3.07	5.32	3.4	2.92	1.96
	[2.42,3.72]	[4.37,6.26]	[2.57,4.5]	[2.38,3.54]	[1.8,2.13]

Tabella 5.2: Marcatori endometrioide

Mucinoso

Modello	Gene	Descrizione
SVM	IGFL2	IGF like family member 2
	EDNRB	endothelin receptor type B
RF	CKLF	chemokine like factor
	OR1K1	olfactory receptor family 1 subfamily K member 1
PAM	PHGR1	proline, histidine and glycine rich 1
	REG4	regenerating family member 4

	Cc	End	Muc	Sier	St. Avanz.
IGFL2	3.61	4.18	4.15	2.58	4.62
	[2.86,4.39]	[3.49,4.91]	[3.26,5.04]	[2.21,2.97]	[4.19,5.03]
REG4	2.54	2.48	5.91	2.34	1.81
	[2.08,3.07]	[2.15,2.86]	[4.18,7.72]	[2.03,2.76]	[1.67,1.97]
CKLF	2.18	2.35	2.25	2.37	4.53
	[1.84,2.51]	[2.03,2.69]	[2.03,2.47]	[2.14,2.61]	[4.43,4.65]
EDNRB	2.84	3.13	2.89	2.69	3.02
	[2.61,3.05]	[2.54,4.05]	[2.54,3.25]	[2.43,2.98]	[2.88,3.15]
PHGR1	3.02	3.64	6.63	2.99	1.74
	[2.51,3.58]	[3.07,4.37]	[4.91,8.44]	[2.64,3.44]	[1.62,1.88]
OR1K1	4.36	4.49	4.69	4.05	2.16
	[4,4.79]	[3.99,5.04]	[4.32,5.11]	[3.78,4.31]	[2.02,2.3]

Tabella 5.3: Marcatori mucinoso

Sieroso

Modello	Gene	Descrizione
SVM	SERPINA9	serpin family A member 9
	XAGE3	X antigen family member 3
RF	CKLF	chemokine like factor
	C9orf84	chromosome 9 open reading frame 84
PAM	OR3A3	olfactory receptor family 3 subfamily A member 3
	CKLF	chemokine like factor

	Cc	End	Muc	Sier	St. Avanz.
OR3A3	4.43	4.04	4.36	4.3	2.09
	[4.23,4.64]	[3.68,4.44]	[4.1,4.6]	[4.11,4.48]	[1.95,2.22]
XAGE3	5.25	3.15	3.14	3.44	3.9
	[4.4,6.12]	[2.85,3.44]	[2.68,3.7]	[3.05,3.88]	[3.67,4.12]
SERPINA9	1.93	2.34	2.17	2.4	1.97
	[1.67,2.28]	[2.06,2.65]	[1.93,2.49]	[2.08,2.75]	[1.85,2.1]
CKLF	2.18	2.35	2.25	2.37	4.53
	[1.85,2.53]	[2.03,2.67]	[2.03,2.49]	[2.14,2.62]	[4.43,4.65]
C9orf84	1.92	2.77	2.23	1.86	3.36
	[1.71,2.16]	[2.26,3.41]	[1.98,2.55]	[1.73,1.97]	[3.28,3.45]

Tabella 5.4: Marcatori sieroso

St. avanzati

Modello	Gene	Descrizione
SVM	C9orf84	chromosome 9 open reading frame 84
	OR56A4	olfactory receptor family 56 subfamily A member 4
RF	CKLF	chemokine like factor
	ZBTB49	zinc finger and BTB domain containing 49
PAM	SIX3	SIX homeobox 3
	CKLF	chemokine like factor

	Cc	End	Muc	Sier	St. Avanz.
ZBTB49	2.11 [1.89,2.31]	2.32 [2.08,2.58]	2.25 [2.07,2.47]	2.22 [2.06,2.38]	3.92 [3.84,4]
CKLF	2.18 [1.87,2.51]	2.35 [2.05,2.66]	2.25 [2.03,2.48]	2.37 [2.13,2.61]	4.53 [4.44,4.64]
SIX3	5 [4.65,5.43]	4.95 [4.38,5.58]	5.35 [5.01,5.76]	4.59 [4.24,4.94]	2.15 [1.96,2.32]
C9orf84	1.92 [1.71,2.17]	2.77 [2.29,3.35]	2.23 [1.97,2.53]	1.86 [1.74,1.96]	3.36 [3.28,3.46]
OR56A4	2.64 [2.4,2.86]	2.48 [2.31,2.64]	2.62 [2.37,2.9]	2.96 [2.77,3.12]	1.76 [1.66,1.86]

