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

Statistica per l'Economia e l'Impresa

Analisi statistiche per lo studio della Neurofibromatosi di Tipo 2 in dati di espressione genica derivanti da RNA-Seq

Relatrice: prof.ssa Giovanna Menardi
Dipartimento di Scienze Statistiche

Laureando: Ilenia Franco
Matricola n. 1220652

Anno Accademico 2021/2022

A Pier,
il più bel regalo
che mamma e papà
potessero farmi.

Indice

Introduzione

Nel corso degli anni, la statistica ha assunto un ruolo sempre più importante nella ricerca biomedica e epidemiologica, in particolare nello studio delle malattie genetiche, dove la necessità di estrarre conoscenza utile a fini diagnostici, prognostici e terapeutici si scontra con il problema di dover gestire ingenti quantità di dati.

Tra le malattie genetiche, la neurofibromatosi di tipo 2 è una patologia rara, caratterizzata dalla tendenza allo sviluppo di cellule tumorali. I tumori che comporta sono perlopiù benigni e tipicamente si tratta di meningiomi e schwannomi che si formano a livello di encefalo, midollo spinale, nervi cranici e nervi periferici. Nonostante la loro natura benigna, i tumori possono comprimere i nervi associati e causare disfunzioni nervose e pressioni intracraniche, per questo occorre procedere con la loro escissione per via chirurgica. Chi soffre di neurofibromatosi di tipo 2 può contare soltanto su terapie che alleviano i sintomi e sul monitoraggio delle complicanze, in quanto, al momento, non esiste alcuna cura che permetta la completa guarigione. L'aspettativa di vita per un soggetto affetto da questa malattia non raggiunge i quarant'anni.

L'obiettivo della relazione è quella di raccogliere in un unico documento informazioni generali di carattere biomedico riguardanti la malattia ed alcune evidenze statistiche ottenute in seguito alla lavorazione di dati di espressione genica relativi ai suoi due tumori principali, per i quali sono oggetto di studio diversi trattamenti. La relazione rappresenta inoltre un esempio di applicazione di metodi statistici appropriati per lo studio di dati genetici.

Nel Capitolo 1 sarà introdotto il contesto biologico di riferimento con una descrizione della malattia ed una panoramica sull'analisi di espressione genica rilevata mediante il sequenziamento dell'RNA. Inoltre, vengono presentati i dati e gli obiettivi dello studio.

Nel Capitolo 2 verranno presentate alcune analisi preliminari volte a preparare i dati su cui verranno effettuate le successive metodologie statistiche. Inoltre, viene fornita una prima analisi esplorativa dei dati.

Il Capitolo 3 è mirato all'illustrazione ed applicazione di tecniche statistiche per l'individuazione dei geni differenzialmente espressi.

Nel Capitolo 4 si valuterà l'effetto di alcuni trattamenti alternativi sull'espressione genica di pazienti affetti da neurofibromatosi di tipo 2.

Lo studio si conclude con una riflessione generale sui risultati ottenuti.

Mio fratello è uno dei pazienti che hanno subìto una mutazione *de novo* al seguito della quale è sorta la neurofibromatosi di tipo 2. Con questa relazione vorrei offrire un piccolo contributo al suo futuro.

Capitolo 1

Il contesto biologico di riferimento

Per comprendere al meglio i dati su cui verte questo studio, è necessaria un'introduzione generale al contesto biomedico di riferimento. Nel primo paragrafo ci si concentrerà sulla patologia in questione, la Neurofibromatosi, dandone una panoramica generale. Nel secondo paragrafo, invece, si introdurranno concetti prettamente biologici in merito alla cellula, al DNA, ai geni, all'espressione genica e al sequenziamento dell'RNA.

1.1 La Neurofibromatosi

1.1.1 Caratteri generali della patologia

La neurofibromatosi (NF) è una malattia genetica facente parte delle facomatosi, un gruppo di malattie ereditarie (quali anche la sclerosi tuberosa e alcune angiomatosi) caratterizzate dalla tendenza allo sviluppo di formazioni neoplastiche (cellule anomale, tumorali) a lento accrescimento. Questa facomatosi è a trasmissione ereditaria autosomica dominante, ossia è sufficiente che si alteri anche solo una copia del gene coinvolto (carattere dominante), localizzato su una delle 22 coppie di cromosomi (detti autosomi). Generalmente i soggetti affetti nascono da unioni tra un genitore malato e uno sano. Nonostante ciò, sono molto alti i tassi delle mutazioni *de novo*, ossia non ereditate da nessuno dei genitori, e ciò comporta anche un più alto rischio di formazione di tumori.

La neurofibromatosi è comparsa per la prima volta nel 1882, quando il patologo tedesco Friedrich Daniel von Recklinghausen descrisse

una serie di pazienti aventi una combinazione di tumori e lesioni cutanee nel sistema nervoso periferico e centrale. Solo verso la fine del XX secolo essa fu classificata in due sottotipi geneticamente distinti: la neurofibromatosi di tipo 1, chiamata malattia di Recklinghausen, e la neurofibromatosi di tipo 2, precedentemente denominata neurofibromatosi acustica o centrale. Di seguito viene fornita una breve descrizione per entrambi i sottotipi.

