Università degli Studi di Padova
Corso di Laurea in Statistica e Tecnologie Informatiche

UNO STUDIO SULLA DIFFERENZIAZIONE GENETICA FRA LINFOMA A GRANDI CELLULE B E LINFOMA FOLLICOLARE

Relatore: Ch.ma Prof.ssa Ventura

Dipartimento di Scienze Statistiche

Laureando: Umberto Simola

ANNO ACCADEMICO 2011/2012

Alla mia famiglia

Indice

INTRODUZIONE

Lo scopo di questo lavoro è analizzare tre tipologie di pazienti (affetti dal linfoma diffuso a grandi cellule B, affetti dal linfoma follicolare, sani) attraverso l'espressione di alcuni geni caratteristici. Lo scopo principale è stabilire se e in quale misura i geni a disposizione sono in grado di discriminare tra le tre tipologie di pazienti. Particolare attenzione viene riservata al confronto tra pazienti affetti dal Linfoma Diffuso a Grandi Cellule B (DLBCL) e pazienti colpiti dal Linfoma Follicolare (FL), sempre utilizzando i geni a disposizione. I dati sono stati forniti dal Dipartimento di Scienze Medico Diagnostiche e Terapie Speciali dell'Università degli Studi di Padova.

Il primo capitolo si occupa di presentare brevemente i linfomi non-Hodgkin, introducendo il DLBCL e il FL. Vengono inoltre presentati i geni che fungono da marcatori nelle analisi dell'elaborato. La parte finale del capitolo accenna all'imputazione dei dati mancati, operata al fine di poter svolgere le analisi successive.

Nel secondo capitolo vengono condotte delle analisi esplorative sui geni. Seguono dei test sulla bontà di adattamento e dei confronti sia tra le tre tipologie di pazienti che solamente tra pazienti affetti dal DLBCL e pazienti colpiti dal FL.

Il terzo capitolo si occupa di approfondire il confronto tra le prime due tipologie di pazienti, valutando la bontà delle classificazioni operate dai geni nel discriminare tra i pazienti affetti dal DLBCL e quelli colpiti dal FL. Dopo una breve introduzione teorica, si procede all'applicazione di un'analisi statistica basata sulla curva ROC.

Nel quarto ed ultimo capitolo si utilizzano degli strumenti di statistica multivariata allo scopo di raggruppare il campione a disposizione in gruppi utilizzando tutti i geni a disposizione. Dopo un paragrafo teorico introduttivo, si utilizzano i geni per svolgere delle analisi di

raggruppamento, utilizzando l'intero campione, basate su metodi gerarchici agglomerativi. Sempre ricorrendo a metodi gerarchici agglomerativi, si svolge un'analisi dei gruppi privando il dataset dei soggetti sani, e utilizzando solamente quei geni che hanno prodotto dei valori significativi nei test di confronto tra le prime due tipologie di pazienti. I risultati ottenuti vengono confrontati con i reali gruppi di pazienti che compongono il dataset.

L'ultimo passaggio è dedicato alle conclusioni, dove vengono discussi i risultati ottenuti, assieme ad alcune considerazioni sulle tecniche applicate.

Capitolo 1. I linfomi non-Hodgkin

Scopo di questo capitolo è introdurre le variabili presenti nel dataset. Dopo una sintetica spiegazione riguardante le tipologie di linfoma cui sono affetti i pazienti, vengono presentati alcuni geni, che fungono da marcatori per il confronto tra i gruppi presenti. Viene, infine, fatto un accenno all'imputazione dei dati mancati, necessaria per poter condurre le analisi successive.

1.1 Introduzione

I tumori delle cellule linfatiche costituiscono un'ampia famiglia di neoplasie che vanno dalle forme indolenti a quelle più aggressive. Tali neoplasie possono manifestarsi sotto forma di leucemia, se vi è un interesse del midollo osseo e del sangue, oppure sotto forma di linfomi, ossia tumori solidi del sistema immunitario. Non è esclusa la possibilità che la neoplasia presenti entrambi gli aspetti appena descritti, oppure possa manifestare aspetti leucemici, seppure inizialmente si tratti di un linfoma (Harrison, 2011, parte 6).

La classificazione istologica dei linfomi non-Hodgkin ha rappresentato una delle questioni più controverse in ambito oncologico (Harrison, 2011, parte 6). Nel 1999, l'Organizzazione Mondiale della Sanità (OMS) ha classificato le neoplasie a cellule linfatiche tenendo conto di aspetti morfologici, clinici, immunologici e sopratutto genetici, fornendo un'accuratezza diagnostica più elevata rispetto alle suddivisioni passate. Come si suol dire: "Imperfetti sistemi morfologici sono stati soppiantati da imperfetti sistemi immunologici" (Harrison, 2011, parte 6). Con questa frase, si vuole sottolineare come l'evoluzione della ricerca medica abbia portato da una classificazione delle neoplasie basata sul riconoscimento cellulare (in particolare della cellula Reed-Sternberg) ad una basata sugli aspetti anche genetici, tutt'ora in

fase di sviluppo (Rachel, E. et al., 2010). In Figura 1.1 è riportata una classificazione delle neoplasie linfatiche.

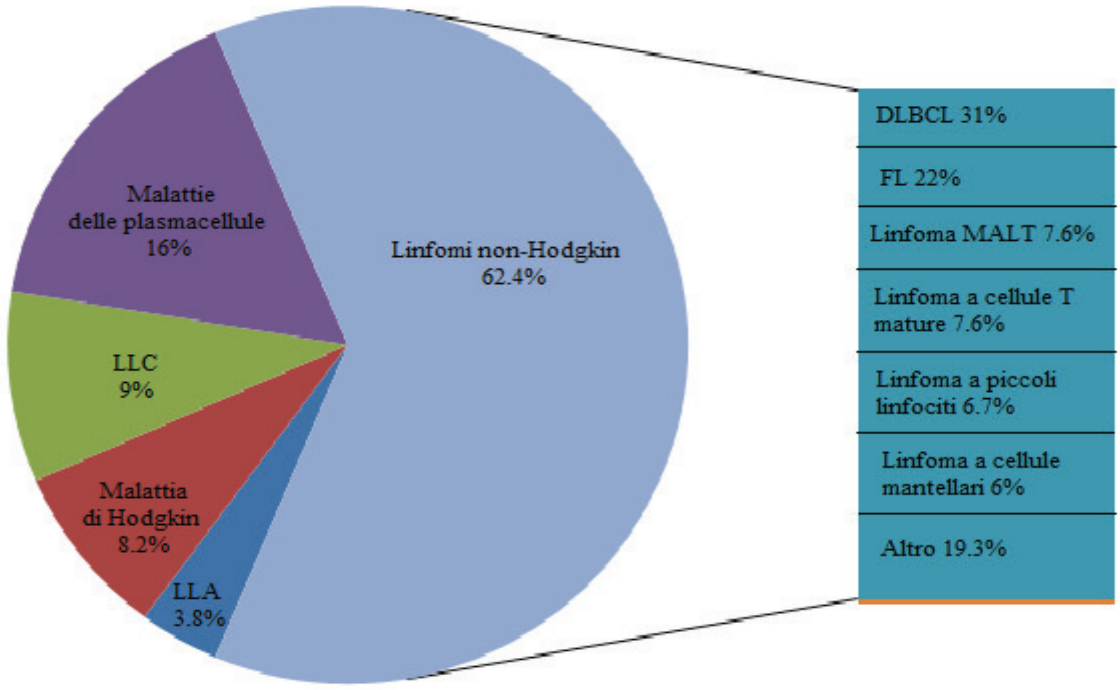

Figura 1.1: Frequenza relativa delle neoplasie linfatiche

Oggetto di questo studio sono i linfomi non-Hodgkin e una loro possibile classificazione, utilizzando come marcatori dei particolari geni. Un gruppo di pazienti è affetto dal linfoma diffuso a grandi cellule B, mentre un secondo gruppo presenta il linfoma follicolare. Vi sono infine anche 5 pazienti sani.

1.2 Linfoma Diffuso a Grandi Cellule B e Linfoma Follicolare

Il Linfoma Diffuso a Grandi Cellule B (DLBCL) è il più comune dei linfomi aggressivi non-Hodgkin (Harrison, 2011, parte 6). Si tratta di un tipo di linfoma che cresce molto rapidamente, con conseguente propagazione in più parti del corpo e la cui diversità dal punto di vista morfologico, clinico e genetico rispetto alle altre tipologie di linfomi lo rende unico.

Il modo in cui il DLBCL si presenta, collegato alla risoluzione clinica e alla sua eterogeneità patologica e biologica, suggerisce l'esistenza di più tipologie distinte, le quali richiedono

differenti approcci terapeutici. In particolare, si possono individuare 3 sottogruppi principali: Germinal Center B-cell (GCB-DLBCL), Activate B-cell (ABC-DLBCL) ed infine Primary Mediastinal DLBCL (PMBCL). A questi 3 sottogruppi corrispondono trattamenti clinici differenti e le probabilità di sopravvivenza a 5 anni dalla diagnosi risultano, rispettivamente, pari a 59%, 30% e 64% (Roehle, A., 2008).

Il DLBCL può manifestarsi sia come malattia primitiva dei linfonodi, che coinvolgere sedi extralinfonodali; più del 50% dei pazienti, al momento della diagnosi, presenta questo secondo aspetto (Harrison, 2011, parte 6).

Il FL è invece la seconda forma più diffusa dei linfomi non-Hodgkin. Al contrario del DLBCL esso rientra della categoria dei cosiddetti linfomi "indolenti", dei quali è il più comune, con una incidenza di circa 2 casi ogni 100.000 persone in Occidente (Roehle, A., 2008).

Mentre il DLCBL, se non trattato in maniera tempestiva, porta in breve tempo ad uno stato degenerativo (da qui il termine aggressivo), nel FL la sopravvivenza è valutabile in termini di anni (Ott, G. et al., 2002).

Morfologicamente, il FL è classificato dall'OMS in 3 gradi, in base ai centroblasti per grado di ingrandimento (HPF). A sua volta, il terzo grado viene ulteriormente suddiviso in 2 classi, una della quali dal punto di vista clinico è riconducibile al DLBCL (Groves, F. D. et al., 2000). Infatti, i pazienti con FL presentano un tasso di trasformazione verso il DLBCL pari al 7% per anno (Harrison, 2011, parte 6). La suddivisione del FL è sintetizzata nella Tabella 1.1.

Grado	**Valore**
1	<5 centroblasti / HPF
2	6-15 centroblasti / HPF
3	> 15 centroblasti / HPF Il grado 3 è suddiviso in: **3a** **3b** (qualità rara)

Tabella 1.1: Morfologia del linfoma follicolare (FL)

La diagnosi per il DLBCL non può prescindere dalla biopsia, in quanto tale malattia può interessare qualsiasi organo. La terapia prevede uno o più cicli polichemioterapici (solitamente quattro). Il FL è, invece, uno dei tumori più sensibile alla chemioterapia. In questo caso, con un adeguato trattamento, la completa remissione è raggiunta in una percentuale variabile tra il 50% e il 75% dei pazienti, anche se si contano situazioni di recidività in almeno nel 20% dei casi (Harrison, 2011, parte 6). Per questo tipo di linfoma sono disponibili inoltre una serie di terapie che differiscono dalla chemioterapia, quali ad esempio il trattamento con nuovi farmaci citotossici. La Figura 1.2 illustra il DLBCL e il FL.

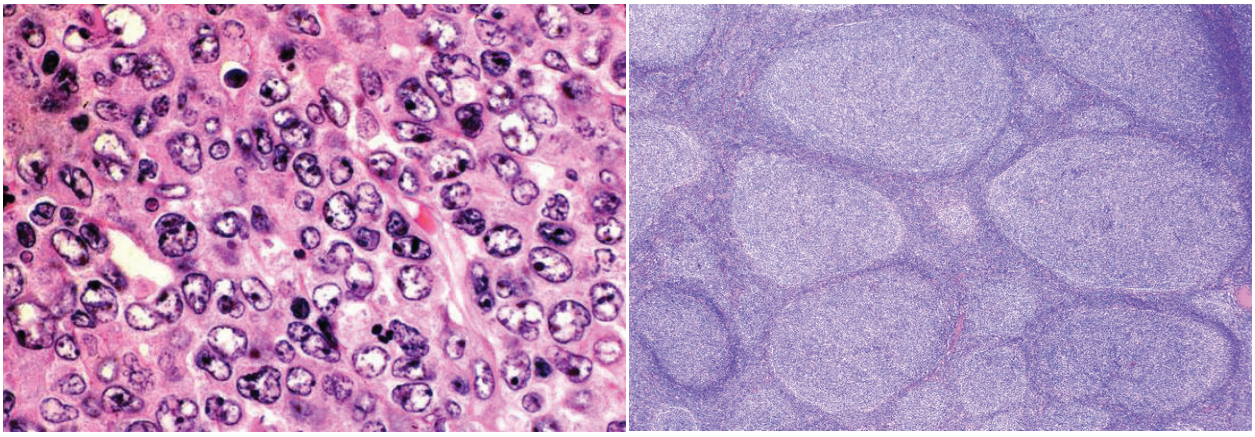

Figura 1.2: Linfoma diffuso a grandi cellule B (sinistra) e Linfoma follicolare (destra)

Come si vede dalla Figura 1.2 (sinistra), le cellule tumorali sono eterogenee, ma si nota un predominio delle grandi cellule, da cui il nome del linfoma. La Figura 1.2 (destra) evidenzia come i noduli siano di dimensioni variabili.

1.3 Il dataset

Il dataset fa riferimento a 60 pazienti suddivisi in 3 gruppi: 37 (61%) colpiti dal DLBCL (gruppo 0), 18 (30%) colpiti dal FL tipo 3b (gruppo 1) e gli ultimi 5 soggetti (9%) sono invece sani (gruppo 2).

Per ciascun paziente sono disponibili 10 variabili esplicative, riguardanti i valori relativi ad alcuni marcatori (miR18b, miR19b, miR20a, miR92a, miR93, miR99, miR106a, miR150, miR210NEW, miR135a).

Tali geni (miR e miRNA) sono dei regolatori negativi di espressioni genetiche, i quali ricoprono particolare importanza nell'individuazione del cancro. Infatti, i geni giocano un ruolo importante in quei processi biologici che riguardano lo sviluppo e la crescita delle cellule, ed espressioni alterate dei miRNA possono condurre alla diagnosi di cancro (Rachel, E. et al., 2010).

Nel corso dell'elaborato si usa il termine marcatore come sinonimo di gene. Tuttavia è utile operare una distinzione tra i due termini. Mentre un gene è definito come un'unità di eredità in un organismo vivente presente normalmente in una tratto di DNA, il marcatore, in particolare quello tumorale, consiste in proteine, ormoni o altre sostanze sintetizzate dalla cellula tumorale, le quali ne segnalano la presenza. Espressioni alterate dei geni, segnalate dai marcatori tumorali, possono indicare, indirettamente (ossia senza condurre l'esame istologico, la biopsia oppure la risonanza magnetica), la presenza di un tumore. Dal punto di vista clinico, i marcatori risultano di fondamentale importanza in quanto possono segnalare la presenza di un tumore prima che una identificazione sia possibile con gli strumenti diagnostici classici (Harrison, 2011, parte 6).