Tabella 5.5: Marcatori st. avanzati

5.2 Apprendimento non supervisionato

Nella sezione precedente ho studiato le caratteristiche dei campioni in modo supervisionato, ovvero sfruttando l'informazione a disposizione riguardo l'istotipo e lo stadio del tumore. In questa sezione utilizzo invece dei metodi che si basano esclusivamente sulla conoscenza del livello di espressione dei geni, senza sapere a che tipologia appartiene ogni campione. Questo approccio consente di individuare dei raggruppamenti che poi confronterò tra loro, per esempio in termini di sopravvivenza delle pazienti, oppure cercando delle caratteristiche cliniche che siano significativamente diverse nei gruppi individuati.
Per analizzare i dati ho utilizzato un metodo di cluster non gerarchico, nello specifico l'algoritmo delle k-medie, e modelli mistura di gaussiane, implementati rispettivamente con le funzioni *kmeans* e *Mclust* dei pacchetti *stats* [24] e *mclust* [31].
Inizialmente ho applicato tali metodi su tutti i campioni a disposizione, utilizzando una selezione dei 2000 geni con maggior coefficiente di variazione. Successivamente ho provato ad applicare tali metodi ai soli campioni sierosi (di stadio iniziale e di stadi avanzati), ricalcolando su queste osservazioni i geni con maggior coefficiente di variazione.
La funzione *Mclust* effettua internamente un confronto per selezionare il modello con miglior numero di componenti di mistura e miglior struttura di covarianza. La scelta può essere fatta in diversi modi, io ho utilizzato il criterio BIC. La definizione delle possibili strutture per la matrice di covarianza è la seguente. Sia Σ_k la matrice di covarianza della componente di mistura k-esima, parametrizzata nel seguente modo: $\Sigma_k = \lambda_k \boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^T$, dove λ_k è uno scalare che controlla il volume dell'ellissoide, $\boldsymbol{A}_k$ è una matrice diagonale che ne specifica la forma (con $\det(\boldsymbol{A}_k) = 1$), mentre $\boldsymbol{D}_k$ è una matrice ortogonale che ne determina l'orientamento. In base alla struttura della matrice di covarianza, si ottengono modelli diversi; per il caso $p > n$ ci sono sei possibilità: $\Sigma_k = \lambda \boldsymbol{I}$, $\Sigma_k = \lambda_k \boldsymbol{I}$, $\Sigma_k = \lambda \boldsymbol{A}$, $\Sigma_k = \lambda_k \boldsymbol{A}$, $\Sigma_k = \lambda \boldsymbol{A}_k$, $\Sigma_k = \lambda_k \boldsymbol{A}_k$.
Per il metodo *kmeans* bisogna, invece, specificare il numero di gruppi desiderato.
Applicando il metodo *Mclust* a tutti i dati, il modello con BIC minore è quello con 11 gruppi e struttura di covarianza $\Sigma_k = \lambda_k \boldsymbol{A}$. Anche con l'algoritmo *kmeans* la scelta di 11 gruppi appare buona: ogni tipologia di campioni è in buona parte rappresentata da uno o più gruppi. I raggruppamenti ottenuti con questi due metodi

sono riportati nelle tabelle 5.6 e 5.7. Nella tabella 5.6 si osserva che i campioni sierosi di stadi avanzati sono divisi in tre gruppi. Tale divisione rimane valida anche se si diminuisce il numero di componenti della mistura, ovvero il numero di gruppi voluti, fino a 8 componenti. Per i modelli che prevedono invece 6 o 7 gruppi si ottiene una divisione in due parti dei campioni di stadi avanzati. Per gli istotipi di stadio iniziale vi è un gruppo che rappresenta ciascuno di essi, ed inoltre è presente una componente aggiuntiva che raccoglie insieme molte unità dei diversi istotipi. I tessuti sani di tube ed hose costituiscono un unico gruppo, mentre quelli di rene e colon formano dei gruppi a sè. Si osserva inoltre che 7 campioni sierosi di stadio iniziale vengono accorpati ai gruppi di stadi avanzati. Il raggruppamento ottenuto con il metodo *kmeans* è molto simile a quello appena descritto.
Gli aspetti che ho deciso di approfondire sono la suddivisione dei campioni di stadi avanzati in più gruppi, e la somiglianza di alcuni campioni sierosi di stadio iniziale agli stadi avanzati. Ho inoltre osservato come si dispongono i campioni imponendo la presenza di soli due o tre gruppi, sia per tutti i campioni, sia limitando l'analisi a quelli sierosi. La scelta fra due o tre gruppi è stata fatta in modo da avere almeno un gruppo che rappresentasse maggiormente gli stadi avanzati, e almeno uno per lo stadio iniziale (ed eventualmente i tessuti sani).
Con il metodo *kmeans* (tabella 5.9) ho utilizzato due gruppi: uno rappresenta gli stadi avanzati con l'aggiunta di alcuni campioni di stadio iniziale, perlopiù sierosi; l'altro è formato dai rimanenti campioni di stadio iniziale e dai tessuti sani. Per il metodo *Mclust* (tabella 5.8) si forma, invece, un gruppo in più, che contiene esclusivamente i tessuti sani di hose e tube. Per i raggruppamenti effettuati solo sui campioni sierosi (tabella 5.10), si ha una divisone in due gruppi con *Mclust*, mentre in tre gruppi con *kmeans*. Nel primo caso un gruppo rappresenta gli stadi avanzati, ai quali si aggiunge una buona porzione dei campioni di stadio iniziale, mentre il secondo gruppo contiene i rimanenti campioni di stadio iniziale. Con il secondo metodo invece, i campioni di stadi avanzati vengono divisi in due gruppi, e quelli di stadio iniziale sono sempre divisi circa a metà: una parte viene accorpata agli stadi avanzati, mentre una parte forma un gruppo a sè. Confrontando questi ultimi raggruppamenti, nei quali si è imposto un massimo di 3 gruppi, con le precedenti suddivisioni in 11 gruppi, si osserva che persiste la divisione dei campioni sierosi di stadio iniziale, dove alcuni formano un gruppo a sè, mentre altri vengono accorpati a quelli di stadi avanzati. Proprio per il numero limitato di gruppi non

è invece evidente la suddivisione in 3 gruppi degli stadi avanzati.

	Cc	End	Muc	Sier	St. Avanz.	Colon	Hose	Rene	Tube	Bord.
1	13		1	1						
2	1	10	1							1
3			7							1
4			2	7						1
5	3	11	5	8						1
6				5	21					
7		1		2	27					
8		1			40					
9							3		6	
10								4		
11						4				

Tabella 5.6: Mclust - 11 gruppi

	Cc	End	Muc	Sier	St. Avanz.	Colon	Hose	Rene	Tube	Bord.
1	14		1	1						
2	2	9	2							
3			7							1
4		2	2	7						2
5	1	10	3	7	1					1
6			1	5	20					
7		1		2	24					
8		1		1	43					
9						4				
10								4		
11							3		6	

Tabella 5.7: Kmeans - 11 gruppi

	Cc	End	Muc	Sier	St. Avanz.	Colon	Hose	Rene	Tube	Bord.
1	1	8	3	12	88					
2	16	15	13	11		4		4		4
3							3		6	

Tabella 5.8: Mclust - 3 gruppi

	Cc	End	Muc	Sier	St. Avanz.	Colon	Hose	Rene	Tube	Bord.
1	17	21	14	14	1	4	3	4	6	4
2		2	2	9	87					

Tabella 5.9: Kmeans - 2 gruppi

	St. I	St.Avanz.
1	13	87
2	10	1

(a) Mclust - 2 gruppi

	St. I	St.Avanz.
1	1	43
2	12	1
3	10	44

(b) Kmeans - 3 gruppi

Tabella 5.10: Solo sierosi

Le analisi proseguono con l'approfondimento dei raggruppamenti individuati. In particolare analizzerò se i sottogruppi individuati in base a caratteristiche di espressione genica riflettono anche una differenza in termini di sopravvivenza delle pazienti o di caratteristiche cliniche. Per i risultati più interessanti approfondirò le analisi cercando di individuare quali geni sono maggiormente coinvolti nel determinare le differenze tra gruppi, considerando anche le relazioni che intercorrono tra geni all'interno dei pathway biologici.