1.1.2 Neurofibromatosi di tipo 1

La neurofibromatosi di tipo 1 (NF1) ha un'incidenza alla nascita di circa 1 individuo su 2500/3500 e si presenta con una varietà di anomalie della pelle, delle ossa e del sistema nervoso periferico.

La sindrome è causata da mutazioni sul cromosoma 17q11.2 nel gene NF1 che codifica per la proteina *neurofibromina 1*, coinvolta nella trasduzione di segnali all'interno delle cellule.

Le manifestazioni cutanee sono solitamente i primi sintomi osservati nei pazienti con NF1. In tutti i casi esse includono macule iperpigmentate, denominate "Café-Au-Lait Macule", che si sviluppano principalmente durante l'infanzia e hanno all'incirca un diametro tra i 10 e i 40 millimetri. Altro tratto caratteristico del fenotipo cutaneo della NF1 è la presenza di lentiggini alle ascelle, all'inguine, intorno al collo, sulle palpebre e sotto il seno che si generano a partire dalle macule stesse.

Si manifestano inoltre disfunzioni ortopediche che sorgono a causa di anomalie nel mantenimento della struttura ossea dovute alla riduzione della densità minerale (BMD). Gli individui con NF1 hanno, quindi, un rischio maggiore di sviluppare osteopenia (lieve riduzione della BMD) e, in casi più gravi, anche osteoporosi (consistente riduzione della BMD): quest'ultime risultano molto dolorose perché colpiscono prevalentemente parti del corpo portanti come la zona lombare.

Diffuso in quasi tutti i pazienti, inoltre, è lo sviluppo di neurofibromi intorno o sui nervi periferici e di amartomi pigmentati nell'iride, i cosiddetti noduli di Lisch. I neurofibromi sono tumori prevalentemente benigni che si manifestano come lesioni cutanee e sottocutanee o come tumori plessiformi. Quest'ultimi sono generalmente asintomatici ma, in caso di prurito e bruciore intermittenti, la risposta agli antistaminici è scarsa e l'asportazione chirurgica è associata a cicatrici ipertrofiche. I neurofibromi sottocutanei, soprattutto quelli che generano pressione sul midollo spinale della zona cervicale alta causano, invece, dolore e *deficit* neurologici e occasionalmente subiscono alterazioni maligne.

La disabilità cognitiva, infine, è il sintomo neurologico più comune nei bambini affetti da NF1 e non vi sono miglioramenti in età adulta. Tipicamente, il quoziente intellettivo è nell'intervallo medio-basso, ma vi sono rari casi di ritardo mentale (ossia quoziente intellettivo inferiore a 70). I bambini con NF1 hanno difficoltà di apprendimento specifiche che includono problemi visivi spaziali, saccadi oculari anormali, *deficit* del linguaggio e disturbo delle funzioni esecutive (ovvero delle capacità di anticipare, progettare, stabilire obiettivi, attuare progetti finalizzati a uno scopo, monitorare e autoregolare il proprio comportamento per adeguarlo a nuove condizioni). Per approfondimenti sulla NF1 si vedano Gerber *et al.* (2009) e Ferner (2007).

1.1.3 Neurofibromatosi di tipo 2

La neurofibromatosi di tipo 2 (NF2), su cui si concentra questo lavoro, ha un'incidenza alla nascita di circa 1 individuo su 25.000, ossia inferiore alla NF1, e un'incidenza annuale di nuove diagnosi di uno su 2.355.000. La discrepanza tra le due sorge perché molti individui non sviluppano problemi clinici fino all'età adulta e altri, invece, muoiono prima che venga diagnosticata.

La sindrome è causata da mutazioni sul cromosoma 22q12.2 nel gene NF2 che codifica per la proteina merlina (anche nota come neurofibromina 2 o schwannomina). Si ritiene che la merlina agisca come una regolatrice della crescita, della motilità e del rimodellamento cellulare e che una sua mancata espressione causi un' iperproliferazione delle cellule di *Schwann* originando lo schwannoma, un tumore benigno incapsulato e attaccato ai nervi periferici. Proprio per questo, il primo segno clinico di NF2 è spesso un'improvvisa perdita dell'udito dovuta allo sviluppo di schwannomi vestibolari bi o unilaterali, ossia tumori che si verificano intorno ai nervi vestibolari che compongono entrambi i nervi uditivi.

Oltre agli schwannomi, i pazienti affetti da NF2 tendono a sviluppare ulteriori tumori del tessuto nervoso come meningiomi o gliomi ma, a differenza della NF1, essi sono uniformemente benigni. Tuttavia, i tumori possono comprimere i nervi associati e causare dolore considerevole, disfunzione nervosa e pressione intracranica. Oltre a quelli del tessuto nervoso, si verificano anche tumori spinali (i più devastanti e difficili da gestire) e tumori della pelle (di solito al di sopra della vita). Le stigmate cutanee sono meno pronunciate rispetto a quelle dei pazienti affetti da NF1, ma possono essere comunque un campanello di allarme per la presenza della malattia. Ad esempio, le macchie Café-Au-Lait

che caratterizzano il primo sottotipo, sono riportate anche nel 40% dei pazienti con NF2.