Al momento, nel genoma umano sono stati identificati ben 711 miRNA e vi è tutt'ora uno studio vivace al fine di stabilire la concreta possibilità di utilizzare i suddetti come marcatori prognostici e clinici (Harrison, 2011, parte 6).

Il dataset è riportato nella Tabella A.1 in Appendice A.

1.4 Imputazione dei dati mancanti

La prima operazione da svolgere per poter confrontare i vari gruppi consiste nell'imputazione dei dati mancanti presenti nel dataset.

Si nota che il paziente DLBCL42 può essere eliminato dalle indagini in quanto presenta solamente 4 misurazioni, sulle 10 variabili esplicative totali.

Per imputare i dati mancanti si sono considerate le correlazioni tra le 10 variabili esplicative. Una volta note le correlazioni, la variabile della quale si vuole imputare il dato mancante diventa la variabile risposta, mentre la variabile maggiormente correlata con essa costituisce la variabile indipendente. Si considera quindi un modello di regressione lineare semplice e si imputano i dati mancanti.

Per quanto concerne miR135, si nota che 17 pazienti sui 60 totali (28.3%) non dispongono di una misurazione con tale marcatore. Si decide, pertanto, di eliminare tale variabile dalle analisi che seguiranno nei capitoli successivi.

La matrice di correlazione tra i marcatori è riportata nella Tabella A.2 in Appendice A.

Va detto fin da subito che sulle variabili numeriche non si lavorerà con i dati in scala originale, bensì verrà applicata la trasformata logaritmica.

Nel prossimo capitolo si svolgono alcuni test sui marcatori al fine di stabilire se e quali geni discriminano tra i tre gruppi di pazienti. Sempre utilizzando i geni, verranno poi eseguite delle analisi per confrontare i pazienti affetti dal DLBCL e quelli colpiti dal FL.

Capitolo 2. Analisi Esplorative

Obiettivo di questo capitolo è, sulla base dei geni presentati al capitolo precedente, confrontare i gruppi presenti nel dataset (pazienti affetti da DLBCL, pazienti affetti dal FL e pazienti sani), con particolare attenzione al confronto tra le prime due tipologie di pazienti.

Le analisi sono svolte in due fasi. La prima operazione consiste nella verifica degli assunti che stanno alla base dei test che si vogliono utilizzare per il confronto tra i gruppi. Successivamente, in base ai risultati ottenuti, si applicano opportuni test sui marcatori al fine di confrontare i tre gruppi di pazienti e, in seconda istanza, i pazienti affetti dal DLBCL con quelli colpiti dal FL.

L'approccio seguito per accettare o rifiutare le ipotesi alla base dei test usa un livello di significatività α fissato al 5%. Valori del *p-value* inferiori a tale livello risultano significativi e pertanto conducono al rifiuto dell'ipotesi nulla, mentre valori superiori portano all'accettazione della medesima.

Alcuni testi di riferimento riguardanti i metodi utilizzati in questo capitolo sono Piccolo (1998), Pace e Salvan (2001) e Azzalini (2000).

Il software utilizzato per lo svolgimento delle analisi è R (www.r-project.org).

2.1 Relazione tra le variabili

Con riferimento alla Tabella A.2 in Appendice A e alla Figura 2.1, è possibile analizzare le relazioni tra le variabili (i geni) in esame.

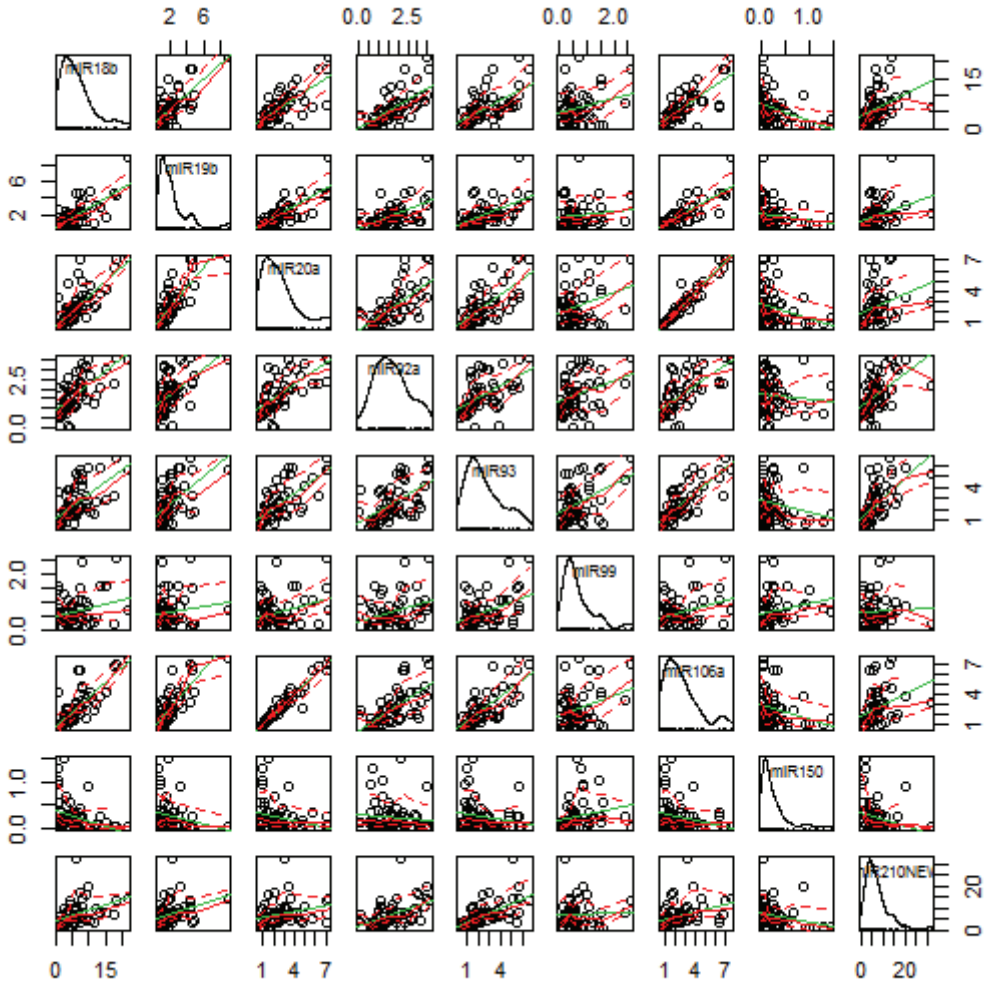

Figura 2.1: Grafici di dispersione tra i marcatori

Si nota che vi è una forte correlazione positiva tra i geni miR106a e miR20a (r = 0.9691). Inoltre, vi sono altri geni con una correlazione significativa, quali miR19b e miR20a (r = 0.8074), miR19b e miR106a (r = 0.8053) ed infine miR18b e miR106a (r = 0.7807). In definitiva, il gene miR106a sembra essere quello maggiormente correlato con gli altri.

Alcune statistiche di sintesi dei marcatori sono presentate nella Tabella 2.1.

Marcatori	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Standard deviation
miR18b	-0.49	1.01	1.60	1.46	2.05	3.07	0.85
miR19b	-0.56	-0.02	0.49	0.46	0.80	2.19	0.61
miR20a	-0.55	0.19	0.73	0.70	1.14	1.98	0.66
miR92a	-3.37	0.06	0.47	0.33	0.80	1.27	0.80
miR93	-0.93	0.31	0.74	0.72	1.23	1.89	0.66
miR99	-2.73	-1.16	-0.70	-0.71	-0.16	0.94	0.79
miR106a	-0.51	0.28	0.83	0.73	1.16	2.01	0.65
miR150	-5.18	-2.80	-1.97	-1.99	-1.16	0.38	1.19
miR210NEW	-0.96	1.21	1.69	1.61	2.09	3.46	0.84

Tabella 2.1: Statistiche di sintesi dei marcatori

2.2 Test sulla bontà di adattamento

Prima di procedere con il test più appropriato per gli scopi dell'analisi è necessario verificare alcuni assunti.

Nel caso in cui gli assunti di normalità e omoschedasticità non vengano rifiutati, si procederà con l'analisi della varianza a 1 fattore nel caso del confronto tra i tre gruppi, mentre si applicherà il test t di Student a 2 campioni quando si confronteranno i pazienti affetti dal DLBCL con quelli aventi il FL. In caso di rifiuto degli assunti, si procederà utilizzando test non parametrici: in particolare si ricorrerà al test di Kruskal-Wallis per confrontare i tre gruppi di pazienti e al test di Mann-Whitney per confrontare i pazienti affetti da DLBCL con quelli colpiti dal FL. Si rimarca, infine, che il terzo campione, ossia quello dei pazienti sani, conta solo 5 unità.

La verifica dell'ipotesi di normalità avviene ricorrendo al test di Shapiro-Wilk, mentre per verificare l'ipotesi di omoschedasticità si fa ricorso al test di Bartlett nel caso in cui i gruppi

siano tre oppure al test di omogeneità delle varianze per due gruppi, quando si analizzano solamente i pazienti affetti dal DLBCL e quelli affetti dal FL. Le Tabelle 2.2, 2.3 e 2.4 sintetizzano i risultati delle analisi.

Uno strumento grafico particolarmente utile per il confronto tra gruppi è il box-plot. In Figura 2.2 sono riportati i box-plot riguardanti ogni gene, suddivisi in base alla tipologia del paziente (DLBCL, FL, sani).

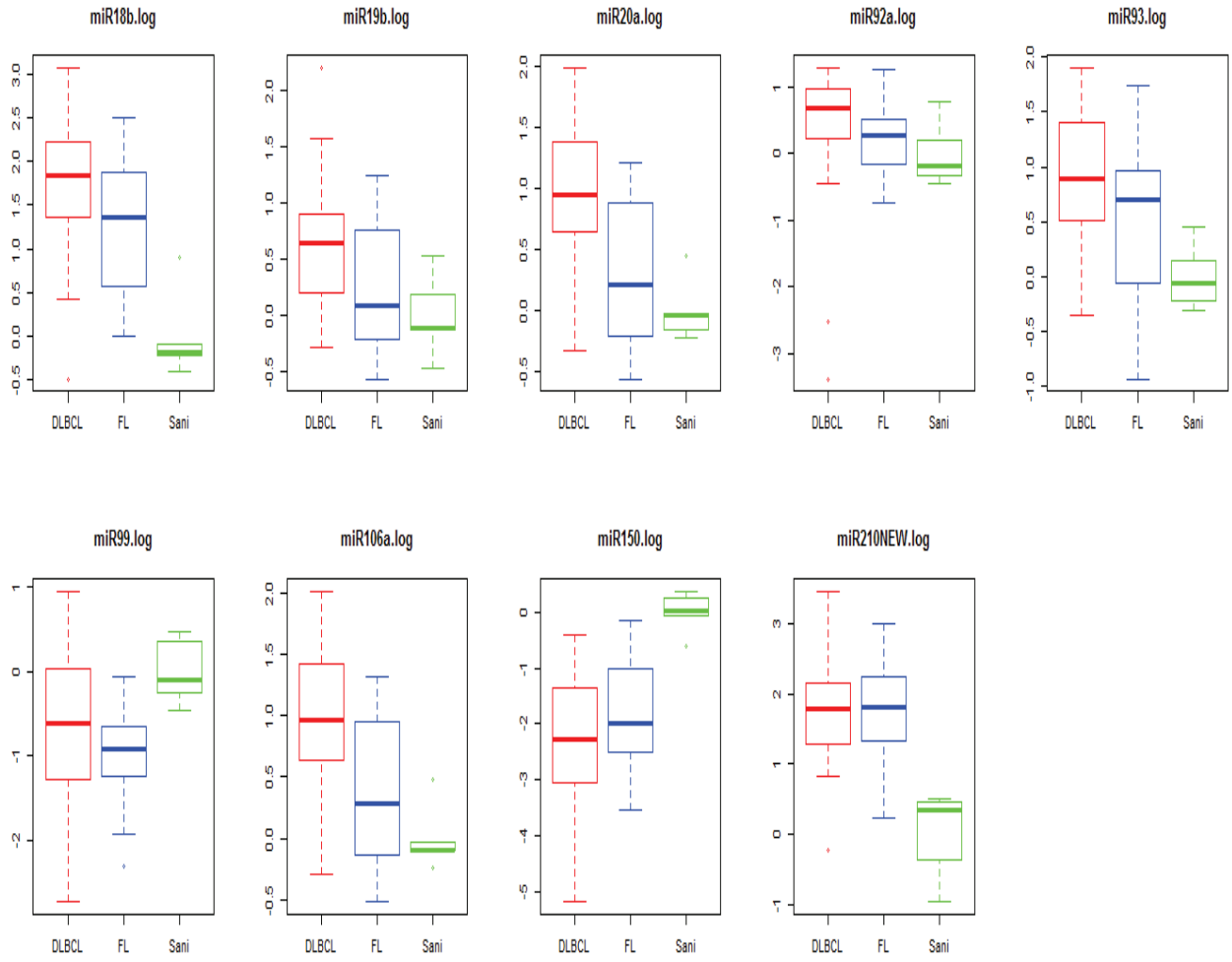

Figura 2.2: Box-plot dei geni suddivisi in base alla tipologia del paziente

Gene	*p-value* shapiro.test	accetto ip. di normalità?
miR18b	DLBCL: 0.18 FL: 0.80 Sani: 0.036	ACCETTO
miR19b	DLBCL: 0.12 FL: 0.14 Sani: 0.88	ACCETTO
miR20a	DLBCL: 0.77 FL: 0.16 Sani: 0.13	ACCETTO
miR92a	DLBCL: 1.45e-07 FL: 0.92 Sani: 0.43	RIFIUTO
miR93	DLBCL: 0.33 FL: 0.88 Sani: 0.69	ACCETTO
miR99	DLBCL: 0.96 FL: 0.41 Sani: 0.52	ACCETTO
miR106a	DLBCL: 0.71 FL: 0.32 Sani: 0.080	ACCETTO
miR150	DLBCL: 0.19 FL: 0.81 Sani: 0.56	ACCETTO
miR210NEW	DLBCL: 0.44 FL: 0.73 Sani: 0.17	ACCETTO

Tabella 2.2: Verifica dell'ipotesi di normalità nel confronto tra i tre gruppi

L'ipotesi di normalità viene accettata per tutti geni, eccetto miR92a. Per tale gene il test di Shapiro-Wilk, relativo al gruppo di pazienti affetti da DLBCL, ha fornito un valore del *p-value* prossimo a 0. Pertanto l'ipotesi di normalità viene rifiutata.

Per quanto riguarda il gruppo dei pazienti sani, data la limitata numerosità campionaria di sole 5 unità, non vi sono elementi sufficienti per rifiutare l'assunto di normalità, che viene pertanto accettato.

Per quanto riguarda la verifica dell'assunzione di normalità nel caso si voglia confrontare il gruppo di pazienti affetti da DLBCL con quello aventi il FL si ha che, sempre facendo riferimento alla Tabella 2.2, si rifiuta tale assunto solamente per il gene miR92a, per gli stessi motivi descritti sopra nel caso del confronto tra i tre gruppi.