5.2.1 Caratterizzazione stadio iniziale

I metodi presentati in questa sezione forniscono possibili raggruppamenti dei campioni sierosi. Come primo approfondimento ho deciso di confrontare i campioni sierosi di stadio I più simili agli stadi avanzati con i rimanenti campioni sierosi di stadio iniziale; per fare ciò inizio con l'individuare una precisa suddivisione in due gruppi. Ho innanzitutto escluso i risultati dei raggruppamenti in 11 gruppi, in quanto la classe dei campioni sierosi di stadio iniziale risulta molto frammentata. I sottoinsiemi di campioni di stadio iniziale simili a quelli di stadio avanzato individuati dagli altri quattro raggruppamenti proposti sono abbastanza simili tra loro, nel senso che sono uno contenuto nell'altro. Tra questi ho scelto di utilizzare i due gruppi individuati dal metodo *Mclust* applicato ai soli campioni sierosi (Tabella 5.10 (a)). Sia indicato con "Gruppo 1" il sottoinsieme di 13 campioni raggruppati agli stadi avanzati, e con "Gruppo 2" i 10 campioni rimanenti. In figura 5.2 sono rappresentati i due gruppi appena definiti, nelle coordinate ottenute dallo scaling

multidimensionale, calcolato sulla selezione dei 2000 geni con maggior coefficiente di variazione. Soprattutto nei primi due grafici si osserva una maggior vicinanza dei campioni sierosi del Gruppo 1 ai campioni di stadi avanzati.

Per confrontare i due gruppi ed individuare eventuali differenze a livello clinico, ho selezionato le variabili ritenute di maggior interesse e, per ciascuna, ho valutato tramite il test esatto di Fisher se la distribuzione è significativamente diversa nei due gruppi. In particolare ho utilizzato le variabili: Grado, Stato, FIGO e Recidiva, che hanno le seguenti distribuzioni nei due gruppi:

	Grado		
	G1	G2	G3
Gruppo 1	1	2	10
Gruppo 2	7	5	11

	FIGO		
	IA	IB	IC
Gruppo 1	3	2	8
Gruppo 2	7	4	12

	Stato	
	DOD	NED
Gruppo 1	3	10
Gruppo 2	5	18

	Recidiva	
	si	no
Gruppo 1	7	6
Gruppo 2	9	14

Nessun test è risultato significativo, ad eccezione del test unilaterale che confronta il grado G3 rispetto ai rimanenti, con l'ipotesi che nel Gruppo 1 il grado G3 è più frequente. Tale test è marginalmente significativo, con p-value associato pari a 0.087.

Con numerosità così basse è difficile rilevare differenze significative. In tali condizioni, e con le informazioni cliniche a disposizione, le caratteristiche a livello genico che hanno determinato la divisione in Gruppo 1 e Gruppo 2 non sembrano trovare riscontro in una diversa caratterizzazione a livello clinico dei due gruppi.