Il trattamento di tumori gravi e deturpanti viene solitamente eseguito chirurgicamente con tempi piuttosto lunghi che superano le sei ore. SAeppur benigna, l'escissione completa dei tumori rimane un grande sforzo terapeutico a causa della stretta associazione con i nervi. La rimozione totale di uno schwannoma vestibolare, ad esempio, provoca quasi spesso la sordità totale dell'orecchio colpito, la perdita della funzione cocleare e la disfunzione del nervo facciale. Una parziale soluzione è l'inserimento, per chi può, di un impianto cocleare nel tronco cerebrale, il quale migliora la lettura labiale e l'identificazione dei suoni ambientali. Infine però, va notato che l'asportazione del tumore comporta, comunque, un alto rischio di recidiva e non porta ad una cura permanente della malattia ma solo ad un allungamento della sopravvivenza. La durata media della vita rimane comunque a 36 anni.

Per ultieriori approfondimenti sulla patologia si vedano Riccardi (1981) e Asthagiri *et al.* (2009).

1.2 Introduzione all'analisi dell'espressione genica

1.2.1 Cenni di biologia genetica

L'Acido Desossiribonucleico (DNA) è una macromolecola presente nel nucleo delle cellule animali, costituita dal pentosio desossiribosio, dall'acido ortofosforico e da quattro basi azotate: adenina (A), guanina (G), citosina (C) e timina (T). Il DNA è costituito da due catene di polinucleotidi - unità formate dall'unione di una base azotata con il pentosio - che formano una doppia elica. Queste catene "antiparallele" hanno polarità opposta e le basi azotate, affacciate all'interno della catena, contribuiscono a mantenere rigida la struttura grazie alla formazione di legami idrogeno tra di esse, in particolare tra l'adenina e la timina e tra la citosina e la guanina. Il numero di molecole di DNA presenti nel nucleo delle cellule è fisso e corrisponde a quello dei cromosomi, ossia strutture filamentose costituite da una sola lunga molecola di DNA ripiegata su sé stessa.

I cromosomi sono presenti in numero, forma, grandezza costanti per ogni specie di animali e si presentano in coppie di elementi omologhi. Nel caso della specie umana, i cromosomi sono 23 e ognuno di questi è

presente in doppia copia, per un totale di 46. I cromosomi sono portatori dei geni, ossia segmenti di DNA che costituiscono l'unità di informazione ereditaria. Il complesso dei geni di un organismo è chiamato genoma. Un gene è un'entità stabile, ma è soggetto a cambiamenti di sequenza occasionali, detti mutazioni genetiche, che possono portare a malattie genetiche quali la fibrosi cistica, l'emofilia e la sopracitata Neurofibromatosi. Oltre ad essere i vettori dell'eredità, i geni sono i protagonisti del processo di "espressione genica" di cui viene fornita una breve descrizione nel paragrafo successivo.

Il DNA è responsabile della sintesi dell'Acido Ribonucleico (RNA), una macromolecola biologica che funge da passaggio tra il DNA e le proteine. L'RNA ha la stessa struttura del DNA ma è formato da un'unica catena di polinucleotidi e contiene lo zucchero ribosio al posto del desossiribosio, la base azotata uracile (U) al posto della timina. L'RNA è coinvolto nella formazione delle proteine, macromolecole formate da una o più catene di aminoacidi di diverse specie, ossia composti organici costituiti da un insieme di tre basi azotate (detto codone). Essendoci quattro basi, i possibili codoni sono $4^3 = 64$, ma gli aminoacidi risultano circa venti: ciò significa che la maggior parte degli aminoacidi è codificata da più di un codone. Il corpo umano utilizza le proteine per ogni attività svolta al suo interno, quali il trasporto di ossigeno, la sintesi degli ormoni, il corretto funzionamento degli anticorpi e il mantenimento di ossa, cartilagine, organi, pelle, capelli e unghie. In poche parole, le proteine sono essenziali per la crescita ed il mantenimento delle cellule e dei tessuti umani.

1.2.2 Espressione genica e RNA-Seq

L'espressione genica è il processo attraverso cui il DNA viene copiato in RNA e vengono sintetizzate le proteine. Tale processo è l'insieme di due fasi essenziali: la trascrizione, in cui si ha la formazione dell'RNA messaggero (mRNA) a partire da una sequenza DNA stampo, e la traduzione, in cui si ha la sintesi della proteina da parte di uno specifico mRNA. Per ogni gene del DNA possono essere create molte copie di RNA, e quindi di proteine.

In termini molecolari il genoma è identico in tutte le cellule di un individuo, ma solo una piccola frazione di esso compone anche il trascrittoma. Quest'ultimo, insieme al proteoma, cambia a seconda delle attività della cellula: si parla, dunque, di "espressione genica differenziale" in

quanto soltanto una percentuale limitata del genoma viene espressa in ogni cellula per la sintesi dei prodotti specifici di un determinato tessuto. I geni non utilizzati nelle cellule differenziate non vengono distrutti (o mutati), ma mantengono il loro potenziale di essere espressi.