Gene	***p-value* bartlett.test**	**accetto ip. di omoschedasticità?**
miR18b	0.73	ACCETTO
miR19b	0.53	ACCETTO
miR20a	0.26	ACCETTO
miR92a	0.0098	RIFIUTO
miR93	0.20	ACCETTO
miR99	0.079	ACCETTO
miR106a	0.29	ACCETTO
miR150	0.058	ACCETTO
miR210NEW	0.97	ACCETTO

Tabella 2.3: Verifica dell'ipotesi di omoschedasticità nel confronto tra i tre gruppi

Per quanto concerne la verifica del secondo assunto alla base dei test, ossia quello di omoschedasticità tra i tre gruppi, si nota come il valore del *p-value* riguardante miR92a (0.0098) risulti significativo per ogni livello di α usuale (1%, 5%, 10%). Si decide pertanto di rifiutare l'ipotesi di omoschedasticità solamente per il marcatore miR92a.

Gene	*p-value* var.test	accetto ip. di omoschedasticità?
miR18b	0.90	ACCETTO
miR19b	0.74	ACCETTO
miR20a	0.86	ACCETTO
miR92a	0.0050	RIFIUTO
miR93	0.34	ACCETTO
miR99	0.091	ACCETTO
miR106a	0.78	ACCETTO
miR150	0.23	ACCETTO
miR210NEW	0.83	ACCETTO

Tabella 2.4: Verifica dell'ipotesi di omoschedasticità nel confronto tra i due gruppi di pazienti (DLBCL e FL)

Il test per la verifica dell'omogeneità delle varianze per due gruppi porta all'accettazione dell'ipotesi nulla per tutte le variabili eccetto miR92a, il cui valore del *p-value* (0.0050) risulta significativo per tutti i livelli di α usuali.

Nelle analisi che seguono, si è deciso di utilizzare il test non parametrico di Kruskal-Wallis per confrontare i tre gruppi di pazienti nel caso in cui il marcatore sia miR92a, mentre nei restanti confronti si utilizza l'analisi della varianza a 1 fattore. Una maniera per pervenire a quale confronto conduce al rifiuto dell'ipotesi di uguaglianza delle medie delle tre distribuzioni, nel caso in cui si faccia uso dell'analisi della varianza ad 1 fattore, consiste nell'avvalersi di test post-hoc, quali ad esempio il test di Fisher oppure il test di Bonferroni (Piccolo, 1998).
Per quanto riguarda il confronto tra pazienti affetti da DLBCL e pazienti colpiti dal FL, il test usato è, invece, quello non parametrico di Mann-Whitney nel caso in cui il marcatore sia miR92a, mentre per i restanti geni si utilizzerà il test t di Student a 2 campioni.

2.3 Confronto tra gruppi

Come si nota dalla Tabella 2.5, i test di confronto tra le tre tipologie di pazienti portano al rifiuto dell'ipotesi di uguaglianza dei 3 gruppi per tutti i nove geni considerati.

Gene	*p-value*	accetto l'ip. di uguaglianza tra i 3 gruppi?
miR18b	7.28e-06	RIFIUTO
miR19b	0.014	RIFIUTO
miR20a	1.52e-05	RIFIUTO
miR92a	0.013	RIFIUTO
miR93	0.0024	RIFIUTO
miR99	0.031	RIFIUTO
miR106a	8.702e-05	RIFIUTO
miR150	3.48e-05	RIFIUTO
miR210NEW	5.64e-06	RIFIUTO

Tabella 2.5: Test per il confronto tra i tre gruppi

Per i geni miR19b, miR92a, miR99 il valore del test ha prodotto dei *p-values* significativi al 5% ma non all'1%. Per i restanti marcatori, si rifiuta l'ipotesi nulla di uguaglianza delle medie per le tre tipologie di pazienti per ogni livello di α usuale.

Passando al confronto tra pazienti affetti da DLBCL e pazienti colpiti dal FL, dalla Tabella 2.6 si può notare che per tre geni (miR99, miR150 e miR210NEW) il valore del *p-value* porti all'accettazione dell'ipotesi di uguaglianza delle due medie. In particolare, il test fornisce per il marcatore miR210NEW un *p-value* particolarmente alto (*p-value* = 0.76). Per i restanti geni, invece, il valore del *p-value* conduce al rifiuto dell'ipotesi nulla.

Gene	*p-value*	accetto l'ipotesi di uguaglianza tra i 2 gruppi?
miR18b	0.027	RIFIUTO
miR19b	0.029	RIFIUTO
miR20a	0.00022	RIFIUTO
miR92a	0.015	RIFIUTO
miR93	0.04	RIFIUTO
miR99	0.093	ACCETTO
miR106a	0.00026	RIFIUTO
miR150	0.056	ACCETTO
miR210NEW	0.76	ACCETTO

Tabella 2.6: Confronto tra i due gruppi di pazienti (DLBCL e FL)

2.4 Conclusioni sui confronti tra i gruppi

I test svolti nel paragrafo precedente hanno evidenziato come, per quanto riguarda il confronto tra i tre gruppi di pazienti, l'ipotesi di uguaglianza delle medie dei gruppi venga rifiutata per tutti e nove i geni. Nell'analisi di confronto tra i tre gruppi aventi come marcatore i geni miR19b, miR92a e miR99, il test ha prodotto un valore del *p-value* non significativo all'1% ma significativo al 5%. Per i restanti geni, i valori dei *p-values* prodotti dal test risultano significativi per ogni livello di α usuale e pertanto l'ipotesi di uguaglianza delle medie dei tre gruppi viene rifiutata.

Il discorso è leggermente diverso per i confronti tra il gruppo formato dai pazienti con il DLBCL e quello composto dai pazienti affetti dal FL. Utilizzando come marcatore il gene miR210NEW, il valore del test produce un *p-value* elevato (*p-value* = 0.76), il quale conduce all'accettazione dell'ipotesi di uguaglianza delle medie dei due gruppi. Per i marcatori miR99 (*p-value* = 0.093) e miR150 (*p-value* = 0.056) il test produce dei *p-values* non significativi al 5% e pertanto, anche in questo caso, l'ipotesi di uguaglianza delle medie delle due distribuzioni di pazienti non viene rifiutata.

L'ipotesi di uguaglianza delle medie per i due gruppi di pazienti viene invece rifiutata utilizzando i rimanenti geni. In particolare, i valori dei *p-values* per le analisi svolte sui marcatori miR20a (*p-value* = 0.00022) e miR106a (*p-value* = 0.00026) conducono al rifiuto dell'ipotesi di uguaglianza delle medie dei due gruppi per ogni livello di α usuale.

Nel prossimo capitolo si effettueranno ulteriori analisi su quei marcatori il cui test t di Student ha portato al rifiuto dell'ipotesi di uguaglianza delle medie delle due distribuzioni di pazienti. In particolare, si cercherà di valutare la bontà della regola di classificazione utilizzata dai singoli geni per assegnare il paziente ad uno dei due possibili gruppi (DLBCL e FL).

Capitolo 3. Analisi con la curva ROC

Obiettivo di questo capitolo è valutare l'accuratezza dei test diagnostici basati sui marcatori nel discriminare tra pazienti affetti da DLBCL e pazienti colpiti dal FL. Per fare ciò si utilizzano quei marcatori il cui test (Tabella 2.5) ha prodotto un valore significativo nel confronto tra i due gruppi di pazienti. Si svolge quindi un'analisi statistica basata sulla curva ROC. Per uno studio approfondito della curva ROC si rimanda a Krzanowski ed Hand (2009).

3.1 Introduzione alla curva ROC

La curva ROC (*Receiver Operating Characteristic*) è una tecnica statistica nata durante la Seconda Guerra Mondiale e utilizzata largamente in ambito scientifico. Tale tecnica ha trovato una notevole applicazione nella valutazione della bontà di test diagnostici discriminatori, soprattutto in campo medico.

La natura dei dati analizzati in questa tesi fa si che i test diagnostici utilizzati siano di tipo quantitativo. Quando si ricorre a questa tipologia di test è necessario, come primo passo, individuare quel valore in grado di discriminare tra pazienti "positivi" e pazienti "negativi" al test. Tale valore prende il nome di soglia (o *cut-off*) e risulta di fondamentale importanza in quanto, nel nostro caso, discrimina tra pazienti affetti da DLBCL e pazienti affetti dal FL.

E' noto che per una particolare scelta del *cut-off* si possono ottenere tre situazioni distinte.

Nella prima situazione (Figura 3.1) il valore di soglia fissato discrimina perfettamente tra pazienti affetti da una patologia e pazienti affetti da un'altra. Si ha in questo caso quella che viene definita la "situazione ideale".

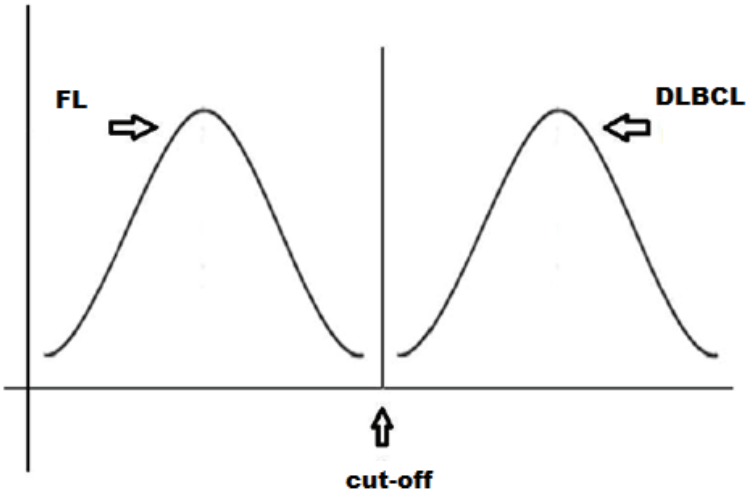

Figura 3.1: Discriminazione perfetta

La seconda situazione (Figura 3.2) riguarda il caso in cui il test è privo di potere diagnostico e non è possibile attribuire il soggetto ad una o all'altra categoria.

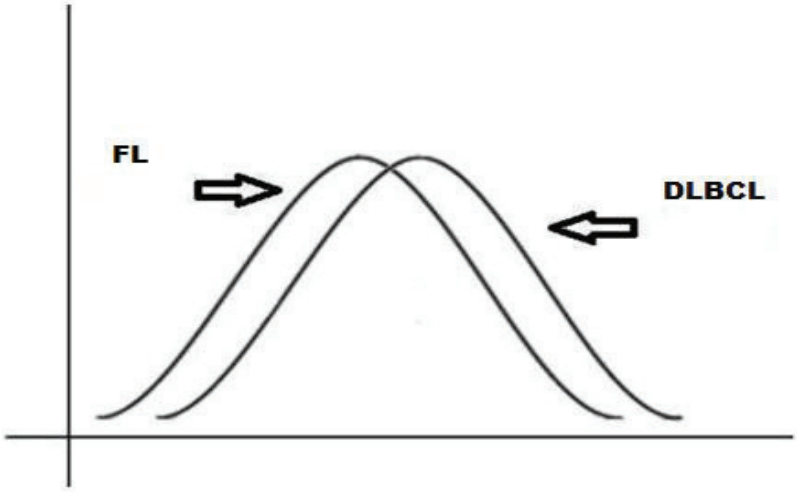

Figura 3.2: Allocazione casuale

Infine, la terza situazione (Figura 3.3), quella più comune, tratta il caso in cui, per un determinato valore di soglia, si ha una zona di sovrapposizione delle due distribuzioni. In altre parole, non tutti i soggetti affetti da DLBCL vengono classificati nel gruppo corretto e non tutti i pazienti affetti dal FL finiscono nel loro gruppo di appartenenza. Si hanno allora dei Falsi Negativi (FN) e dei Falsi Positivi (FP).

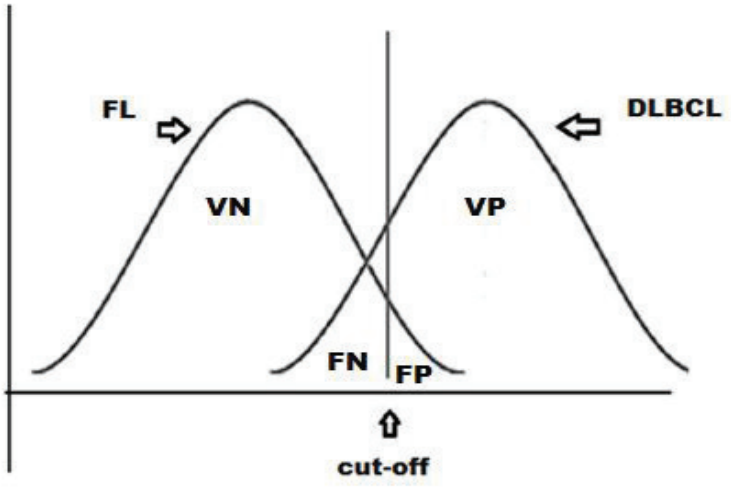

Figura 3.3: Situazione comune

Nelle analisi che seguono viene confrontato, di volta in volta, l'output prodotto dal test eseguito con un particolare marcatore con il vero stato del paziente.

Tale confronto può dare luogo a quattro situazioni distinte, riconducibili allo schema utilizzato in Figura 3.3:

- ✓ Veri Positivi (VP): pazienti affetti da DLBCL correttamente classificati.
- ✓ Falsi Negativi (FN): pazienti affetti da DLBCL classificati erroneamente come portatori del FL.
- ✓ Veri Negativi (VN) : pazienti affetti dal FL correttamente classificati.
- ✓ Falsi Positivi (FP): pazienti affetti dal FL classificati erroneamente come portatori del DLBCL.

Questi valori possono essere sintetizzati in una tabella a doppia entrata che prende il nome di matrice di confusione (o tabella di errata classificazione), riportata in Tabella 3.1.

	Paziente DLBCL	**Paziente FL**	**totale**
Test Positivo	VP (Veri Positivi)	FP (Falsi Positivi)	VP + FP
Test Negativo	FN (Falsi Negativi)	VN (Veri Negativi)	FN + VN
totale	VP + FN	FP + VN	

Tabella 3.1: Matrice di Confusione

Utilizzando la matrice di confusione è possibile reperire degli indici sintetici recanti informazioni sulla qualità della classificazione operata. I principali sono l'accuratezza, la sensibilità e la specificità. L'accuratezza misura la validità del test, tenendo conto della proporzione di Falsi Negativi e Falsi Positivi: tanto più bassi saranno questi valori tanto maggiore sarà la validità del test.

La sensibilità esprime la proporzione di Veri Positivi (VP) rispetto al numero effettivo di positivi, mentre la specificità esprime la proporzione di Veri Negativi (VN) rispetto al numero effettivo di negativi. Un test è sensibile al 100% quando non presenta Falsi Negativi, mentre è specifico al 100% quando non vi sono Falsi Positivi.

Sensibilità e specificità sono però inversamente correlati tra di loro e quindi una modifica del valore di soglia può generare due effetti come evidenziato in Figura 3.4:

- ✓ aumento della sensibilità e diminuzione della specificità.
- ✓ aumento della specifività e diminuzione della sensibilità.