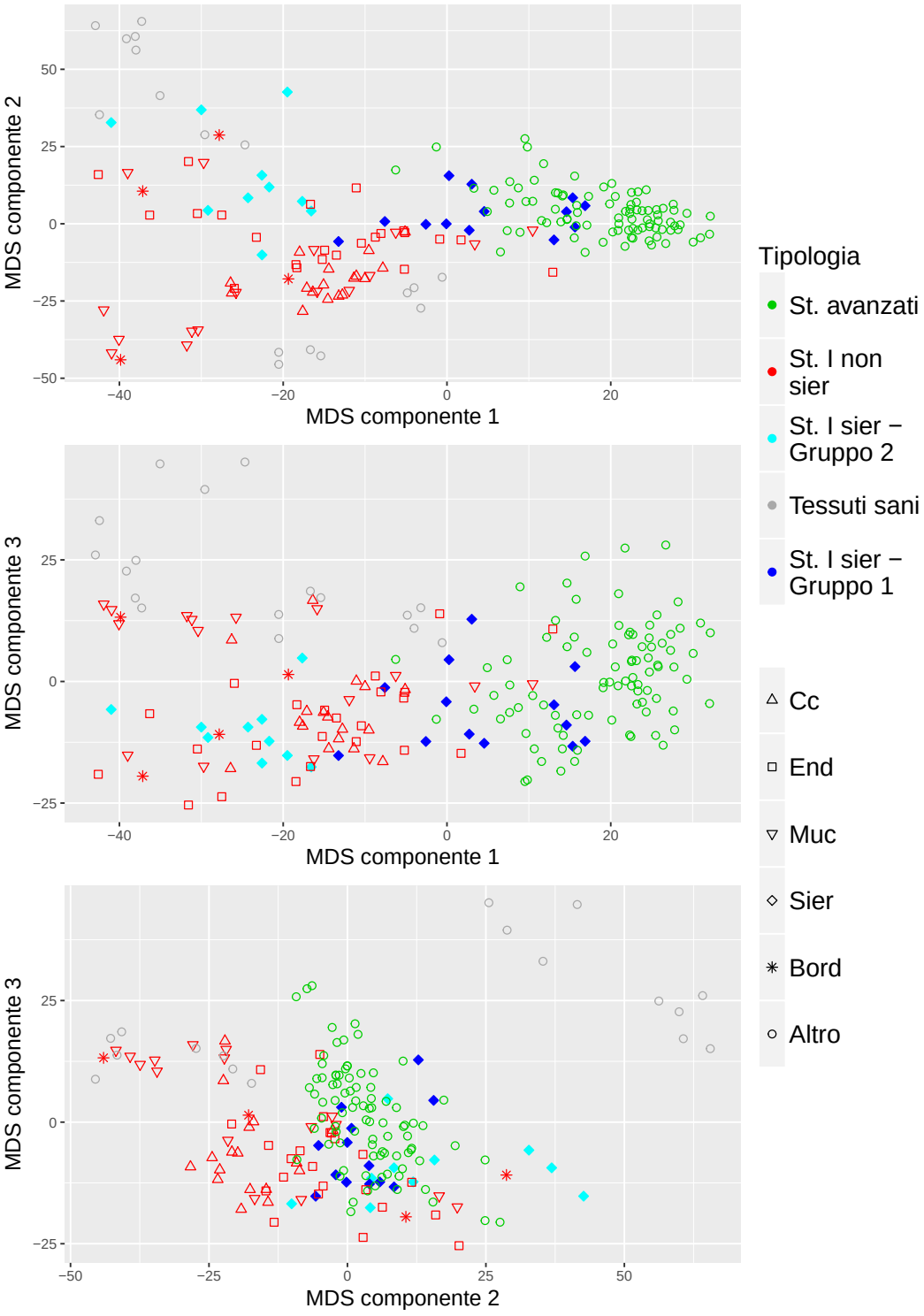


Figura 5.2: Scaling multidimensionale

5.2.2 Caratterizzazione stadi avanzati

Analogamente a quanto fatto per gli stadi iniziali, cerco una caratterizzazione per i diversi gruppi emersi all'interno degli stadi avanzati. Per definire la suddivisione dei campioni ho considerato i raggruppamenti ottenuti con i metodi *mclust* e *kmeans* con 11 gruppi. Confrontandoli tra loro, si osserva che la divisione dei campioni con i due metodi è molto simile, pertanto ho proseguito le analisi solo con quella fornita da *mclust*, che consiste in tre gruppi di numerosità 21, 27 e 40. Ho considerato le caratteristiche cliniche di maggior interesse e ho valutato attraverso il test esatto di Fisher se la distribuzione di queste risulta significativamente diversa nei gruppi individuati. Il test con maggior significatività è quello relativo allo stato di sensibilità o resistenza alla terapia al platino. Ho escluso i campioni delle pazienti parzialmente sensibili, in quanto si tratta di una condizione intermedia, che potrebbe quindi presentare aspetti in comune sia con i campioni sensibili che resistenti. Si ha la seguente distribuzione:

	Sensibili	Resistenti	Tot
Gruppo 1	11	13	24
Gruppo 2	14	4	18
Gruppo 3	9	19	28

Il p-value associato al test di Fisher è 0.011. In particolare il secondo gruppo sembra quello maggiormente diverso dagli altri, infatti si osserva un incremento nella significatività accorpando il primo ed il terzo gruppo, ottenendo un p-value pari a 0.005. Proseguo quindi le analisi con due soli gruppi, così ripartiti:

	Sensibili	Resistenti	Tot
Gruppo 1	20	32	52
Gruppo 2	14	4	18

I soggetti resistenti sono concentrati quasi totalmente nel primo gruppo, mentre i soggetti sensibili sono divisi tra i due gruppi. Ho deciso di approfondire la divisione dei soggetti sensibili per cercare di individuare quali caratteristiche geniche li rendono diversi e se ci sono alcune analogie con ciò che discrimina la divisione clinica in sensibili e resistenti.
Innanzitutto ho confrontato nuovamente le distribuzioni di alcune variabili di maggior interesse tra i soggetti sensibili del Gruppo 1 e del Gruppo 2. Di seguito, con

Gruppo 1 e Gruppo 2 si farà riferimento ai soli soggetti sensibili dei due gruppi. Si ottiene un risultato significativo per il test di Fisher sulla variabile "FIGO", con p-value 0.05. La distribuzione per questa variabile è la seguente:

	IIIA	IIIB	IIIC	IV	Tot
Gruppo 1	0	0	11	9	20
Gruppo 2	2	1	9	2	14

Il Gruppo 1 comprende stadi più avanzati rispetto al Gruppo 2, infatti nel primo gruppo si ha circa un egual numero di campioni negli stadi IIIC e IV, mentre nel secondo gruppo la classe più frequente è IIIC, con qualche osservazione nello stadio IV e negli stadi inferiori (IIIA e IIIB).
Ho inoltre svolto un'analisi di sopravvivenza, che tuttavia non ha evidenziato differenze significative tra i due gruppi.

Analisi di arricchimento e Source Set

Dopo aver cercato una caratterizzazione a livello clinico, ho proseguito cercando di caratterizzare i due gruppi a livello genico. Ho quindi individuato i geni differenzialmente espressi, sui quali ho svolto un'analisi di arricchimento e infine ho applicato il metodo Source Set. Ho svolto queste analisi anche per i due gruppi di pazienti sensibili e pazienti resistenti, individuati dall'annotazione clinica, per confrontare i risultati. Attraverso il test bayesiano empirico ho individuato, tra tutti i geni disponibili, quelli differenzialmente espressi, con una significatività del 5%. Per il confronto tra campioni sensibili e resistenti sono emersi 40 geni differenzialmente espressi, mentre per il confronto tra il Gruppo 2 e il Gruppo 1 sono emersi 25 geni.
Per le analisi di arricchimento ho utilizzato le funzioni *enrichGO* ed *enrichKEGG* del pacchetto *clusterProfiler* [40], ed il database di annotazione *org.Hs.eg.db* [7]. La funzione *enrichGO* ricerca i termini di Gene Ontology (GO) [35] associati alla lista di geni differenzialmente espressi, mentre la funzione *enrichKEGG* cerca pathway biologici dal database KEGG [16]. I termini di Gene Ontology sono raggruppati in tre grandi categorie: funzioni molecolari (MF), componenti cellulari (CC) e processi biologici (BP).

Gene Ontology

ID	Categoria	Descrizione	p-value corretto
GO:0051145	BP	smooth muscle cell differentiation	0.06780027
GO:0000922	CC	spindle pole	0.07188223

Pathway KEGG

ID	Descrizione	p-value corretto
hsa05132	Salmonella infection	0.01614012

Tabella 5.11: Gruppo 1 vs Gruppo 2

Con una significatività del 5% sono risultati 32 termini di Gene Ontology e 3 pathway di KEGG per il confronto sensibili - resistenti, mentre nessun termine di Gene Ontology e un pathway per il confronto Gruppo 1 - Gruppo 2. Riporto qui i primi risultati (fino a cinque) per ogni categoria di GO e per i pathway, ammettendo un p-value corretto fino al 10%.