Esistono due tecnologie differenti per misurare l'espressione genica di un dato campione: i *microarray*, che usano un approccio basato sull'ibridazione del DNA, e l'RNA-Seq, basato sul sequenziamento del DNA.

Dagli anni '90 fino ai primi anni 2000, i *microarray* sono stati la tecnologia preferita per gli studi ad alto rendimento sull'espressione genica. Sebbene abbiano consentito ai ricercatori di avere un quadro globale della cellula a livello molecolare, non raggiungibile con le tecnologie precedentemente disponibili, i *microarray* presentano diversi limiti: i livelli di ibridazione di fondo limitano l'accuratezza delle misurazioni dell'espressione; forniscono una misura relativa dell'espressione piuttosto che una stima assoluta dell'abbondanza di una trascrizione; il design della sonda limita i ricercatori a studiare solo trascrizioni note. I successivi miglioramenti nell'efficienza, nella qualità e nel costo del sequenziamento dell'intero genoma hanno quindi spinto i biologi ad abbandonare rapidamente i *microarray* a favore del sequenziamento ad altissima velocità, noto anche come *Next-Generation Sequencing* (NGS). Il NGS, di cui fa parte il sequenziamento dell'RNA, ha rivoluzionato il modo in cui i ricercatori eseguono gli studi genomici: sono aumentate in modo significativo la quantità e la qualità dei dati generati per ogni ripetizione e contemporaneamente si sono ridotti il costo e la durata per ottenere la risposta. Il NGS è quindi in grado di produrre in poco tempo milioni di brevi sequenze (dette *reads*, tipicamente di 25-100 basi) per ogni campione.

Una delle tecniche di RNA-Seq è il sequenziamento *Illumina*, di cui viene fornita di seguito una breve panoramica non tecnica in quanto i dati di cui si parlerà successivamente in questa relazione proverranno proprio da tale piattaforma. Per approfondimenti si veda Bullard *et al.* (2010).

Un campione di interesse viene sottoposto alla preparazione di una libreria, ossia a una serie di passaggi per convertire l'RNA di input in piccoli frammenti di DNA pronti per essere sequenziati dalla macchina *Illumina*. In particolare, a partire da qualsiasi campione di RNA, il protocollo di preparazione della libreria mRNA-Seq di *Illumina* include la frammentazione dell'RNA, la trascrizione inversa in cDNA tramite primer casuali, la selezione della dimensione, l'amplificazione e

l'arricchimento PCR (per permettere il sequenziamento).

Successivamente alla preparazione della libreria sorge un ulteriore problema: nessuna tecnologia moderna è in grado di leggere l'intera sequenza di un gene (ossia circa 10.000 basi), ma solo piccole sequenze di circa 100 basi in grandi quantità. Non è quindi così facile capire da quale gene provengano le *reads* ottenute. Il primo passo per la quantificazione dell'espressione genica dai dati RNA-Seq è, dunque, quello di allineare (o mappare) le letture al genoma di riferimento, ottenendo le cosiddette "*reads* allineate" o "mappate". A tale scopo, sono stati prodotti numerosi algoritmi di cui è possibile leggerne un approfondimento sulla piattaforma LANGMEAD LAB[1] prodotta dalla Johns Hopkins University.

Una volta compiuto l'allineamento sul genoma, si procede con il contare le *reads* che mappano su ogni gene e tali conteggi rappresentano i livelli di espressione genica. Il fatto che due o più conteggi riferiti allo stesso gene e allo stesso campione biologico, a ripetizioni diverse, risultino difformi, è dovuto alla possibilità di ottenere delle distorsioni nella fase di selezione dei trascritti nella preparazione della libreria. È bene ricordare che l'RNA-Seq non dà una misura assoluta del numero di molecole di RNA presenti in un campione ma solamente una misura relativa di espressione. Queste misure hanno un valore solo se confrontate con altre misure ottenute in maniera simile durante lo stesso esperimento.

1.3 Presentazione dei dati e degli obiettivi

1.3.1 Dati di espressione genica da RNA-Seq

Ai fini di un'analisi statistica, il risultato dell'operazione di sequenziamento descritta è una matrice, esemplificata nella Tabella 1.1, in cui ogni cella riporta l'espressione genica. Le righe sono costituite dai geni e le colonne corrispondono ai campioni biologici. Quest'ultimi dipendono dal confronto che si vuol fare, quindi ad esempio potrebbero esserci confronti tra cellule sane e tumorali, tra individui diversi, tra linee cellulari diverse, tra trattamenti diversi, ecc. In aggiunta ai dati si hanno spesso a disposizione informazioni sui geni e sui campioni biologici, che prendono il nome di metadati. Solitamente il numero di righe p è molto più grande del numero di colonne n. Se si guardasse alla tabella con le solite convenzioni note in statistica secondo cui le unità statistiche ven-

[1] https://langmead-lab.org/teaching-materials/

	Exp 1	Exp 2	...	Exp n
Gene 1				
Gene 2				
...				
Gene p				

Tabella 1.1: Formato tipico in cui si presentano i dati di espressione genica. Ogni colonna è riferita ad un certo campione biologico.

gono disposte per riga e le variabili per colonna, allora una matrice con $p >> n$ risulterebbe un'ottima situazione. In questo contesto, però, le variabili sono costituite dai p geni e le unità statistiche dagli n campioni biologici. Una delle difficoltà nel trattamento di questi tipi di dati è, dunque, l'elevata dimensionalità di p rispetto a n.