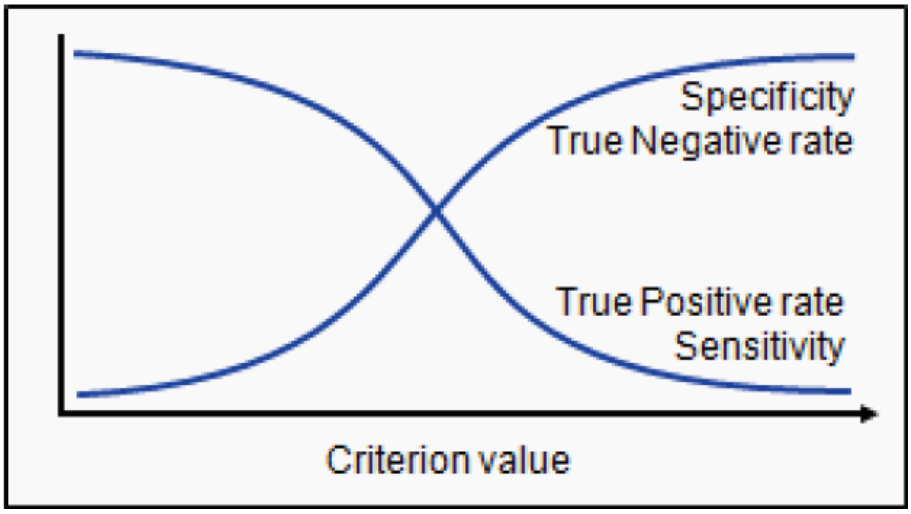

Figura 3.4: Relazione tra specificità e sensibilità al variare della soglia

3.2 Costruzione della Curva ROC

La situazione ideale è rappresentata dal test che fornisce un valore della sensibilità e un valore della specificità pari al 100%. Siccome nella maggior parte dei casi reali ciò non avviene, un'altra via potrebbe essere quella di massimizzare contemporaneamente sensibilità e specificità ma ciò non è possibile.

La soluzione consiste allora nel riportare un sistema di assi cartesiani, per ogni possibile valore del *cut-off*, con in ordinata il corrispondente valore della sensibilità (proporzione di veri positivi) e in ascissa quello del complemento a uno della specificità. L'unione dei vari punti ottenuti genera la curva ROC.

Tanto maggiore è la bontà di un test discriminatorio, tanto la curva ROC è al di sopra della diagonale. La Figura 3.5 sintetizza quest'ultimo aspetto.

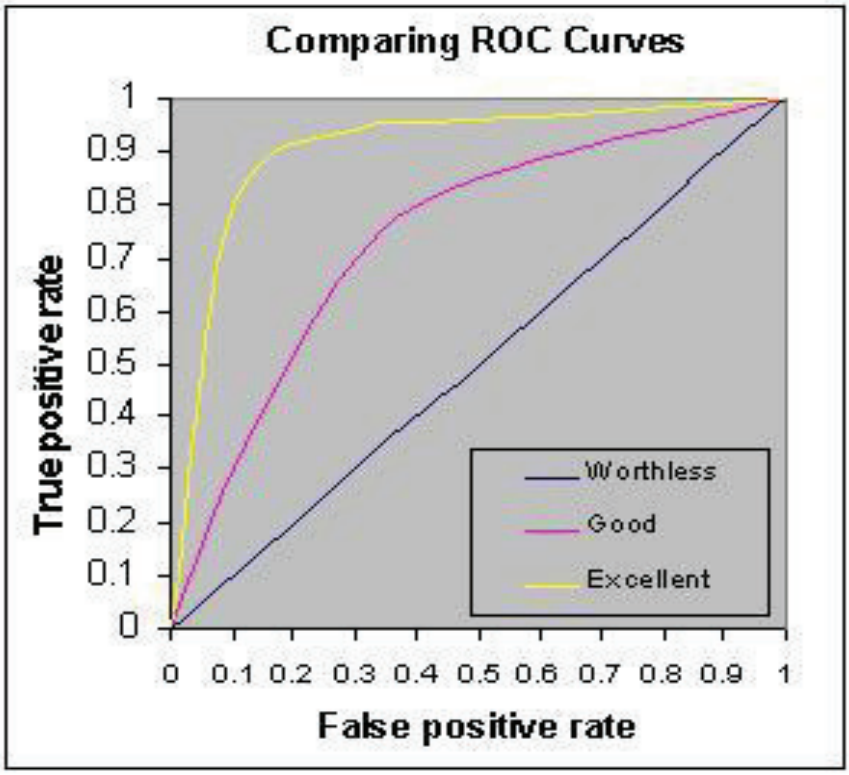

Figura 3.5: Esempi di curva ROC

Una misura sintetica dell'efficacia di una test diagnostico derivato dalla curva ROC, ossia un indicatore in grado di valutare la bontà della regola di classificazione, è l'area sottesa dalla curva ROC (AUC, *Area Under the ROC Curve*), il cui compito è stimare la probabilità di assegnare un'unità statistica al suo reale gruppo di appartenenza. Per l'interpretazione del valore dell'AUC si può far riferimento alla proposta di Swets (1988):

- ✓ AUC = 0.5: test non informativo
- ✓ 0.5 < AUC ≤ 0.7: test poco accurato
- ✓ 0.7 < AUC ≤ 0.9: test moderatamente accurato
- ✓ 0.9 < AUC < 1.0: test altamente accurato
- ✓ AUC = 1.0: test perfetto

Una volta costruita la curva ROC è anche possibile scegliere il valore di soglia ottimale, ossia quel valore che rappresenta il miglior compromesso tra sensibilità e specificità. Un modo per reperire tale quantità consiste nel calcolare la distanza tra l'angolo superiore sinistro del grafico (che rappresenta sensibilità e specificità pari al 100%) e il punto della curva ROC più vicino ad esso (Bottarelli e Parodi, 2003).

3.3 Applicazione della curva ROC al caso di studio

Nella parte finale del Capitolo 2 si è visto come sei dei nove confronti tra le due tipologie di pazienti abbiano portato al rifiuto dell'ipotesi di uguaglianza delle medie per i due gruppi.
In questo capitolo si desidera valutare la bontà della classificazione operata dai sei marcatori che hanno portato al rifiuto dell'ipotesi di uguaglianza delle medie dei due gruppi di pazienti. Ci si aspetta che i marcatori miR20a e miR106a, i quali hanno portato al rifiuto dell'ipotesi di uguaglianza delle medie delle due distribuzioni per ogni livello di α usuale, possiedano un valore dell'AUC maggiore rispetto agli altri geni e di conseguenza abbiano un potere discriminatorio migliore.

In Appendice A.3-8, sono riportate le Tabelle utilizzate per lo svolgimento delle analisi.

La Tabella 3.2 contiene una sintesi dei risultati ottenuti utilizzando i diversi marcatori. Le informazioni riguardano, per ogni gene, il valore dell'AUC e l'interpretazione del medesimo utilizzando la regola di Swets. Si considerano inoltre, sotto ipotesi di normalità, un intervallo di confidenza al 95% per l'area sottesa alla curva ROC ed una verifica di ipotesi basata sul test alla Wald, dove l'ipotesi nulla H_0: AUC = 0.5 indica la mancanza di potere discriminatorio da parte del gene utilizzato.

Gene	**AUC**	**Interpretazione**	**Accettazione Ipotesi nulla H_0: AUC = 0.5?**	**Intervallo di Confidenza al 95% per AUC**
miR18b	0.69	test poco accurato	NO	0.53-0.85
miR19b	0.67	test poco accurato	NO	0.51-0.84
miR20a	**0.79**	test moderatamente accurato	NO	0.66-0.92
miR92a	**0.71**	test moderatamente accurato	NO	0.56-0.85
miR93	0.66	test poco accurato	SI	0.50-0.82
miR106a	**0.79**	test moderatamente accurato	NO	0.66-0.92

Tabella 3.2: AUC per i 6 geni utilizzati nell'analisi basata sulla curva ROC

Conducendo l'analisi basata sulla costruzione della curva ROC, i marcatori miR20a, miR92a e miR106a risultano quelli con il valore dell'AUC più grande e di conseguenza con miglior regola di classificazione.

Le Figure 3.6, 3.7 e 3.8 mostrano le stime delle due distribuzioni, sotto ipotesi di normalità, e la curva ROC rispettivamente per i tre geni sopra citati.

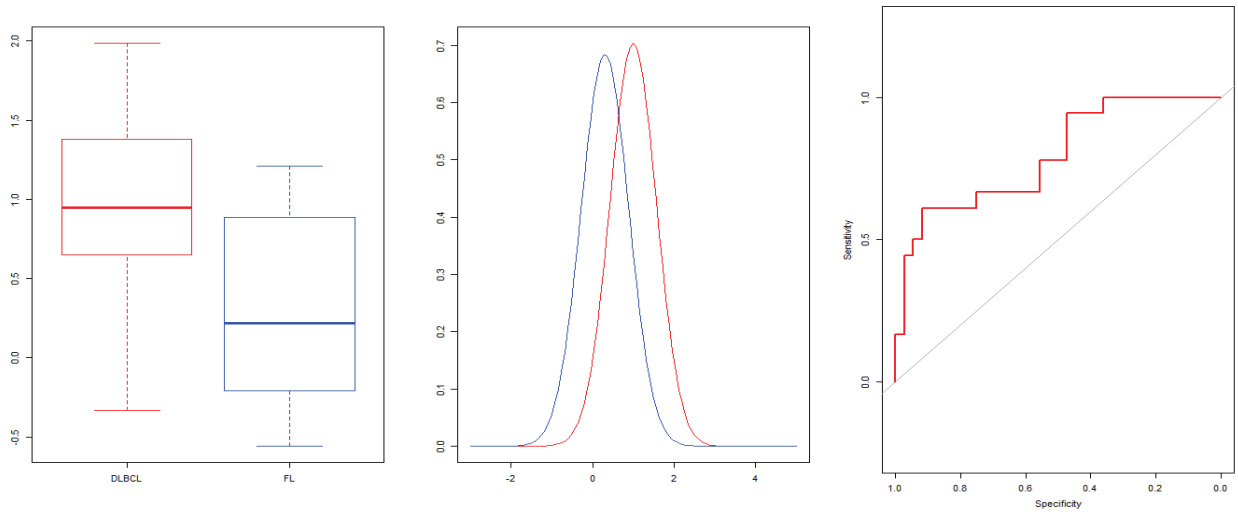

Figura 3.6: Analisi con la curva ROC per miR20a

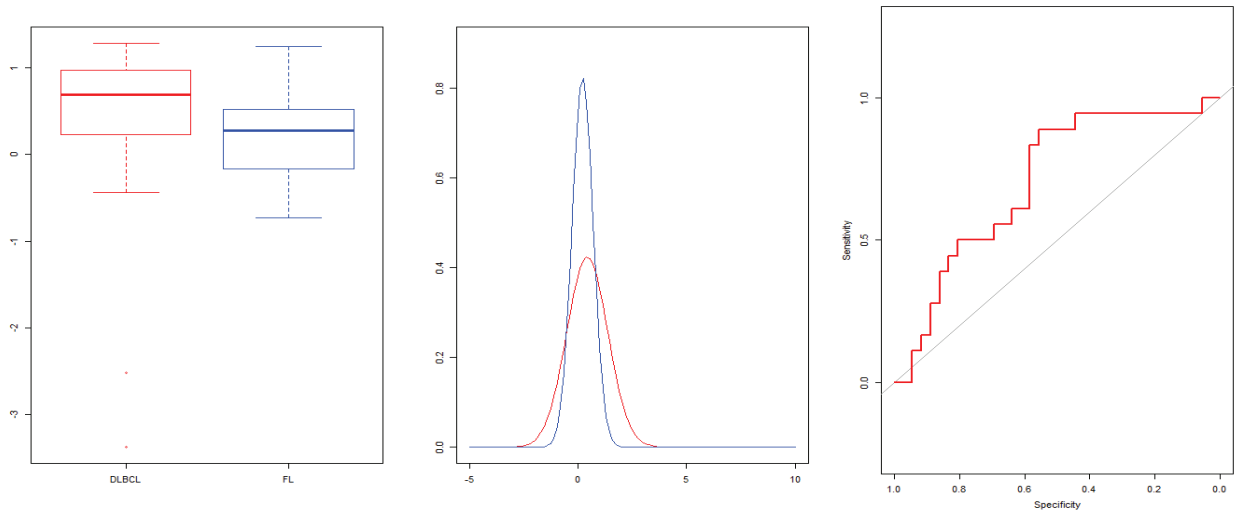

Figura 3.7: Analisi con la curva ROC per miR92a

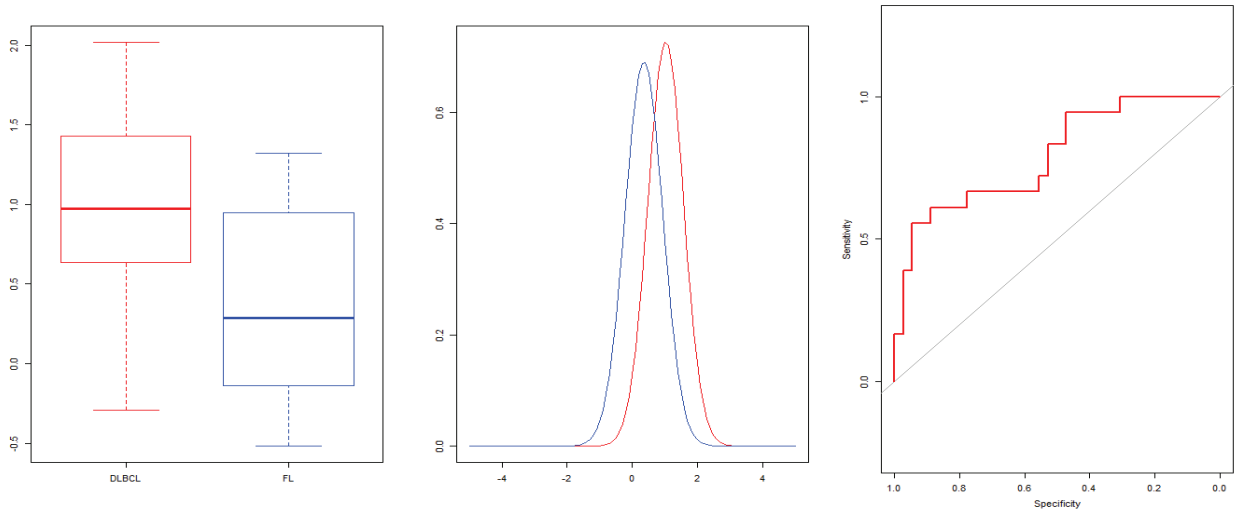

Figura 3.8: Analisi con la curva ROC per miR106a

Guardando le stime delle distribuzioni per i due gruppi di pazienti si nota come non ci sia mai una perfetta separazione tra i gruppi e, di fatto, si vada sempre incontro a dei falsi negativi e a dei falsi positivi.

Per questi tre marcatori può risultare utile determinare il valore di soglia ottimale, ossia quello che rappresenta il miglior compromesso tra sensibilità e specificità.

La Tabella 3.3 indica i valori del *cut-off* ottimale, della sensibilità e della specificità corrispondente a tale soglia, rispettivamente per miR20a, miR92a e miR106a.

Gene	***Cut-off* ottimale**	**Sensibilità**	**Specificità**
miR20a	0.26	0.61	0.92
miR92a	0.58	0.83	0.58
miR106a	0.60	0.78	0.67

Tabella 3.3: *Cut-off* ottimale per i geni che sono risultati migliori nell'analisi con la curva ROC

I risultati ottenuti risultano coerenti con le analisi svolte nel capitolo precedente. I marcatori con le performance migliori sono infatti gli stessi che nel test in cui si saggiava l'ipotesi di uguaglianza delle medie delle due distribuzioni di pazienti avevano prodotto *p-values* che portavano al rifiuto di tale assunto per ogni livello di α usuale.