Nell'individuazione dei geni differenzialmente espressi e nelle analisi di arricchimento si trovano più risultati nel confronto tra sensibili e resistenti, piuttosto che nel confronto dei due gruppi individuati con i metodi non supervisionati. Con questo tipo di analisi sembrano quindi maggiormente diversi, da un punto di vista genico, i primi due gruppi, ovvero i soggetti sensibili e i soggetti resistenti.

Ho proseguito utilizzando il metodo *Source Set*, applicato a tutti i geni a disposizione, sui pathway del database KEGG. Oltre al pacchetto *SourceSet* ho utilizzato il pacchetto *graphite* [28] per acquisire i pathway nel formato richiesto. Per visualizzare i risultati in modo più compatto e facilitarne l'interpretazione ho riassunto l'informazione in un unico grafico, ottenuto con la funzione *sourceUnionCytoscape*, che fornisce l'unione grafica indotta dai source set della collezione di grafi di partenza.

Ogni nodo del grafo rappresenta un gene, e si ha che colore e dimensione del nodo descrivono le principali caratteristiche emerse dall'analisi: la dimensione del nodo cresce al crescere del numero di grafi di partenza nei quali il gene appare nel relativo source set; il colore del nodo si scurisce all'aumentare della percentuale di grafi di partenza tali che il gene appartiene al loro source set, rispetto al totale di grafi.

A tale notazione ho aggiunto, sul bordo di ogni nodo, informazioni riguardanti

Gene Ontology

ID	Categoria	Descrizione	p-value corretto
GO:0006334	BP	nucleosome assembly	4.023253e-04
GO:0031497	BP	chromatin assembly	4.023253e-04
GO:0034728	BP	nucleosome organization	4.051773e-04
GO:0006333	BP	chromatin assembly or disassembly	4.742865e-04
GO:0006323	BP	DNA packaging	6.405799e-04
GO:0000786	CC	nucleosome	4.301112e-09
GO:0044815	CC	DNA packaging complex	4.301112e-09
GO:0032993	CC	protein-DNA complex	2.684216e-07
GO:0046982	MF	protein heterodimerization activity	1.012281e-03
GO:0045236	MF	CXCR chemokine receptor binding	3.254909e-02
GO:0001664	MF	G-protein coupled receptor binding	9.692411e-02

Pathway KEGG

ID	Descrizione	p-value corretto
hsa05322	Systemic lupus erythematosus	7.464835e-06
hsa05034	Alcoholism	2.976823e-05
hsa05203	Viral carcinogenesis	6.861499e-04
hsa04620	Toll-like receptor signaling pathway	9.210056e-02

Tabella 5.12: Sensibili vs Resistenti

l'analisi di differenziale espressione sopra descritta. Al bordo più spesso corrispondono i geni significativamente differenzialmente espressi (p-value corretto < 0.1), mentre al diminuire dello spessore si hanno geni non significativi. Il bordo verde rappresenta geni up-regolati nei soggetti sensibili (o del Gruppo 2), rispetto ai resistenti (o Gruppo 1), mentre il bordo rosso rappresenta geni down-regolati. Quest'ultima distinzione è significativa per i geni differenzialmente espressi, mentre per i rimanenti è solo indicativa, infatti, per questi ultimi, le intensità di espressione (e quindi l'indicazione up o down regolati) non sono risultate significativamente diverse nei gruppi confrontati.

Nelle figure 5.3 e 5.4 sono riportati i grafi così ottenuti, visualizzati attraverso il software *Cytoscape* [32]. Si osserva che, contrariamente ai risultati precedenti, il confronto tra Gruppo 1 e Gruppo 2 dà luogo ad un grafo più esteso rispetto all'altro. Inizio con la discussione dei risultati per la divisione tra soggetti sensibili e resistenti, in riferimento alla figura 5.3, presentando gli aspetti ritenuti più inte-

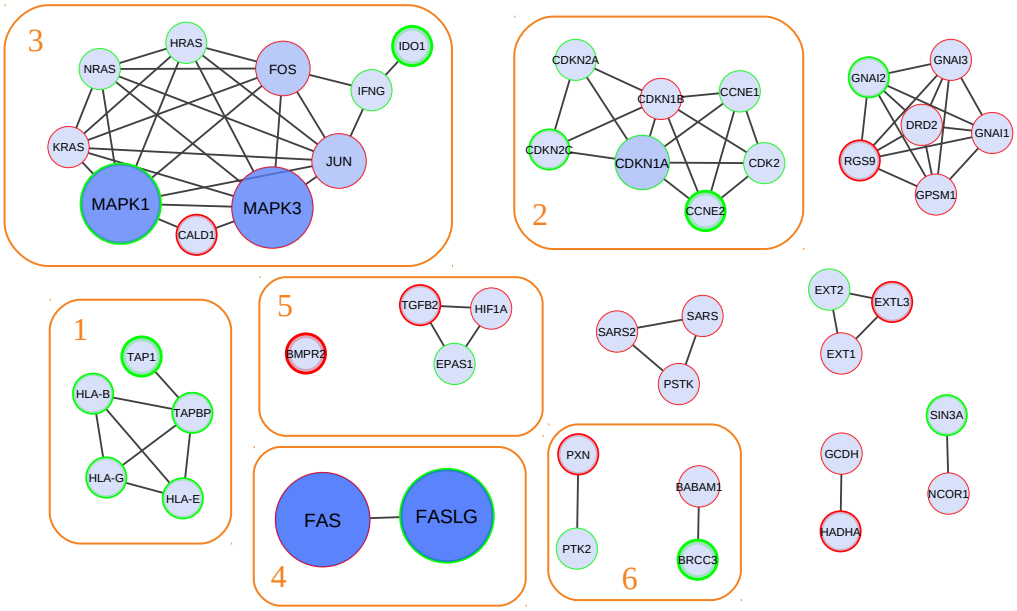


Figura 5.3: Source Set - Sensibili vs Resistenti

ressanti.