A partire da tale matrice le analisi statistiche sono tipicamente volte a rispondere a domande quali:

- Vi sono geni che mostrano un'espressione molto alta per alcuni campioni biologici rispetto ad altri, ossia che sono differenzialmente espressi?
- Quali campioni biologici sono più simili tra loro in termini di espressione dei geni?
- Vi sono trattamenti che influenzano l'espressione genica e in che modo?

Le risposte a tali domande risulteranno poi utili a biologi, medici e ricercatori. Ad esempio, la scoperta di geni differenzialmente espressi tra cellule sane e tumorali può essere usata sia per la diagnosi del tumore sia per indirizzare la ricerca verso farmaci e trattamenti che agiscano proprio su quei geni. Inoltre, la ricerca di campioni biologici più o meno simili tra loro in termini di espressione dei geni potrebbe essere d'aiuto nel momento in cui si vogliono valutare cellule non trattate con quelle trattate oppure cellule appartenenti a tessuti diversi.

1.3.2 I dati Synodos NF2

I dati su cui si lavorerà in questa relazione sono stati presi dall'*NF Data Portal*[2], un portale creato per aiutare a esplorare e condividere pubbli-

[2]https://nf.synapse.org/

LC	Tipo di LC	Stato di Merlina
Syn1	Meningioma	Wildtype
Syn2	Meningioma	Wildtype
Syn3	Meningioma	Assente
Syn4	Meningioma	Assente
Syn5	Meningioma	Assente
Syn6	Meningioma	Assente
Syn10	Meningioma	Assente
HS01	Schwannoma	Assente
HS11	Schwannoma	Wildtype

Trattamento
untreated
DMSO
Panobinostat
CUDC.907
GSK2126458

Tabella 1.2: Linee cellulari e Trattamenti.

camente set di dati, strumenti di analisi, risorse e pubblicazioni relative alla Neurofibromatosi. Lo studio in questione è il *Synodos NF2*[3], una collaborazione di ricerca dedicata a sconfiggere la Neurofibromatosi di tipo 2 che riunisce un *team* multidisciplinare di scienziati provenienti da dodici laboratori di primo ordine con medici di eccellenza. L'obiettivo finale di tale studio è trovare nuovi approcci alla diagnosi e al trattamento di due tumori primari correlati alla NF2: il meningioma e lo schwannoma. L'aspettativa è quella di scoprire informazioni utili per lo sviluppo di nuove terapie farmacologiche efficaci. Per maggiori informazioni su tale studio si veda Allaway *et al.* (2018).

I dati in questione per questa relazione sono costituiti da due dataset, uno in cui sono state selezionate cellule aracnoidali per il tumore meningioma e uno in cui sono state selezionate cellule di *Schwann* per il tumore schwannoma, entrambi riferiti a organismi umani. Ognuno di essi è presentato come in Tabella 1.1, dove ciascuna riga rappresenta uno specifico gene e ciascuna colonna corrisponde ad una coppia formata da una linea cellulare (LC) e un trattamento, secondo lo schema indicato in Tabella 1.2.

Per quanto riguarda i trattamenti, la condizione *untreated*, presente solamente nel dataset del meningioma, indica che non è presente alcun trattamento, mentre *DMSO* è un composto organico che viene inserito quando si vuol misurare l'espressione genica in cellule contenenti particolari farmaci, quindi può comunque essere considerato come una condizione di riferimento di non trattamento. I farmaci studiati per entrambi i tumori sono, invece, *CUDC.907*, *Panobinostat* e *GSK2126458*.

[3]https://nf.synapse.org/Explore/Studies/DetailsPage?studyId=syn2343195

Per rappresentare la biologia del tumore meningioma sono state selezionate 7 linee cellulari (Tabella 1.2, a sinistra). La Tabella 1.1 è qui formata da $p = 56.736$ geni e $n = 74$ combinazioni LC-trattamento. Queste 74 colonne sono state pulite togliendo le LC non di interesse per lo studio in quanto riferite a specie animali e non confrontabili con le altre LC. Per rappresentare la biologia del tumore schwannoma, invece, sono state selezionate 2 linee cellulari (Tabella 1.2, a sinistra). La Tabella 1.1 è qui formata da $p = 28.262$ geni e $n = 26$ combinazioni LC-trattamento. In questo caso non è stata praticata alcuna operazione di pre-trattamento dei dati del dataset. Lo stato di merlina specificato per ogni linea cellulare, sia del meningioma che dello schwannoma, indica la presenza (*Wildtype*) o meno (*Assente*) della proteina merlina, ossia rispettivamente l'assenza o presenza della NF2.