Per quanto riguarda invece i restanti marcatori, l'interpretazione dell'AUC conduce a risultati poco soddisfacenti; in particolare per il gene miR93 il valore dell'AUC è pari a 0.66 e la verifica d'ipotesi porta all'accettazione dell'ipotesi nulla H_0: $AUC_{miR93} = 0.5$. Data, in questo caso, la dualità tra intervalli di confidenza e verifiche di ipotesi, si nota che l'intervallo di confidenza al 95% per miR93 contiene il valore 0.5 (Pace e Salvan, 2001).

Una stima delle due distribuzioni di pazienti, sempre sotto ipotesi di normalità, e il grafico della curva ROC corrispondente alle analisi condotte con il marcatore miR93 sono disponibili in Figura 3.9.

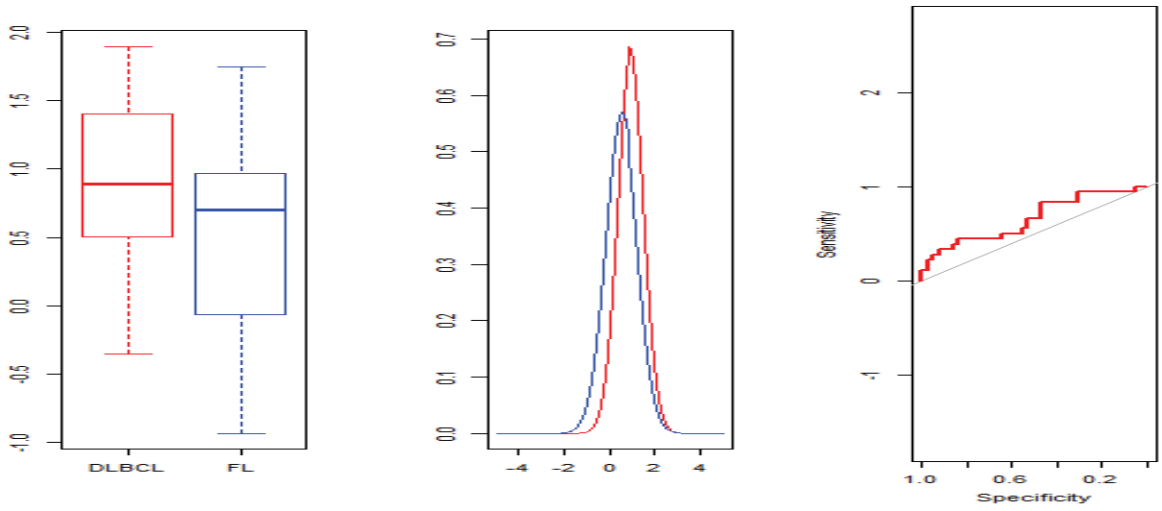

Figura 3.9: Curva ROC per il gene miR93

3.4 Conclusioni

Le analisi svolte con la curva ROC hanno evidenziato come nessun marcatore sia in grado di effettuare, stando alla regola di classificazione dell'AUC di Swets, delle performance altamente accurate. Tuttavia i marcatori miR20a, miR92a e miR106a hanno prodotto una regola di classificazione moderatamente accurata. Sempre guardando il valore dell'AUC, si nota come i geni miR20a e miR106a siano quelli che possiedono la regola di classificazione migliore, in accordo con quanto evidenziato nel Capitolo 2.

Nel prossimo capitolo di questo elaborato si testeranno ulteriormente i geni a disposizione utilizzando alcune tecniche di classificazione; rientrerà nelle analisi anche il gruppo di pazienti sani che in questo capitolo non è stato oggetto di studio.

Capitolo 4. Analisi dei gruppi

In questo capitolo si conducono alcune analisi di statistica multivariata, con l'obiettivo di suddividere in categorie i pazienti presenti nel dataset. Inizialmente si ignora lo stato del paziente (DLBCL, FL oppure sano) e si utilizzano i marcatori a disposizione per raggruppare pazienti simili all'interno dello stesso gruppo e pazienti diversi in altri gruppi.

Le analisi che vengono svolte sono suddivise in due fasi. La prima fase tiene conto di tutti i pazienti a disposizione e le analisi hanno lo scopo di valutare la bontà del raggruppamento svolto utilizzando i nove geni. Nella seconda fase sono oggetto di studio solamente i pazienti affetti dal DLBCL e quelli colpiti dal FL; in questo caso i geni utilizzati sono solamente quei sei che, nelle analisi svolte al Capitolo 2, hanno portato al rifiuto dell'ipotesi di uguaglianza delle medie delle due distribuzioni di pazienti. Nella parte finale del Capitolo si confronta il raggruppamento ottenuto con i reali gruppi di appartenenza dei pazienti, al fine di valutare la bontà della classificazione operata e l'eterogeneità tra i due cluster.

Per un approfondimento sulle tecniche di raggruppamento, si rimanda ad esempio ad Azzalini e Scarpa (2004).

4.1 Analisi di raggruppamento

L'analisi di raggruppamento (o analisi *cluster*) rientra nell'ambito della statistica multivariata non-supervisionata dove, a differenza del caso supervisionato, non si suppone che la classe di appartenenza di un individuo sia osservabile direttamente. Tutti i geni sono, pertanto, posti sullo stesso piano. L'obiettivo consiste nello stabilire se e in quanti gruppi è possibile suddividere il campione alla luce delle variabili osservate (i geni). Dato che suddividere un campione in gruppi significa assegnare allo stesso gruppo pazienti con caratteristiche simili

tra loro ma dissimili da quelle di altri gruppi, risulta fondamentale reperire qualche strumento in grado di misurare la somiglianza tra le varie unità. Tali strumenti prendono il nome di misure di similarità e cambiano a seconda della tipologia delle variabili in gioco. In particolare, nel caso di variabili quantitative, si ricorre alla misura di Mincowski, definita come

$$d_p(x_i, x_j) = \left[\sum_{h=1}^{q} |\mathrm{x_{ih}} - \mathrm{x_{jh}}|^{\mathrm{p}}\right]^{\frac{1}{p}},$$

dove, facendo uso della notazione a doppio indice, $\mathrm{x_{ih}}$ indica l'*i*-esima unità statistica valutata per l'*h*-esima variabile esplicativa (gene), *q* indica il numero di variabili esplicative e *p* è un numero reale positivo. Al variare di *p*, corrispondono differenti distanze di similarità. Una delle più importanti, utilizzata anche nelle analisi che seguono, per $p = 2$, è la distanza euclidea.

Le analisi di raggruppamento si suddividono in due tipologie distinte. Vi sono innanzitutto i metodi di partizione, dove il numero di gruppi in cui si vuole suddividere il campione viene fissato a priori. L'algoritmo più utilizzato per operare l'analisi di raggruppamento con il metodo di partizione è quello delle *k*-medie (Azzalini e Scarpa, 2004).

Si può poi fare uso di metodi gerarchici, dove si individua una sequenza di partizioni. I metodi gerarchici si dicono agglomerativi se, attraverso una serie di successive fusioni, si raggruppano i singoli individui in gruppi sempre più ampi sino ad avere un solo cluster contenente tutte le unità statistiche. In caso contrario, i metodi gerarchici vengono detti scissori. Il risultato finale dell'analisi consiste in una struttura ad albero che prende il nome di dendrogramma. Il numero ottimo di gruppi può essere scelto ispezionando il dendrogramma risultante dall'analisi gerarchica e sezionandolo all'altezza del massimo salto tra livelli di somiglianza ai quali sono avvenute le aggregazioni cluster. Nelle analisi con il dendrogramma l'altezza indica il livello di eterogeneità tra i gruppi; tanto maggiore è la differenza di altezza

tra i gruppi, tanto più dissimili saranno tra loro. Vi sono vari algoritmi per costruire il dendrogramma, i quali si differenziano per il diverso criterio che regola la valutazione delle distanze tra i gruppi. Vengono riportate, di seguito, le regole solitamente utilizzate per agglomerare i diversi gruppi:

- ✓ **Metodo del legame singolo**: la misura della distanza tra i due gruppi è la minima distanza tra tutte le coppie di punti di cui il primo elemento è nel primo gruppo e il secondo elemento è nel secondo gruppo.
- ✓ **Metodo del legame completo**: la misura della distanza tra due gruppi è la massima distanza tra tutte le coppie di punti di cui il primo elemento è nel primo gruppo e il secondo elemento è nel secondo gruppo.
- ✓ **Metodo del legame medio**: in questo caso si considerano tutte le distanze tra le coppie di punti di cui il primo elemento è nel primo gruppo e il secondo elemento è nel secondo gruppo e si calcola la media non ponderata di tutte le distanze.
- ✓ **Metodo di Ward**: questo algoritmo aggrega i gruppi in modo che l'incremento di varianza nei nuovi gruppi sia il minimo possibile. A differenza degli altri algoritmi, il metodo di Ward non richiede il calcolo di distanze.

La procedura statistica, nel caso di metodi gerarchici agglomerativi, utilizzata al fine di raggruppare i dati, impiega ricorsivamente i metodi illustrati in precedenza.
L'algoritmo ricorsivo consiste nei seguenti 5 passi:

1. **inizializzazione**: ogni paziente rappresenta un cluster. Si hanno cioè in totale *n* gruppi;
2. **selezione**: si selezionano i due cluster più simili rispetto alla misura di similarità che si è deciso di utilizzare. Al termine di questo passo si contano *n-1* gruppi.

3. **aggiornamento**: si aggiornano, rispettivamente, il numero dei cluster e la matrice delle distanze;
4. **ripetizione**: si applicano i passi 2 e 3 per *n-1* volte;
5. **arresto**: l'algoritmo termina una volta che tutti gli elementi appartengono ad un unico cluster.

Le analisi che seguono fanno uso di metodi gerarchici agglomerativi, in cui vengono utilizzate la distanza euclidea e il metodo di Ward.

4.2 Applicazione delle analisi cluster a tutti i pazienti

Utilizzando tutti i 59 pazienti presenti nel dataset le analisi di raggruppamento hanno prodotto il dendrogramma di Figura 4.1.

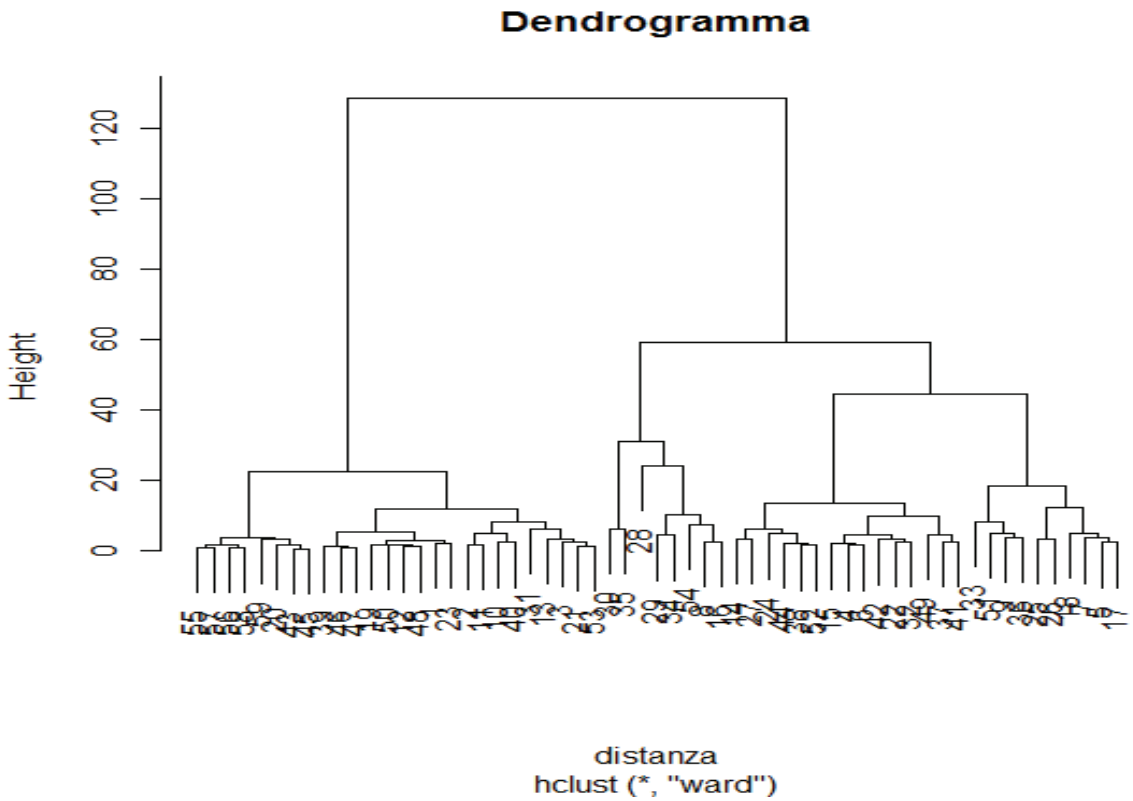

Figura 4.1: Dendrogramma utilizzando il metodo di Ward con il campione completo

Il primo aspetto che si nota guardando il dendrogramma è che i 5 pazienti sani, identificati dai numeri dal 55 al 59, appaiono molto simili tra loro. Si nota inoltre come, per una distanza

pari a 60, si vengano a formare due gruppi eterogenei tra loro. Non si nota, tuttavia, una separazione secondo quelli che sono i reali gruppi di appartenenza dei pazienti. I sani infatti risultano simili ai pazienti affetti dalle neoplasie già ad un'altezza pari a 10. Una suddivisione in tre gruppi è possibile per una distanza compresa tra 40 e 60, ma non vi è corrispondenza tra i risultati delle analisi di raggruppamento e la reale classe di appartenenza dei pazienti. La scelta di utilizzare la tecnica Ward piuttosto che un'altra, risulta determinante nei risultati dell'analisi di raggruppamento. La Figura 4.2 mostra le stesse analisi svolte utilizzando il metodo del legame medio.