Il grafo 1 contiene geni che sono coinvolti nel processo di risposta immunitaria della cellula. In particolare partecipano al trasporto degli antigeni e alla loro presentazione ai linfociti T, i quali possono eventualmente innescare una risposta immunitaria. Tali geni risultano tutti up-regolati nei soggetti sensibili, con una significatività inferiore al 10% per il gene TAP1. Ciò è coerente col fatto che nei soggetti sensibili alla terapia si osserva una risposta immunitaria, contrariamente a quanto avviene per i soggetti resistenti.

Il grafo 2 contiene geni coinvolti nel ciclo cellulare. Le chinasi ciclina dipendente (CDK) sono una classe di proteine che partecipa alla regolazione del ciclo cellulare, mentre le cicline (che includono le proteine CCNE) agiscono come regolatori delle chinasi. Tutte queste proteine risultano solitamente sovraespresse in situazioni tumorali, nelle quali le cellule tumorali si riproducono molto in fretta.

La famiglia di chinasi MAP (MAPK) è anch'essa coinvolta a monte del ciclo cellulare. Tale famiglia di proteine partecipa ad un'ampia varietà di processi cellulari, quali proliferazione, differenziazione, regolazione della trascrizione e sviluppo. Queste proteine si ritrovano nel grafo 3 assieme ad altri geni che sono

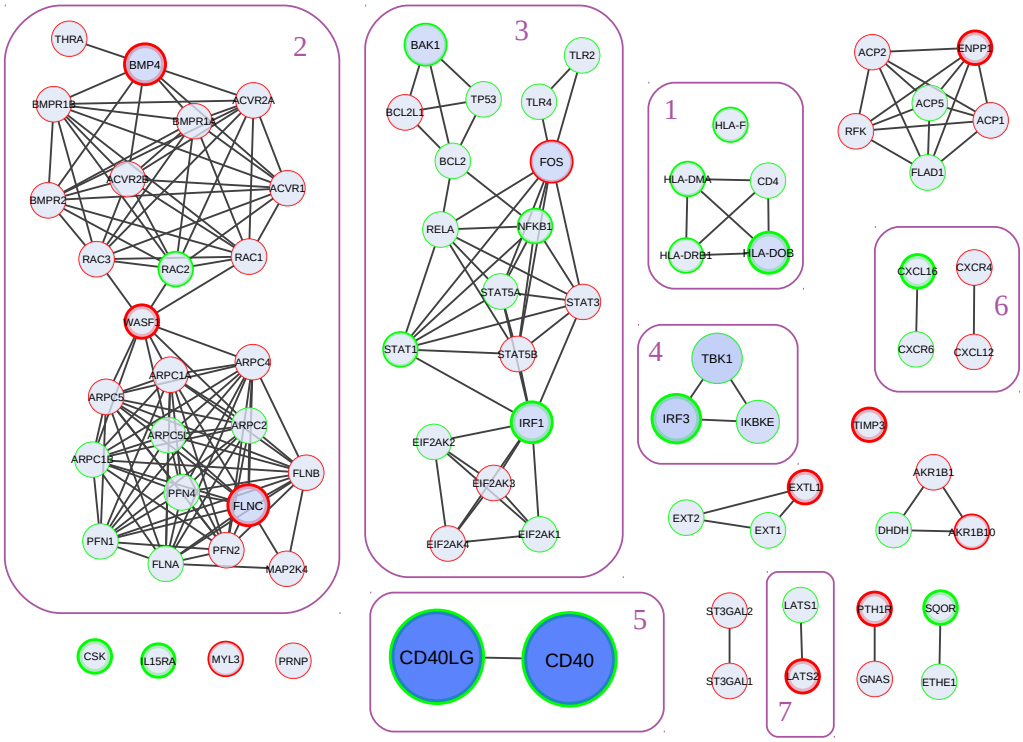


Figura 5.4: Source Set - Gruppo 2 vs Gruppo 1

proto-oncogeni (HRAS, KRAS, FOS, JUN), ovvero geni normali che, a causa di mutazioni o di un aumento di espressione, possono diventare oncogeni, cioè contribuire alla possibilità che lo sviluppo di una cellula si indirizzi verso una situazione tumorale.

La proteina FAS, assieme al suo ligando FASLG, è coinvolta nel processo di apoptosi della cellula (grafo 4). L'interazione del recettore FAS con il suo ligando permette, infatti, la formazione di un complesso che può scatenare una cascata di segnali portando alla morte cellulare.

Nel quinto riquadro si trovano i geni TGFB2 e BMPR2, che fanno entrambi parte della famiglia di TGFbeta. Tra i ruoli svolti da tale famiglia di proteine vi è il controllo della proliferazione, differenziazione ed altre funzioni. In condizioni normali tali proteine sono in grado di contribuire alla soppressione del tumore, attraverso attività di citostasi, differenziazione ed apoptosi. Tuttavia, ad un certo punto, le cellule tumorali perdono le risposte di TGFbeta in grado di favorire la soppressione