I valori dell'espressione genica sono stati rilevati tramite il sequenziamento *Illumina* illustrato al paragrafo 1.2.2, tuttavia il dataset riferito allo schwannoma è stato creato in un ambiente diverso e con un kit differente rispetto a quello riferito al meningioma e per questo il confronto diretto non è praticabile.

1.3.3 Obiettivi dello studio

In questa relazione, a partire dai dati appena descritti, ci si porrà inizialmente l'obiettivo di ricercare i geni differenzialmente espressi. Tale ricerca avverrà confrontando di volta in volta l'espressione genica relativa a due diversi trattamenti e prendendo per ciascun confronto i geni che si esprimono in modo statisticamente differente, in direzione sia di un trattamento che dell'altro.

Una volta eliminati i geni non differenzialmente espressi dalle tabelle relative al meningioma e allo schwannoma, si passerà alla valutazione di se e come i diversi trattamenti influenzino l'espressione genica. In particolare, si vorrà sapere se essi hanno lo stesso effetto o meno su entrambi i tipi di tumori e su ambedue gli stati di merlina. La speranza è quella di giungere alla conclusione che i trattamenti agiscano allo stesso modo per entrambi i tumori in modo tale che l'utilizzo di uno di essi su un paziente affetto da meningioma e schwannoma non rischi di agire solamente su uno dei due, rischiando invece di peggiorare l'altro. L'ulteriore aspettativa è quella di osservare una risposta differente in base allo stato di merlina, in quanto l'obiettivo è quello di trovare terapie farmacologiche efficaci a sconfiggere entrambi i tipi di tumore quando questi sono causati dalla NF2.

Le analisi condotte per raggiungere gli obiettivi menzionati sono state svolte nell'ambiente di programmazione R (R Core Team, 2022) con l'ausilio del pacchetto edgeR (Chen *et al.*, 2020), che include un insieme di *routine* per l'analisi statistica di dati genomici provenienti da RNA-Seq.

Capitolo 2

Analisi preliminari

2.1 Filtraggio

L'analisi parte considerando il fatto che i geni con conteggi molto bassi in tutte le librerie (ossia, nel caso di studio, i campioni biologici) dovrebbero essere rimossi per motivazioni sia biologiche che statistiche. Dal punto di vista biologico, un gene deve essere espresso a un livello minimo prima che possa essere tradotto in una proteina ed essere biologicamente importante. Dal punto di vista statistico, conteggi troppo bassi non forniscono evidenza sufficiente per formulare un giudizio affidabile sul successivo numero di geni differenzialmente espressi ma, anzi, interferiscono con alcune delle approssimazioni statistiche che verranno utilizzate in seguito. Tali geni "poco importanti" quindi possono essere rimossi dall'analisi senza alcuna perdita di informazione.

Solitamente un gene, per essere considerato espresso in una libreria, deve avere un conteggio in essa almeno pari a 5-10 ma non esistono soglie di filtraggio ottimali universali per tutti gli studi. Il filtraggio, però, è bene che avvenga sui conteggi per milione (CPM) piuttosto che sui conteggi in quanto altrimenti non si terrebbe conto delle differenze tra le dimensioni delle varie librerie. Le dimensioni (*library size*) sono date dal numero totale di letture (*reads*) generate per una data libreria e il CPM riferito ad un gene in una dato campione biologico è dato dal rapporto tra l'espressione genica corrispondente e la dimensione del campione biologico stesso. Con riferimento alla Tabella 1.1, le *library size* corrispondono ai totali di colonna mentre il CPM è dato dal rapporto tra una cella e il relativo totale di colonna.

Una strategia di filtraggio molto diffusa in letteratura basata sui CPM è quella descritta da Chen *et al.* (2016) e qui di seguito esposta. Sia:

DATASET	geni iniziali	N	l	r	geni filtrati
Meningioma	56.736	48,57	0,20	1	29.751
Schwannoma	28.262	31,09	0,30	1	12.249

Tabella 2.1: Operazione di filtraggio sui dati NF2.

- N la dimensione in milioni minima tra tutte le librerie;
- l l'approsimazione del rapporto $10/N$, detto *cutoff*;
- r il numero minimo di repliche avvenute in ciascun campione biologico.

Nel filtraggio vengono mantenuti i geni con CPM superiore a l in almeno r librerie.

La Tabella 2.1 riporta i dettagli principali sull'operazione di filtraggio eseguita sui due dataset oggetto di studio, descritti nel capitolo precedente. A seguito dell'operazione di filtraggio sono stati, dunque, rimossi 29.751 geni dal dataset relativo al meningioma e 12.249 da quello relativo allo schwannoma, pari a circa il 52.4% e rispettivamente 43.3% dei geni inizialmente valutati.

Con un abuso di notazione, si manterrà nel seguito l'uso di p ad indicare il numero di geni considerati, sebbene a seguito del processo di filtraggio ne sia stato appunto selezionato un sottoinsieme.