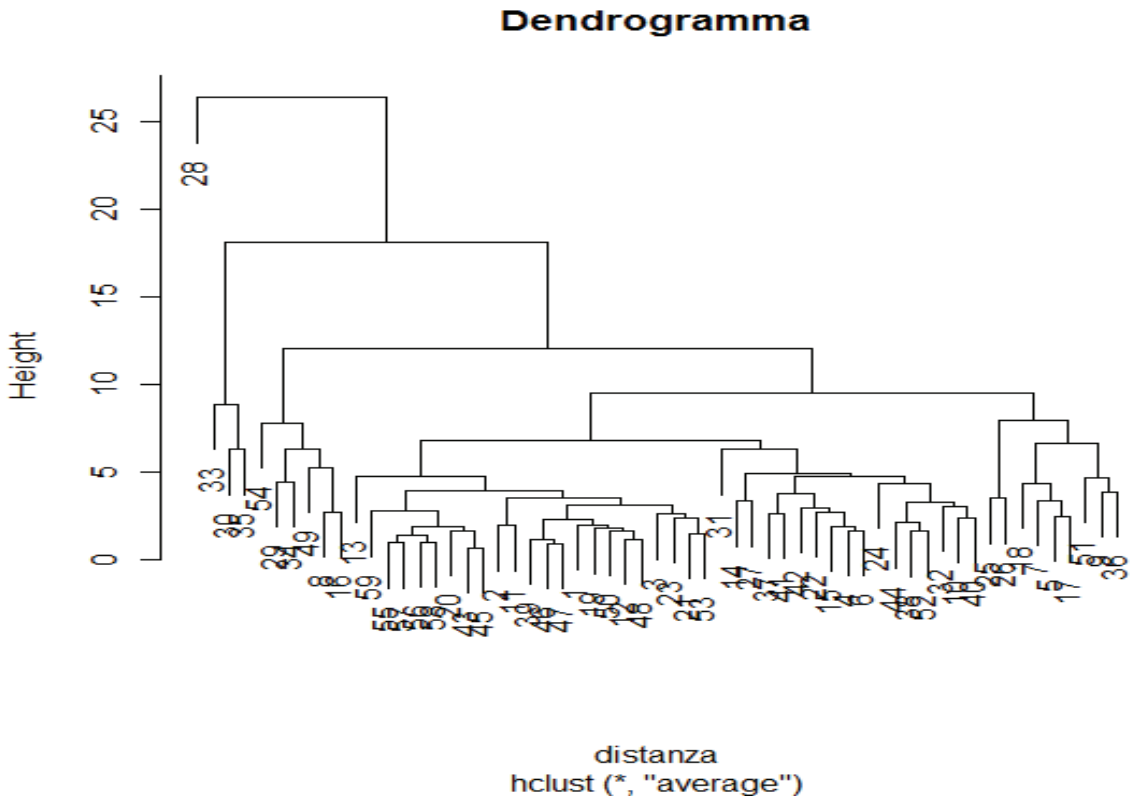

Figura 4.2: Dendrogramma utilizzando il metodo del legame medio con il campione completo

Dal dendrogramma si nota come, in questo caso, vi sia, per una distanza pari a 20, un gruppo formato da 58 pazienti, mentre l'altro presenta solamente una unità. Ciò è dovuto al fatto che tale paziente presenta un valore anomalo per quanto riguarda il gene miR210NEW che lo rende diverso da tutti gli altri con un valore pari a 31.91, mentre la media generale delle

osservazioni per tale gene è 6.76. I pazienti sani continuano, come in precedenza, a risultare simili tra loro, ma non si nota una corrispondenza tra i risultati prodotti dal raggruppamento e la reale tipologia dei gruppi, in quanto pazienti affetti dal DLBCL e pazienti colpiti dal FL vengono raggruppati all'interno dello stesso cluster.

4.3 Applicazione delle analisi cluster alle prime due tipologie di pazienti

Eliminando il gruppo composto dai soggetti sani e utilizzando solamente quei sei geni che hanno prodotto dei valori significativi nel test di confronto tra i primi due gruppi di pazienti (Tabella 2.6), le analisi di raggruppamento hanno prodotto i risultati sintetizzati dalla Figura 4.3.

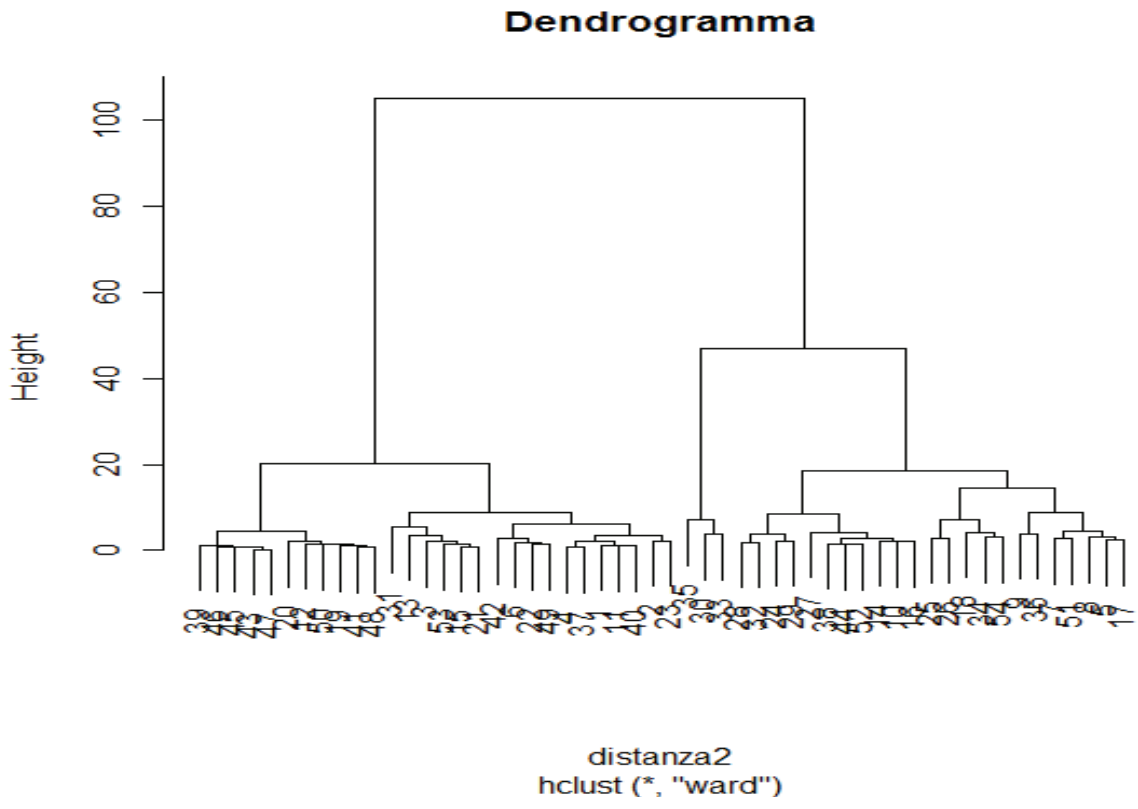

Figura 4.3: Dendrogramma utilizzando il metodo di Ward avente come campione le due tipologie di malati (DLBCL e FL)

Il dendrogramma mostra come, per un'altezza pari circa a 60, ci siano due gruppi eterogenei tra loro. Il primo cluster contiene 28 pazienti, 15 con il DLBCL e 13 colpiti dal FL. Il secondo

invece presenta 26 pazienti, 21 dei quali sono affetti dal DLBCL. Si nota come, entrambi i cluster ottenuti, presentino sia pazienti affetti dal DLBCL che dal FL.

Utilizzando il metodo del legame medio si ottengono i risultati sintetizzati dalla Figura 4.4.

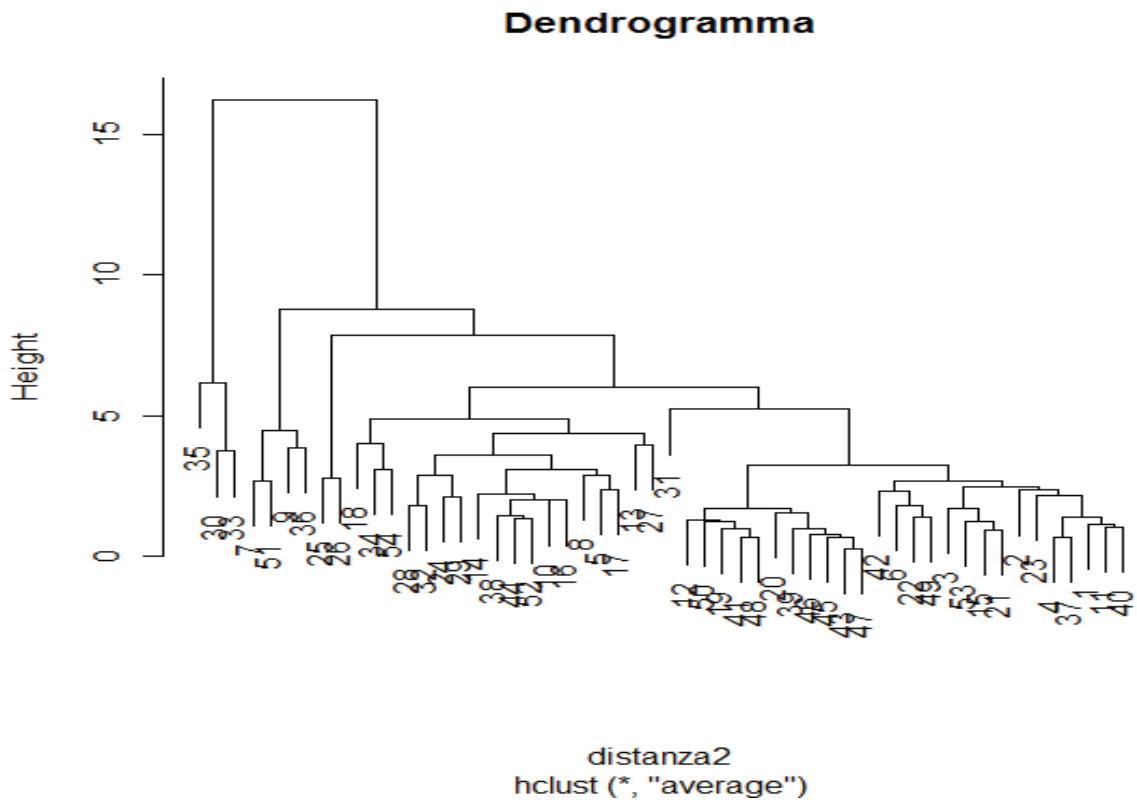

Figura 4.4: Dendrogramma utilizzando il metodo del legame medio avente come campione le due tipologie di malati (DLBCL e FL)

Per una distanza pari circa a 10 si possono notare 2 gruppi; il primo conta però solamente 3 pazienti. Questo aspetto è spiegabile osservando le misurazioni riguardanti il gene miR18b; i tre pazienti possiedono, rispettivamente, un valore di 18.03, 17.81 e 21.56, contro la media generale pari a 5.91 per tale marcatore. Si ottengono risultati del tutto similari applicando la regola del legame singolo oppure quella del legame completo.

Utilizzando il dendrogramma di Figura 4.3 si può valutare la bontà del raggruppamento operato utilizzando metodi gerarchici agglomerativi.

Una volta partizionato il dendrogramma, ad un'altezza tale che evidenzi due cluster, è possibile confrontare il raggruppamento ottenuto con il reale gruppo di appartenenza dei pazienti. La Tabella 4.1 sintetizza tale confronto.

	Gruppo 1	**Gruppo 2**
Pazienti DLBCL	15	21
Pazienti FL	13	5

Tabella 4.1: Tabella a doppia entrata che confronta il raggruppamento ottenuto con la reale partizione dei pazienti

Ciò che ci si propone di svolgere ora è un confronto tra i due cluster individuati tagliando il dendrogramma di Figura 4.3 ad un'altezza pari circa a 60, utilizzando come marcatori i geni. In Figura 4.5 sono riportati i boxplot riguardanti i sei geni, dove Gruppo 1 e Gruppo 2 indicano i due cluster.

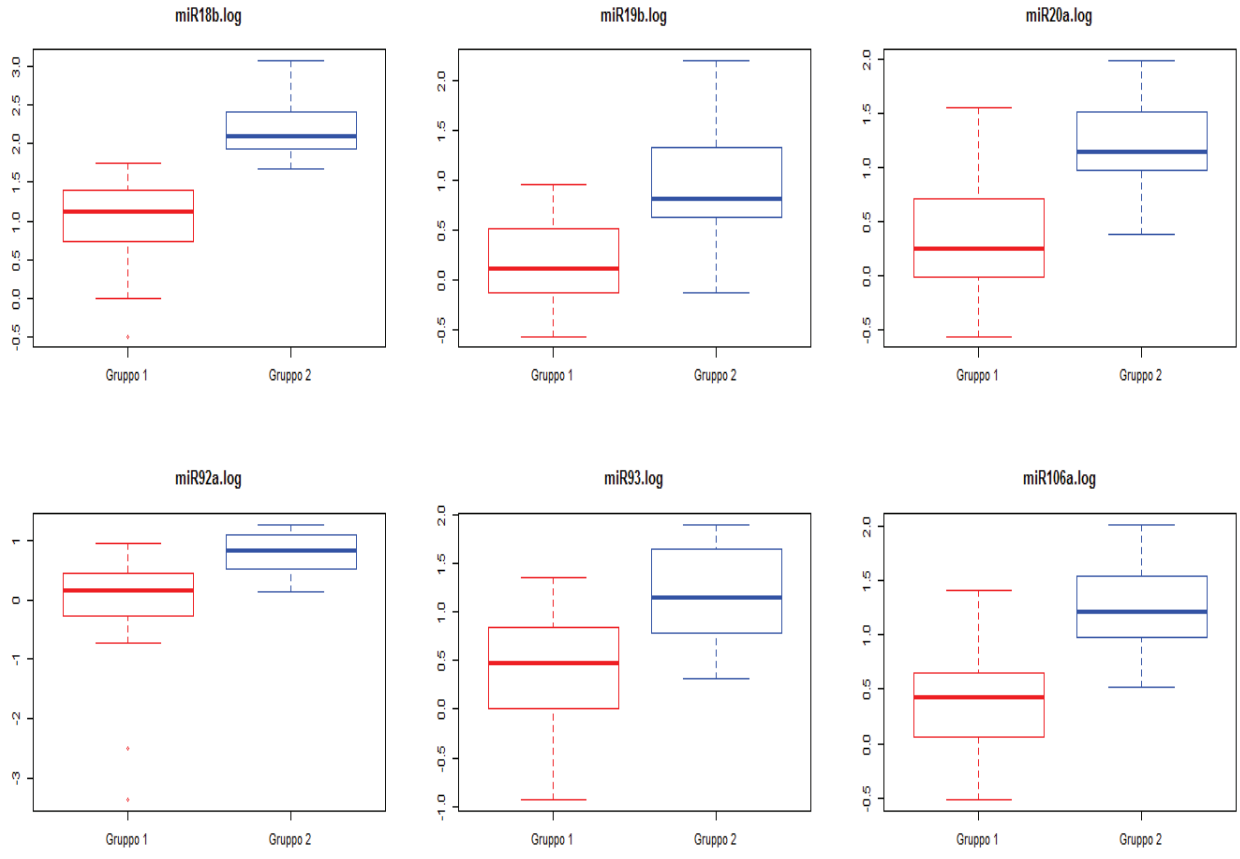

Figura 4.5: Boxplot dei due gruppi utilizzando i sei geni

Dai boxplot si nota come tutti i sei geni sembrano discriminare tra i due gruppi risultanti dall'analisi cluster. Si nota inoltre come i valori dei sei geni per il Gruppo 2 siano maggiori rispetto a quelli del Gruppo 1.

La Tabella 4.2 sintetizza i risultati ottenuti confrontando i due gruppi, utilizzando come marcatori i sei geni sotto ipotesi di normalità ed omoschedasticità ed applicando il test t di Student a 2 campioni, similmente a quanto svolto al Capitolo 2. In questo caso, alla luce delle considerazioni effettuate analizzando i boxplot di Figura 4.5, si utilizza un'ipotesi alternativa unilaterale a sinistra, fermo restando l'ipotesi nulla di uguaglianza delle medie delle distribuzioni dei due gruppi.

Gene	***p-value***	**accetto l'ipotesi di uguaglianza tra i 2 gruppi?**
miR18b	1.353e-12	RIFIUTO
miR19b	9.938e-07	RIFIUTO
miR20a	3.092e-09	RIFIUTO
miR92a	2.308e-05	RIFIUTO
miR93	7.545e-06	RIFIUTO
miR106a	3.248e-10	RIFIUTO

Tabella 4.2: Confronto tra Gruppo 1 e Gruppo 2 utilizzando i sei geni

L'ipotesi di uguaglianza delle medie delle due distribuzioni viene rifiutata per tutti i sei geni. Il test ha prodotto in tutti i casi dei *p-values* prossimi a 0.