del tumore, ed anzi le cellule tumorali possono usare le TGFbeta a loro vantaggio per favorire l'evasione dal sistema immunitario, la produzione di fattori di crescita, la differenziazione in fenotipi invasivi e la diffusione di metastasi. Nel gruppo di pazienti sensibili tali geni risultano down-regolati, ad indicare forse il contenimento delle attività volte alla diffusione del tumore.
Infine, nel riquadro 6, si individuano altri geni associati a diverse situazioni tumorali: la disregolazione del gene PXN è coinvolta in diversi tumori, tra i quali quello al seno; il gene PTK2 si è osservato essere sovra espresso in tumori metastatici al seno e al colon; infine i geni BRCCB e BABAM1 sono associati ai geni BRCA1/2, i quali hanno un ruolo importante nel tumore al seno.
Nel confronto tra Gruppo 1 e Gruppo 2 (figura 5.4) si ritrovano molti fattori in comune con quanto appena descritto; spesso risultano significativi geni diversi, ma coinvolti nei processi biologici già incontrati.
Nel grafo 1 si ritrova la famiglia di geni HLA, con prevalenza di geni appartenenti alla classe II.
Nel grafo 2 si trovano diversi geni BMP che appartengono alla famiglia dei TGF-beta. La maggior parte dei geni nella parte inferiore del grafo, dai geni RAC a scendere, è coinvolta nei processi di formazione e regolazione del citoscheletro di actina. In diversi tumori si sono osservati alti livelli di alcune di queste proteine.
Nel terzo grafo si osservano molti geni coinvolti nei processi tumorali. Le famiglie STAT ed EIF2AK sono coinvolte, rispettivamente, nella trascrizione e traduzione dei trascritti, che portano alla sintesi delle proteine. Il gene FOS è considerato un proto-oncogeno e ha una funzione centrale nella proliferazione e differenziazione delle cellule, come anche nella trasformazione delle stesse e nella progressione tumorale. La famiglia STAT comprende fattori di trascrizione che mediano diversi processi di immunità, proliferazione, differenziazione e apoptosi della cellula. Nel processo di apoptosi giocano un ruolo fondamentale anche i geni BAK e TP53, e in particolare quest'ultimo ricopre la funzione di soppressore tumorale. I geni IRF (riquadri 3 e 4) hanno anch'essi un ruolo nella risposta immunitaria della cellula, nel regolare l'apoptosi e la soppressione dei tumori.
I geni CD40 e CD40LG (grafo 5) sono coinvolti nel ciclo cellulare. Il recettore codificato fa parte della famiglia TNF (recettore del fattore di necrosi tumorale) e contribuisce a determinare risposte cellulari di tipo immunitario e infiammatorio.
Le chemochine CXC (riquadro 6) sono una famiglia di proteine che svolgono fun-

zioni nella regolazione dell'angiogenesi, ovvero moltiplicazione dei vasi sanguigni. Essendo l'angiogenesi una delle caratteristiche fondamentali nello sviluppo di un tumore (inserita tra gli *Hallmarks of Cancer*), questa famiglia di geni ricopre un ruolo importante nella diffusione dei tumori.
Infine, nel riquadro 7, si trovano i geni LATS, che fanno parte di una famiglia di geni soppressori di tumore.
Per concludere l'analisi ho fatto l'intersezione tra i grafi dei due confronti descritti e l'insieme risultante è costituito da quattro geni. I due più interessanti sono il gene BMPR2 e FOS. Come già osservato, ci sono molti altri elementi in comune tra i due confronti, che coinvolgono geni diversi, appartenenti però alla stessa famiglia oppure coinvolti nello stesso processo biologico. In entrambi i confronti sono emersi dall'analisi geni implicati nel ciclo cellulare e nella risposta immunitaria della cellula. Nei soggetti sensibili e, parallelamente, nel Gruppo 2 risultano up-regolati i geni responsabili di una risposta immunitaria. Nonostante i soggetti dei gruppi 1 e 2 siano tutti sensibili, nel Gruppo 2 la risposta immunitaria sembra maggiormente accentuata rispetto al Gruppo 1. Gli altri aspetti sono di più difficile interpretazione e richiedono un'approfondita conoscenza dei processi biologici sottostanti.

Capitolo 6

Conclusioni

L'obiettivo di questa tesi è analizzare diversi aspetti del tumore all'ovaio, per cercare di comprenderne alcune caratteristiche. Il tumore all'ovaio è una malattia molto eterogenea e, anche per questo motivo, molti meccanismi non sono ancora chiari al giorno d'oggi e si cerca di approfondirne sempre di più lo studio per riuscire a migliorarne la prognosi.
Ripercorrendo i tre grandi quesiti che ho deciso di affrontare in questa tesi, espongo, in questo capitolo, le principali e più interessanti conclusioni raggiunte, così come le questioni rimaste aperte, che necessitano di ulteriori studi e approfondimenti.

Primo quesito. Con il primo quesito ho cercato di capire, attraverso due diversi approcci, a quale tessuto sano, tube oppure hose, sono più simili i campioni di tumore ovarico sieroso, per individuare un probabile sito di origine di questo tipo tumore.
Con il primo approccio, di correlazione, ho individuato che più del 70% dei campioni di tumore ovarico sieroso di stadio iniziale assomigliano maggiormente ai tessuti delle hose, mentre tale percentuale sale al 95% per gli stadi avanzati. Il secondo approccio, basato su modelli di classificazione, è quello che ha risentito maggiormente della scarsa numerosità dei tessuti sani. Molti modelli sono infatti risultati instabili al variare dei parametri di regolazione, e sbilanciati verso una classe. Per limitare tale effetto ho rimosso dai modelli i tessuti del rene e l'istotipo clear cell a loro collegato, in quanto risultavano i più problematici. Selezionati i modelli

migliori, che mostravano cioè una buona stabilità e una buona classificazione dei rimanenti istotipi (endometrioide e mucinoso), ho osservato, complessivamente, una classificazione dei campioni sierosi prevalentemente come tessuto delle hose, piuttosto che delle tube. Tali risultati sono concordi con quanto ottenuto con il primo approccio, anche se, con quest'ultimo, la ripartizione tra hose e tube è molto più bilanciata.
Il sito di origine del tumore ovarico sieroso è argomento di discussione al giorno d'oggi. Un'ipotesi è che esso origini, nella maggior parte dei casi, dalle tube di Falloppio, mentre in passato si credeva originasse dal tessuto dell'ovaio. Sicuramente è una questione aperta, che necessita di molti altri studi per essere approfondita.

Secondo quesito. Il secondo e terzo quesito si concentrano sull'analisi dei soli campioni tumorali.
Il secondo obiettivo è individuare dei marcatori di ogni condizione tumorale: i diversi istotipi di stadio iniziale e gli stadi avanzati sierosi. Per fare ciò ho utilizzato dei modelli di classificazione, scegliendo quelli ad errore minore, e ho estratto per ogni modello le variabili più importanti per predire ogni classe. I marcatori individuati distinguono discretamente bene la classe considerata rispetto alle rimanenti, come si è potuto vedere da una *heatmap* costruita sull'insieme dei marcatori.