2.2 Normalizzazione

2.2.1 Giustificazioni alla normalizzazione

Come già accennato nel paragrafo precedente, vi sono casi in cui la rilevazione su una data coppia LC-trattamento è stata eseguita più di una volta, producendo così un certo numero di repliche utili allo scopo di ridurre il più possibile gli errori sistematici nel rilevamento dei conteggi. Al fine di rendere omogeneo il numero di rilevazioni, una pratica comune (Chen *et al.*, 2016) consiste nel riassumere l'informazione contenuta in più repliche mediante la loro somma. Tuttavia, poiché il numero di repliche non è coerente al variare delle coppie LC-trattamento, una volta eseguita la somma non è più possibile confrontare tra loro le varie librerie e occorre procedere con una normalizzazione del dataset.

Oltre a permettere la confrontabilità tra campioni biologici, la normalizzazione minimizza gli errori sistematici sorti durante le misurazioni,

ossia quegli errori dettati dall'imprecisione degli strumenti, dei metodi, delle persone e delle operazioni con i quali si ha operato. In questo modo le differenze tra due misurazioni relative allo stesso gene ma a librerie differenti rappresentano esclusivamente le differenze biologiche tra i due campioni biologici stessi.

2.2.2 Tecniche di normalizzazione

Sia Y_{ji} l'espressione genica del gene j nel campione biologico i. La normalizzazione consiste nella divisione di ogni Y_{ji} per un fattore di scala S_i specifico per ciascuna libreria.

Nella normalizzazione *Upper-Quartile* (UQ) (Abbas-Aghababazadeh *et al.*, 2018) i conteggi non vengono divisi per la *library size* ma per il quartile superiore della distribuzione empirica della libreria corrispondente, ossia quel valore $S_i^{(UQ)}$ tale che

$$P(Y_{ji} \leq S_i^{(UQ)}) \geq 0,75.$$

In questo modo, si evita di utilizzare i conteggi dei geni troppo espressi nel calcolo dei fattori di scala.

Nella normalizzazione *Relative Log Expression* (RLE) (Anders e Huber, 2010) il fattore di scala $S_i^{(RLE)}$ viene calcolato come mediana del rapporto tra i conteggi di una libreria e quelli di una pseudo-libreria di riferimento data dalla media geometrica tra le librerie:

$$S_i^{(RLE)} = Med_i \frac{Y_{ji}}{(\prod_{v=1}^{n} Y_{jv})^{1/n}}.$$

Il metodo *Trimmed Mean of M-values* (TMM) di normalizzazione parte dal presupposto che la maggior parte dei geni non sia espressa in modo differenziale, come accade nella maggior parte dei conteggi provenienti da RNA-Seq.

Esso poggia su un modello per dati di sequenziamento in cui si assume che il numero medio di *reads* per il gene j nel campione biologico i sia dato da:

$$E[Y_{ji}] = \frac{\mu_{ji} L_j}{T_i} N_i$$

dove:

- μ_{ji} è il livello di espressione genica vero e ignoto;
- L_j è la lunghezza in nucleotidi del gene j;

- $T_i = \sum_{j=1}^{J} \mu_{ji} L_j$ è la quantità totale di RNA nella libreria i;
- N_i è il numero totale di *reads* per la libreria i.

Il problema alla base di questo modello è che mentre N_i è noto, la quantità T_i è sconosciuta e può variare da libreria a libreria a seconda della composizione dell'RNA, quindi risulta difficile da stimare poiché non si conoscono i livelli di espressione e le lunghezze reali di ogni gene. Se una popolazione ha una quantità totale di RNA maggiore rispetto ad un'altra libreria, gli esperimenti di RNA-Seq finiscono per sottocampionare molti geni (ossia ridurre il numero di *reads* di interesse).

Tuttavia, è relativamente più semplice, per ogni coppia di campioni biologici (i,k), stimare la quantità relativa di RNA $\frac{T_i}{T_k}$ mediante una media pesata troncata dei valori di log-rapporto (*Trimmed Mean of M-values*). A tale scopo si definiscono i *log*-rapporti

$$M_j = \log_2 \frac{Y_{ji}/N_i}{Y_{jk}/N_k}$$

e i livelli di espressione assoluta

$$A_j = \frac{1}{2} \log_2 (Y_{ji}/N_i \cdot Y_{jk}/N_k).$$

Nel caso di più di due campioni, si considera come campione di riferimento il campione k e si calcola il fattore TMM per ciascun altro campione i non di riferimento.

L'obiettivo della normalizzazione TMM è quello di trovare un riassunto robusto di M_j tramite una media troncata e pesata. Per fare ciò, quindi, si inizia troncando i geni con valori estremi sia di M_j sia di A_j (di solito il 30% dei primi e il 5% dei secondi). Successivamente, si procede con una media ponderata usando come pesi l'inverso di un'approsimazione della varianza dei *log*-rapporti. In questo modo si tiene conto del fatto che i valori di *log*-rapporto dei conteggi alti hanno una varianza minore di quella dei conteggi bassi. Più precisamente, il fattore di normalizzazione per il campione biologico i rispetto al campione biologico di riferimento i_0 è dato da

$$S_i^{(TMM)} = \frac{\sum_{j=J^*} \omega_{ji} M_{ji}}{\sum_{j=J^*} \omega_{ji}}, \tag{2.1}$$

dove

$$M_{ji} = \log_2(Y_{ji}/N_i) - \log_2(Y_{ji_0}/N_{i_0})$$

e

$$\omega_{ji} = \frac{N_i - Y_{ji}}{N_i Y_{ji}} + \frac{N_{i_0} - Y_{ji_0}}{N_{i_0} Y_{ji_0}},$$

per $Y_{ji} > 0$ e $Y_{ji_0} > 0$, e J^* rappresenta l'insieme dei geni che non sono stati troncati. Per maggiori approfondimenti si veda Robinson e Oshlack (2010).