4.4 Conclusioni

Scopo di questo capitolo è valutare se e in quanti gruppi sia possibile raggruppare i pazienti considerando tutti i geni a disposizione, prima utilizzando il campione completo, poi privandolo dei soggetti sani. Le analisi cluster, utilizzando tutti i pazienti, hanno evidenziato come i 5 pazienti sani siano molto simili tra loro e quindi facilmente raggruppabili in un unico gruppo. Purtroppo però a loro volta i soggetti sani risultano simili agli altri pazienti, mentre

pazienti affetti da DLBCL e pazienti affetti da FL vengono raggruppati all'interno dello stesso cluster.

Le analisi condotte utilizzando solamente le prime due tipologie di pazienti hanno evidenziato come si vengano a creare due cluster eterogenei, ciascuno dei quali contiene sia i pazienti affetti dal DLBCL che quelli colpiti dal FL. In particolare il primo cluster conta 28 pazienti (15 DLBCL e 18 FL) mentre il secondo ne presenta 26 (21 DLBCL e 5 FL). Le analisi di confronto tra i due cluster, utilizzando come marcatori i geni, hanno portato al rifiuto dell'ipotesi di uguaglianza delle medie delle due distribuzioni per tutti i sei miRNA. Tale aspetto sottolinea come i due cluster individuati dal dendrogramma di Figura 4.3 siano eterogenei tra loro.

Inoltre, valori alti nelle misurazioni di tutti i sei geni sembrano individuare il Gruppo 2 e, conseguentemente, la possibilità che il paziente di cui si desidera classificare la tipologia di linfoma sia affetto dal DLBCL. Un discorso simile non vale per il Gruppo 1 in quanto esso contiene, quasi in pari numero, sia pazienti affetti dal DLBCL che pazienti colpiti dal FL.

CONCLUSIONI FINALI

Obiettivo principale di questo elaborato è stabilire se e in quale misura i geni a disposizione siano in grado di discriminare tra le tre tipologie di pazienti presenti nel dataset.

Le analisi svolte consentono di trarre delle conclusioni in merito a questo quesito. Per quanto riguarda la differenziazione tra i tre gruppi di pazienti (DLBCL, FL e sani), le analisi svolte al Capitolo 2 portano al rifiuto dell'ipotesi di uguaglianza delle medie per le tre distribuzioni di pazienti per tutti i nove geni. In particolare per sei geni (miR18b, mir20a, miR93, miR106a, miR150 e miR210NEW) il test di analisi della varianza a 1 fattore ha prodotto dei *p-values* significativi per tutti i livelli di α usuali. Le analisi di statistica multivariata, operata utilizzando metodi gerarchici agglomerativi, confermano che i pazienti sani risultano molto simili tra loro e di conseguenza raggruppabili in un unico cluster. Una classificazione in pazienti sani, pazienti affetti dal DLBCL e pazienti colpiti dal FL, interpretando il dendrogramma, non appare soddisfacente. I pazienti sani risultano non solo simili tra loro, ma anche simili ai portatori di DLBCL e FL, mentre una distinzione tra pazienti affetti da DLBCL e pazienti affetti da FL non risulta apprezzabile.

Passando alle conclusioni riguardanti le analisi svolte utilizzando i pazienti affetti da DLBCL e quelli affetti da FL, il Capitolo 2 rileva come l'ipotesi di uguaglianza delle medie per i due gruppi di malati venga rifiuta utilizzando come marcatori sei geni (miR18b, miR19b, miR20a, miR92a, miR93, miR106a). In particolare, l'ipotesi di uguaglianza delle medie per le due distribuzioni di pazienti viene rifiuta ad ogni livello di α usuale per miR20a e miR106a. L'analisi basata sulla curva ROC, ossia lo strumento utilizzato nel terzo capitolo per valutare la bontà della regola di classificazione operata dal marcatore, sottolinea come i geni miR20a, miR92a e miR106a siano moderatamente accurati nel discriminare tra pazienti affetti dal

DLBCL e pazienti colpiti dal FL. L'interpretazione dell'AUC sottolinea come i geni miR20a e miR106a posseggano la regola di classificazione migliore nel discriminare tra pazienti affetti dal DLBCL e pazienti affetti dal FL. I risultati prodotti risultano pertanto coerenti, in quanto i geni che possiedono la regola di classificazione migliore sono gli stessi che conducono al rifiuto dell'ipotesi di uguaglianza delle due medie per tutti gli α usuali. I geni miR20a e miR106a sono inoltre fortemente correlati positivamente ($r = 0.9691$) e pertanto possiedono comportamenti simili. Le analisi di raggruppamento svolte utilizzando il campione privato dei soggetti sani hanno prodotto, al livello in cui era di interesse partizionare il dendrogramma, due cluster che possiedono, al loro interno, entrambe le tipologie di pazienti. Confronti tra i due cluster, utilizzando come marcatori i geni, conducono al rifiuto dell'ipotesi di uguaglianza delle medie dei due gruppi per tutti i miRNA, sottolineando come i due cluster individuati dall'analisi di raggruppamento siano eterogenei tra loro. Un confronto con i reali gruppi di appartenenza dei pazienti mostra come il raggruppamento ottenuto ricorrendo ai metodi gerarchici agglomerativi presenti, all'interno dei due gruppi, sia casi di DLBCL che di FL. Il Gruppo 1 contiene infatti 15 pazienti affetti dal DLBCL e 13 colpiti dal FL, mentre il Gruppo 2 conta 21 casi di DLBCL e 5 di FL.

Si nota però come il Gruppo 2, contenente prevalentemente pazienti affetti dal DLBCL, presenti misurazioni per tutti i sei geni più elevate rispetto alle corrispondenti del Gruppo 1. Valori alti di tutti i sei geni utilizzati potrebbero dunque segnalare la presenza del DLBCL per il paziente analizzato.

Appendice A

Tabella A.1: Dataset Linfomi

	Paziente	Type	miR18b	miR19b	miR20a	miR92a	miR93	miR99	miR106a	miR150	miR210NEW	miR135a
1	DLBCL1	0	4.3057000	1.2430000	1.1931123	1.56645356	1.3624217	0.84294949	1.1674354	0.549726472	3.8054997	3.5797644
2	DLBCL2	0	5.4389000	2.1171000	2.0790402	0.99316875	0.7020682	0.06483240	1.8784488	0.005575796	2.9450581	0.4727064
3	DLBCL3	0	3.3825000	1.5171000	2.3609666	2.61020922	1.4342449	0.22867897	1.9031839	0.053845120	5.1026398	1.3961942
4	DLBCL4	0	3.9644000	0.9461000	1.3784052	2.04973144	2.3554456	0.27874155	1.5626558	0.069878866	7.6603683	1.5284510
5	DBCL5	0	9.1538000	2.2281000	2.9810055	1.47389832	3.3312770	0.10266090	3.4997118	0.110768093	4.6996033	0.3073705
6	DLBCL6	0	2.9880000	1.1330000	1.5586311	1.26432650	3.3058763	0.61372888	1.7130083	0.042971546	7.5715012	0.7466992
7	DLBCL7	0	11.2415000	0.8732000	2.4397596	1.27746793	1.5829592	0.27239038	2.3965311	0.041002868	3.3497078	0.9727206
8	DLBCL8	0	9.3244000	1.2269000	1.4569991	3.09580604	2.5070660	1.00486382	1.6748130	0.047071577	15.6129292	1.3684377
9	DLBCL9	0	14.9110000	1.6764000	4.1478342	3.00307692	2.8339132	1.54898571	4.2380669	0.116546299	7.4035280	0.4109371
10	DLBCL10	0	6.8528000	1.0160000	2.0199088	1.87264793	1.7192256	0.43087338	2.3193111	0.073509888	5.2371091	1.3111230
11	DLBCL11	0	5.4524000	0.9368000	1.2320731	1.22480554	1.4871603	0.49100483	1.3743140	0.331314658	2.3077594	1.6348444
12	DLBCL12	0	3.0700000	0.7448000	0.7184701	0.73523929	0.8802590	0.25208811	0.7469067	0.176287245	3.2713431	3.2933217
13	DLBCL13	0	3.9272000	1.6472000	4.7502387	1.88406729	2.2283873	0.69351549	2.6371872	0.165625877	3.0514714	0.5026854
14	DBCL14	0	6.3340000	1.8550000	3.1901265	2.90067723	1.3663145	0.57027114	3.0236623	0.300802275	9.5296564	1.0197409
15	DBCL32	0	3.1283000	1.5765000	2.0856757	1.24994897	1.8121653	0.40293968	2.3575691	0.672682223	7.7626321	1.5159303
16	DLBCL33	0	7.0217000	1.9119000	2.0820432	2.95506198	3.0440668	0.68160130	1.8908042	0.259894859	16.4662545	1.2746817
17	DLBCL34	0	9.2436000	1.3149000	3.5259206	3.15322701	2.1081828	0.43437109	3.1470498	0.033984012	4.1062302	0.9859435
18	DLBCL35	0	10.1228000	4.8130000	4.5489543	2.67431331	1.7353815	0.22937541	4.2610126	0.007488748	4.3104434	1.5846472
19	DLBCL36	0	2.7182000	1.1949000	1.2943180	1.10724418	2.0200409	1.26885396	1.5267094	0.213267518	3.4054331	0.6137705
20	DLBCL37	0	1.5322000	1.3020000	1.6677800	0.64300702	1.0890942	0.56278992	1.5889816	0.324010171	0.7938680	1.6298777
21	DLBCL38	0	2.9820000	2.3111000	2.5350229	1.45038761	1.8725041	0.56857983	2.5605155	0.172427226	3.1914805	0.6033575
22	DLBCL39	0	3.8431000	2.1433000	1.8356739	0.08097867	3.8365590	1.10000167	2.3085409	0.154485881	6.8176069	0.6053092
23	DLBCL40	0	4.0540718	0.9718000	2.0008927	0.03425824	1.7190834	0.52093989	1.9393872	0.138675383	4.2750358	1.6775033
24	DLBCL41	0	7.8457000	1.1450000	2.3005989	2.29044360	4.7441721	2.42693159	3.4787438	0.268621040	7.9257428	3.6892817
25	DLBCL43	0	7.0267000	4.5234000	7.0872512	2.36344939	5.5959233	1.48390146	6.4513107	0.257976295	3.3552399	1.6913369
26	DLBCL44	0	6.5152000	2.2390000	5.8244509	2.28280585	4.8326412	1.08686554	6.4037539	0.343389264	5.4327984	1.0565314
27	DLBCL45	0	5.8242000	4.5002000	3.0801487	1.15791638	1.5977038	0.14328902	3.2243313	0.014487977	10.0991439	1.5167469
28	DLBCL46	0	6.2489000	2.2045538	3.0267738	2.14244665	5.2233592	0.18878053	3.2158632	0.047684322	31.9110170	1.0758014
29	DLBCL47	0	8.1574000	2.0235000	3.6778831	1.94380737	5.1952294	0.35511119	4.6840372	0.030358656	13.9058221	0.4841677
30	DLBCL48	0	18.0317000	4.3525523	7.1470295	3.58578386	6.6660890	2.58305646	6.8866428	0.047550316	12.8860764	6.3596556
31	DLBCL49	0	0.6115000	2.6027000	3.3741691	2.02030944	3.5794614	0.35812056	4.1182414	0.056228049	6.3380211	1.8531164
32	DLBCL50	0	5.3737000	1.9756170	2.6353257	1.61056679	3.9292937	0.41557038	2.6459400	0.069471507	5.6043489	0.5693595
33	DLBCL51	0	17.8169000	4.5871000	6.2869446	2.29544043	3.2939170	0.21093194	6.7810055	0.050737165	6.3960320	1.7046316
34	DLBCL52	0	8.1934000	3.7709000	5.0070983	3.31463276	4.1815314	1.06911757	4.7854068	0.095950083	10.5101376	1.6138044
35	DLBCL53	0	21.5628000	8.9374000	7.2777607	3.47523996	5.6717703	0.73155796	7.5209210	0.043845958	13.8682097	1.7325907
36	DLBCL54	0	13.0454000	2.9141000	3.7873273	2.10581775	5.7694870	1.56173179	3.8537483	0.256461086	7.3242899	1.1010748
37	FL15	1	4.0360161	0.5943000	1.0690581	1.68621109	2.2649121	0.47882692	1.2101351	0.247100000	11.1380055	4.0730337
38	FL16	1	7.5974992	3.4342617	2.7856230	1.88168463	2.2438870	0.42543187	2.5829147	0.419300000	6.2115529	0.7741265
39	FL17	1	1.0024411	0.6323361	0.5715787	0.48169282	0.3934361	0.15930964	0.5956399	0.147900000	3.5070310	1.4811703
40	FL18	1	4.9869207	0.9794790	1.2002229	2.06535998	1.8694182	0.93858343	1.3581069	0.128200000	5.8777049	1.3241119
41	FL19	1	2.4259070	0.9501149	0.8116140	1.22013606	1.4848810	0.38198264	0.9896202	0.100600000	9.3779086	1.2598828
42	FL20	1	5.7113620	1.0942937	1.2905625	1.03657313	3.6325937	0.56136071	1.5304968	0.028900000	8.2216710	1.3492035
43	FL21	1	1.7507380	0.5692970	0.6643525	0.72067667	0.7118029	0.09969749	0.6503363	0.124200000	1.2630399	1.1794723
44	FL22	1	7.7991244	2.1181352	2.4238617	1.69608688	2.1808443	0.49255932	2.6544286	0.153700000	8.0037658	2.0965518
45	FL23	1	1.7795578	1.0665291	0.9144947	0.63411207	0.6696522	0.14589110	0.8700438	0.082000000	1.5691641	1.9338641
46	FL24	1	1.5039671	0.8066418	0.7463892	0.99541890	1.3764956	0.54036270	0.8201729	0.403600000	3.5202009	0.7389723
47	FL25	1	1.5267503	0.5668563	0.6181966	0.79761300	0.8265809	0.33827529	0.6596443	0.080200000	3.9973487	1.1821083
48	FL26	1	2.4962716	0.8321987	1.2860975	0.84800898	1.2887746	0.25121595	1.1439310	0.114300000	3.8012904	1.0823668
49	FL27	1	4.3628391	1.6701758	1.9291964	1.15000168	3.8131942	0.33727563	1.7764536	0.064700000	13.3821303	1.3811619
50	FL28	1	2.7913040	1.6963447	1.1870209	1.55272110	0.9405556	0.28652361	1.3082682	0.450700000	4.3898755	1.3163579
51	FL29	1	12.2407851	3.0268034	2.6382598	1.63843273	2.6396616	0.40844980	2.9316828	0.080300000	9.9576297	3.2069528
52	FL30	1	6.5672656	2.3619853	2.7075726	1.64523719	2.5508888	0.52195596	2.5829147	0.363700000	6.6747308	1.4648078
53	FL31	1	3.7458592	2.0037901	2.3742918	1.42458642	2.7546931	0.34749612	2.4995895	0.269600000	3.9596123	0.6121382
54	FL55	1	9.9051750	3.1512725	3.3525193	3.47513444	5.7193541	0.84813435	3.7462075	0.878500000	20.0351222	0.5285909
55	SANO56	2	0.8000443	0.6241653	0.9780635	0.72204560	0.9401742	0.77217513	0.9686182	1.293248932	1.5896505	1.2556445
56	SANO57	2	0.9185518	1.2016361	0.7971925	0.83246162	0.7350937	1.59881166	0.7840405	0.545631939	0.6904462	1.6035000
57	SANO58	2	0.8287731	0.8950251	0.9579352	1.20950517	1.1534856	1.43594451	0.9037527	1.028826708	1.4196281	0.7719000
58	SANO59	2	0.6687546	0.8796491	0.8514537	0.63375055	0.7988519	0.89626670	0.9068903	0.936921447	0.3826731	0.4394000
59	SANO60	2	2.4551605	1.6934906	1.5724346	2.17042647	1.5702562	0.62937859	1.6065880	1.470187336	1.6771272	1.8384000