Terzo quesito. Il terzo quesito si sviluppa a partire da un'analisi dei dati non supervisionata. Ho studiato i raggruppamenti dei campioni così emersi, confrontandone sia le caratteristiche cliniche che quelle geniche.
Il risultato più interessante riguarda gli stadi avanzati: sono stati individuati due gruppi che presentano caratteristiche cliniche significativamente diverse. In particolare ho approfondito lo studio su un sottoinsieme di questi campioni, sensibili alla terapia al platino, confrontando i risultati delle analisi svolte sui due gruppi da me individuati (Gruppo 1 e Gruppo 2) con i risultati ottenuti sulla divisione dei soggetti in sensibili e resistenti. Utilizzando il metodo *SourceSet* sono emersi diversi elementi in comune da un punto di vista genico, mentre altri processi biologici sembrano caratterizzare maggiormente un confronto piuttosto che l'altro. Nel confronto tra il Gruppo 1 ed il Gruppo 2 sono risultati coinvolti più geni rispetto all'altro confronto, e molti di essi sono geni tipicamente coinvolti nelle situazioni tumorali.

In questa tesi ho affrontato lo studio di diverse questioni riguardanti il tumore ovarico. Ho cercato di affiancare ai risultati statistici, dove possibile, anche una prima interpretazione a livello biologico dei risultati.
Come proseguimento del mio lavoro ritengo interessante un approfondimento della relazione tra tessuti sani e tessuti tumorali, con numerosità più elevate.

Bibliografia

[1] A. Azzalini and B. Scarpa. *Data analysis and data mining: An introduction.* Oxford University Press, 2012.

[2] K. V. Ballman, D. E. Grill, A. L. Oberg, and T. M. Therneau. Faster cyclic loess: normalizing rna arrays via linear models. *Bioinformatics*, 20(16):2778–2786, 2004.

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, 57(1):289–300, 1995.

[4] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. Go::termfinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004.

[5] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, 573(1):83 – 92, 2004.

[6] E. Calura, L. Paracchini, R. Fruscio, A. DiFeo, A. Ravaggi, J. Peronne, P. Martini, G. Sales, L. Beltrame, E. Bignotti, G. Tognon, R. Milani, L. Clivio, T. Dell'Anna, G. Cattoretti, D. Katsaros, E. Sartori, C. Mangioni, L. Ardighieri, M. D'Incalci, S. Marchini, and C. Romualdi. A prognostic regulatory pathway in stage i epithelial ovarian cancer: new hints for the poor prognosis assessment. *Annals of Oncology*, 27(8):1511–1519, 2016.

[7] M. Carlson. *org.Hs.eg.db: Genome wide annotation for Human*, 2018. R package version 3.6.0.

[8] Y. Chang and C. Lin. Feature ranking using linear svm. In I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J. Pellet, P. Spirtes, and A. Statnikov, editors, *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*, volume 3 of *Proceedings of Machine Learning Research*, pages 53–64, Hong Kong, 03–04 Jun 2008. PMLR.

[9] G. Chu, B. Narasimham, R. Tibshirani, and V. G. Tusher. *SAM "Significance Analysis of Microarrays" Users guide and technical document*, 2005.

[10] V. Djordjilović and M. Chiogna. Searching for a source of difference: a graphical model approach. Working Paper Series 4, Università degli studi di Padova, 2017.

[11] J. Ducie, F. Dao, M. Considine, N. Olvera, P. A. Shaw, R. J. Kurman, I. Shih, R. A. Soslow, L. Cope, and D. A. Levine. Molecular analysis of high-grade serous ovarian carcinoma with and without associated serous tubal intra-epithelial carcinoma. *Nature Communications*, 8, 2017.

[12] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[13] R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, 2005.

[14] T. Hastie, R. Tibshirani, and J. Fridman. *The Elements of Statistical Learning: Data Mining,Inference, and Prediction*. Springer, 2009.

[15] T. Hastie, R. Tibshirani, B. Narasimhan, and G. Chu. *pamr: Pam: prediction analysis for microarrays*, 2014. R package version 1.55.

[16] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2017.

[17] M. Kuhn. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

[18] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004.

[19] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[20] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017. R package version 1.6-8.

[21] E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

[22] G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger. *The Analysis of Gene Expression Data. Methods and Software.* Springer, 2003.

[23] A. Pedro Duarte Silva. *HiDimDA: High Dimensional Discriminant Analysis*, 2015. R package version 0.2-4.

[24] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

[25] B. M. Reid, J. B. Permuth, and T. A. Sellers. Epidemiology of ovarian cancer: a review. *Cancer Biol. Med.*, 14(1):9–32, 2017.

[26] W. Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research.* Northwestern University, Evanston, Illinois, 2018. R package version 1.8.4.

[27] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[28] G. Sales, E. Calura, and C. Romualdi. *graphite: GRAPH Interaction from pathway Topological Environment*, 2018. R package version 1.26.1.

[29] E. Salviato. *Computational methods for the discovery of molecular signatures from Omics Data.* PhD thesis, Università degli studi di Padova, 2018.

[30] E. Salviato, V. Djordjilovic, C. Romualdi, and M. Chiogna. *SourceSet: A Graphical Model Approach to Identify Primary Genes in Perturbed Biological Pathways*, 2018. R package version 0.1.1.

[31] L. Scrucca, M. Fop, T. M. Brendan, and A. E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233, 2017.

[32] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[33] G. K. Smith. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 2004.

[34] G. K. Smyth and T. P. Speed. Normalization of cdna microarray data. *Methods*, 31:265 – 273, 2003.

[35] The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338, 2017.

[36] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, 99:6567–6572, 2002.

[37] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98:5116–5121, 2001.

[38] Z. Wang and X. Xue. Multi-class support vector machine. In Y. Ma and G. Guo, editors, *Support Vector Machines Applications*, chapter 2, pages 23–48. Springer, Cham, 2014.

[39] J. W. Wragg, J. P. Finnity, J. A. Anderson, H. J.M. Ferguson, E. Porfiri, R. I. Bhatt, P. G. Murray, V. L. Heath, and R. Bicknell. Mcam and lama4

are highly enriched in tumor blood vessels of renal cell carcinoma and predict patient outcome. *Cancer Research*, 76(8):2314–2326, 2016.

[40] G. Yu, L. Wang, Y. Han, and Q. He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 2012.