	miR18b	**miR19b**	**miR20a**	**miR92a**	**miR93**	**miR99**	**miR106a**	**miR150**	**miR210 NEW**	**miR135a**
miR18b	1.00000	0.69312	0.7538	0.6862	0.6163	0.2715	0.7807	-0.3291	0.4203825	0.26640
miR19b	0.69312	1.0000	0.8074	0.5119	0.5537	0.02682	0.8053	-0.17149	0.3975254	-0.24330
miR20a	0.7538	0.80745	1.0000	0.6735	0.7111	0.30589	0.9691	-0.25032	0.2994557	0.12970
miR92a	0.68622	0.51192	0.67353	1.0000	0.5757	0.31411	0.6634	-0.10535	0.5084182	0.136708
miR93	0.61633	0.55370	0.7111	0.5757	1.0000	0.45780	0.7711	-0.20363	0.6090577	0.15701
miR99	0.27154	0.02682	0.30589	0.3141	0.4578	1.00000	0.3213	0.21192	0.029192	0.39276
miR106a	0.78075	0.80536	0.9691	0.66342	0.7711	0.32137	1.0000	-0.25151	0.3523917	0.15500
miR150	-0.32910	-0.17149	-0.2503	-0.1053	-0.2036	0.21192	0.2515	1.0000	-0.2443	-0.05892
miR210 NEW	0.4203	0.39752	0.2994	0.5084	0.6090	0.02919	0.3523	-0.24430	1.0000000	0.11099
miR135a	0.2664	-0.24330	0.1297	0.1367	0.1570	0.39276	0.1550	-0.05892	0.1109901	1.0000

Tabella A.2: Matrice di correlazione tra le variabili esplicative

Pazienti DLBCL		Pazienti FL	
Id paziente	gene miR18b	Id paziente	gene miR18b
DLBCL1	1.4599397	FL15	1.395258096
DLBCL2	1.6935768	FL16	2.027819138
DLBCL3	1.2186151	FL17	0.002438149
DLBCL4	1.3773545	FL18	1.606818631
DBCL5	2.2141691	FL19	0.886205459
DLBCL6	1.0946043	FL20	1.742457516
DLBCL7	2.4196123	FL21	0.560037387
DLBCL8	2.2326346	FL22	2.054011469
DLBCL9	2.7020992	FL23	0.576364920
DLBCL10	1.9246573	FL24	0.408106324
DLBCL11	1.6960559	FL25	0.423141471
DLBCL12	1.1216776	FL26	0.914798256
DLBCL13	1.3679267	FL27	1.473123012
DBCL14	1.8459319	FL28	1.026508865
DBCL32	1.1404897	FL29	2.504773417
DLBCL33	1.9490054	FL30	1.882097548
DLBCL34	2.2239314	FL31	1.320651015
DLBCL35	2.3147903	FL55	2.293057353
DLBCL36	0.9999699		
DLBCL37	0.4267046		
DLBCL38	1.0925942		
DLBCL39	1.3462793		
DLBCL40	1.3997218		
DLBCL41	2.0599656		
DLBCL43	1.9497172		
DLBCL44	1.8741379		
DLBCL45	1.7620217		
DLBCL46	1.8324054		
DLBCL47	2.0989255		
DLBCL48	2.8921313		
DLBCL49	-0.4918403		
DLBCL50	1.6815167		
DLBCL51	2.8801474		
DLBCL52	2.1033290		
DLBCL53	3.0709696		
DLBCL54	2.5684356		

Tabella A.3: valori del marcatore miR18b nei due gruppi di malati

Pazienti DLBCL		Pazienti FL	
Id paziente	gene miR19b	Id paziente	gene miR19b
DLBCL1	0.21752781	FL15	-0.52037104
DLBCL2	0.75004723	FL16	1.23380198
DLBCL3	0.41680062	FL17	-0.45833421
DLBCL4	-0.05540701	FL18	-0.02073449
DBCL5	0.80114920	FL19	-0.05117239
DLBCL6	0.12486898	FL20	0.09010913
DLBCL7	-0.13559065	FL21	-0.56335302
DLBCL8	0.20449066	FL22	0.75053608
DLBCL9	0.51664864	FL23	0.06440958
DLBCL10	0.01587335	FL24	-0.21487563
DLBCL11	-0.06528547	FL25	-0.56764949
DLBCL12	-0.29463955	FL26	-0.18368400
DLBCL13	0.49907688	FL27	0.51292891
DBCL14	0.61788470	FL28	0.52847578
DBCL32	0.45520720	FL29	1.10750709
DLBCL33	0.64809751	FL30	0.85950250
DLBCL34	0.27376062	FL31	0.69504043
DLBCL35	1.57132059	FL55	1.14780635
DLBCL36	0.17806250		
DLBCL37	0.26390154		
DLBCL38	0.83772360		
DLBCL39	0.76234670		
DLBCL40	-0.02860526		
DLBCL41	0.13540464		
DLBCL43	1.50926392		
DLBCL44	0.80602934		
DLBCL45	1.50412184		
DLBCL46	0.79052512		
DLBCL47	0.70482869		
DLBCL48	1.47076242		
DLBCL49	0.95654937		
DLBCL50	0.68088073		
DLBCL51	1.52324802		
DLBCL52	1.32731370		
DLBCL53	2.19024472		
DLBCL54	1.06956102		

Tabella A.4: valori del marcatore miR19b nei due gruppi di malati

Pazienti DLBCL		Pazienti FL	
Id paziente	gene miR20a	Id paziente	gene miR20a
DLBCL1	0.1765653	FL15	0.06677794
DLBCL2	0.7319063	FL16	1.02447153
DLBCL3	0.8590711	FL17	-0.55935306
DLBCL4	0.3209271	FL18	0.18250725
DBCL5	1.0922606	FL19	-0.20873045
DLBCL6	0.4438080	FL20	0.25507816
DLBCL7	0.8918995	FL21	-0.40894240
DLBCL8	0.3763789	FL22	0.88536200
DLBCL9	1.4225863	FL23	-0.08938362
DLBCL10	0.7030524	FL24	-0.29250811
DLBCL11	0.2086982	FL25	-0.48094873
DLBCL12	-0.3306312	FL26	0.25161243
DLBCL13	1.5581949	FL27	0.65710353
DBCL14	1.1600606	FL28	0.17144676
DBCL32	0.7350929	FL29	0.97011952
DLBCL33	0.7333497	FL30	0.99605250
DLBCL34	1.2601416	FL31	0.86469919
DLBCL35	1.5148974	FL55	1.20971209
DLBCL36	0.2579839		
DLBCL37	0.5114934		
DLBCL38	0.9302027		
DLBCL39	0.6074116		
DLBCL40	0.6935934		
DLBCL41	0.8331695		
DLBCL43	1.9582976		
DLBCL44	1.7620647		
DLBCL45	1.1249779		
DLBCL46	1.1074973		
DLBCL47	1.3023374		
DLBCL48	1.9666968		
DLBCL49	1.2161491		
DLBCL50	0.9690068		
DLBCL51	1.8384752		
DLBCL52	1.6108566		
DLBCL53	1.9848232		
DLBCL54	1.3316606		

Tabella A.5: valori del marcatore miR20a nei due gruppi di malati

Pazienti DLBCL		Pazienti FL	
Id paziente	gene miR92a	Id paziente	gene miR92a
DLBCL1	0.448814182	FL15	0.522484052
DLBCL2	-0.006854689	FL16	0.632167454
DLBCL3	0.959430378	FL17	-0.730448677
DLBCL4	0.717708782	FL18	0.725304534
DBCL5	0.387910810	FL19	0.198962378
DLBCL6	0.234539572	FL20	0.035920209
DLBCL7	0.244879937	FL21	-0.327564690
DLBCL8	1.130048306	FL22	0.528323761
DLBCL9	1.099637402	FL23	-0.455529580
DLBCL10	0.627353432	FL24	-0.004591626
DLBCL11	0.202782093	FL25	-0.226131757
DLBCL12	-0.307559267	FL26	-0.164864048
DLBCL13	0.633432893	FL27	0.139763405
DBCL14	1.064944239	FL28	0.440008940
DBCL32	0.223102730	FL29	0.493740132
DLBCL33	1.083519626	FL30	0.497884562
DLBCL34	1.148426376	FL31	0.353881541
DLBCL35	0.983692642	FL55	1.245633165
DLBCL36	0.101874208		
DLBCL37	-0.441599629		
DLBCL38	0.371830837		
DLBCL39	-2.513569517		
DLBCL40	-3.373828013		
DLBCL41	0.828745511		
DLBCL43	0.860122160		
DLBCL44	0.825405321		
DLBCL45	0.146622168		
DLBCL46	0.761948471		
DLBCL47	0.664648613		
DLBCL48	1.276977100		
DLBCL49	0.703250685		
DLBCL50	0.476586159		
DLBCL51	0.830924733		
DLBCL52	1.198346838		
DLBCL53	1.245663528		
DLBCL54	0.744703874		

Tabella A.6: valori del marcatore miR92a nei due gruppi di malati

Pazienti DLBCL		Pazienti FL	
Id paziente	gene miR93	Id paziente	gene miR93
DLBCL1	0.30926374	FL15	0.81753594
DLBCL2	-0.35372468	FL16	0.80820961
DLBCL3	0.36063851	FL17	-0.93283655
DLBCL4	0.85672992	FL18	0.62562728
DBCL5	1.20335570	FL19	0.39533465
DLBCL6	1.19570159	FL20	1.28994690
DLBCL7	0.45929599	FL21	-0.33995422
DLBCL8	0.91911316	FL22	0.77971211
DLBCL9	1.04165852	FL23	-0.40099681
DLBCL10	0.54187395	FL24	0.31954085
DLBCL11	0.39686849	FL25	-0.19045745
DLBCL12	-0.12753908	FL26	0.25369187
DLBCL13	0.80127814	FL27	1.33846721
DBCL14	0.31211694	FL28	-0.06128457
DBCL32	0.59452241	FL29	0.97065072
DLBCL33	1.11319437	FL30	0.93644184
DLBCL34	0.74582637	FL31	1.01330603
DLBCL35	0.55122729	FL55	1.74385588
DLBCL36	0.70311774		
DLBCL37	0.08534632		
DLBCL38	0.62727664		
DLBCL39	1.34457587		
DLBCL40	0.54179123		
DLBCL41	1.55691693		
DLBCL43	1.72203835		
DLBCL44	1.57539314		
DLBCL45	0.46856749		
DLBCL46	1.65314071		
DLBCL47	1.64774078		
DLBCL48	1.89703333		
DLBCL49	1.27521235		
DLBCL50	1.36845970		
DLBCL51	1.19207744		
DLBCL52	1.43067754		
DLBCL53	1.73550130		
DLBCL54	1.75258318		

Tabella A.7: valori del marcatore miR93 nei due gruppi di malati

Pazienti DLBCL		Pazienti FL	
Id paziente	gene miR106a	Id paziente	gene miR106a
DLBCL1	0.1548093	FL15	0.19073199
DLBCL2	0.6304463	FL16	0.94891849
DLBCL3	0.6435282	FL17	-0.51811904
DLBCL4	0.4463868	FL18	0.30609176
DBCL5	1.2526806	FL19	-0.01043402
DLBCL6	0.5382511	FL20	0.42559237
DLBCL7	0.8740223	FL21	-0.43026561
DLBCL8	0.5157015	FL22	0.97622940
DLBCL9	1.4441072	FL23	-0.13921176
DLBCL10	0.8412702	FL24	-0.19824009
DLBCL11	0.3179547	FL25	-0.41605457
DLBCL12	-0.2918150	FL26	0.13447055
DLBCL13	0.9697129	FL27	0.57461901
DBCL14	1.1064688	FL28	0.26870427
DBCL32	0.8576311	FL29	1.07557658
DLBCL33	0.6370023	FL30	0.94891849
DLBCL34	1.1464654	FL31	0.91612653
DLBCL35	1.4495068	FL55	1.32074399
DLBCL36	0.4231147		
DLBCL37	0.4630933		
DLBCL38	0.9402086		
DLBCL39	0.8366157		
DLBCL40	0.6623720		
DLBCL41	1.2466713		
DLBCL43	1.8642833		
DLBCL44	1.8568844		
DLBCL45	1.1707256		
DLBCL46	1.1680958		
DLBCL47	1.5441604		
DLBCL48	1.9295837		
DLBCL49	1.4154262		
DLBCL50	0.9730264		
DLBCL51	1.9141254		
DLBCL52	1.5655710		
DLBCL53	2.0176886		
DLBCL54	1.3490463		

Tabella A.8: valori del marcatore miR106a nei due gruppi di malati.

Riferimenti Bibliografici

Azzalini, A. (2001). *Inferenza Statistica: Una Presentazione Basata Sul Concetto Di Verosimiglianza.* Springer Verlag.

Azzalini, A., Scarpa, B. (2004). *Analisi dei dati e data mining*. Springer Verlag.

Bortot, P., Ventura, L., Salvan, A. (2000). *Inferenza statistica: applicazioni con S-plus e R.* Cedam.

Bottarelli, E., Parodi, S. (2003). *Un approccio per la valutazione della validità dei test diagnostici: le curve R.O.C. (Receiver Operating Characteristic)*. Ann. Fac. Medic. Vet. di Parma, Vol. 23, pag. 49-68.

Groves, F. D., Linet, M. S., Travis, L. B. (2000). *Cancer surveillance series : non-Hodgkin's lymphoma incidence by histologic subtype in the United states from 1978 through 1995*. Biostatistics Branch.

Harrison. (2011). *Principi di medicina interna; 17° edizione*. McGraw-Hill Italia.

Krzanowski,W.J., Hand, D.J. (2009). *Roc Curves for Continuous Data*, Chapman & Hall, London.

Ott, G., Katzenberger, T., Lohr, A. (2002). *Cytomorfologic,immunohistochemical and cytogenetic profiles of follicular lymphoma : 2 types of follicular lymphoma grade 3.* Blood Journal, Vol. 99, pag. 3806-3812.

Pace, L., Salvan, A. (2001). *Introduzione alla statistica. Vol. 2: Inferenza, verosimiglianza, modelli*. Cedam, Padova.

Piccolo, D. (1998). *Statistica*, Il Mulino, Bologna.

Rachel, E., Stephen, J. (2010). *A 9 series microRNA signature differentiates between germinal centre and activated B-cell-like diffuse large B-cell lymphoma cell lines.* International Journal of Oncology, Vol. 37, pag. 367-376

Roehle, A. (2008). *MicroRNA signatures characterize diffuse large B-cell lymphomas and follicolar lymphomas.* British Journal of Haematology, Vol. 142, pag. 732–744.