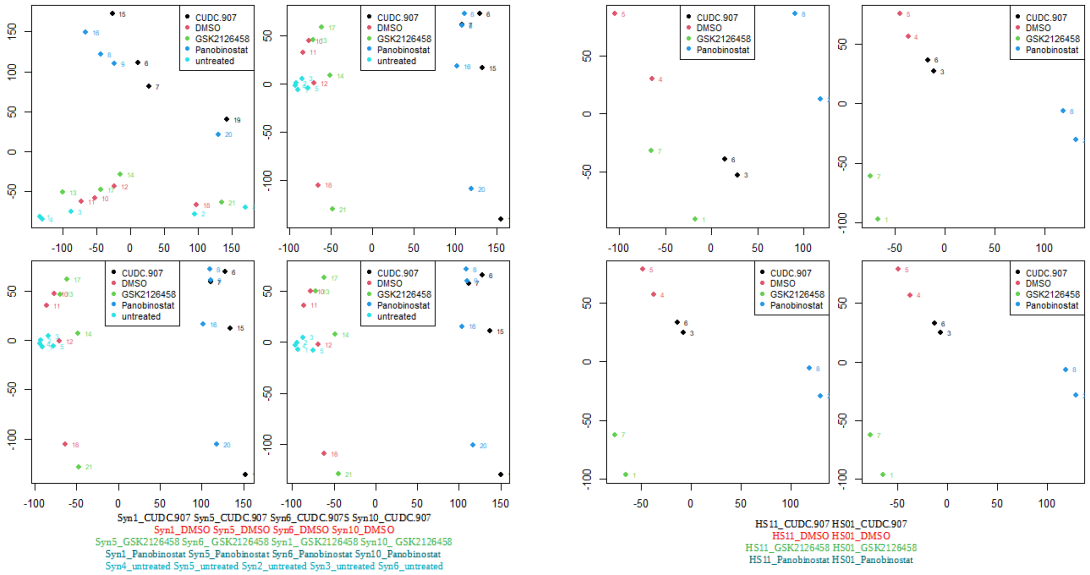
2.2.3 Confronto tra normalizzazioni nei dati NF2

La scelta del metodo di normalizzazione più adatto non è sempre univoca e dipende dal dataset a disposizione. Due strumenti particolarmente utili per il confronto e la successiva scelta della normalizzazione più adatta sono il RLE-*plot* (Gandolfo e Speed, 2018) e l'analisi delle componenti principali (PCA). Non viene utilizzato il classico *boxplot* perché risulterebbe non informativo in quanto si stanno forzando le distribuzioni dei dati ad essere le più simili possibile tra loro.

Sia $\log(Y_{ji})$ l'espressione logaritmica per il gene j nel campione biologico i e sia Y_{j*} la somma dei logaritmi delle espressioni geniche riferite al gene j in tutti gli n campioni biologici. Per ogni gene j viene calcolata la sua espressione mediana $Med(Y_{j*})$ nelle n librerie. Successivamente, vengono calcolate le deviazioni da questa mediana, ossia $log(Y_{ji}) - Med(Y_{j*})$. Infine, per ogni campione biologico, viene generato un *boxplot* di tutte le deviazioni per quella libreria. Come misura robusta si utilizza la mediana in quanto poco influenzata da valori anomali.

La PCA è una tecnica di riduzione della dimensionalità che aiuta nella descrizione e visualizzazione dei dati ad alta dimensionalità. Le componenti principali sono combinazioni lineari ortogonali delle variabili originali che, nell'ordine, minimizzano la varianza dei dati originali. L'idea è quella di proiettare i dati dallo spazio multidimensionale ad uno spazio bidimensionale, preservando il più possibile la struttura dei dati. Se le prime due componenti principali dopo la normalizzazione riflettono le differenze biologiche tra i dati, vuol dire che la normalizzazione è stata in grado di eliminare le distorsioni tecniche presenti senza ridurre l'informazione originariamente inclusa nei dati. In caso contrario, ci si trova di fronte a dati che presentano ancora molto rumore, casuale e/o sistematico. Per la valutazione della miglior normalizzazione, quindi, si utilizza il PCA-*plot* bidimensionale.

I PCA-*plots* ottenuti prima e dopo le normalizzazioni sui dati sulla NF2 sono riportati in Figura 2.1, sia per il dataset meningioma che per quello schwannoma. Tutte e tre le normalizzazioni hanno svolto il loro

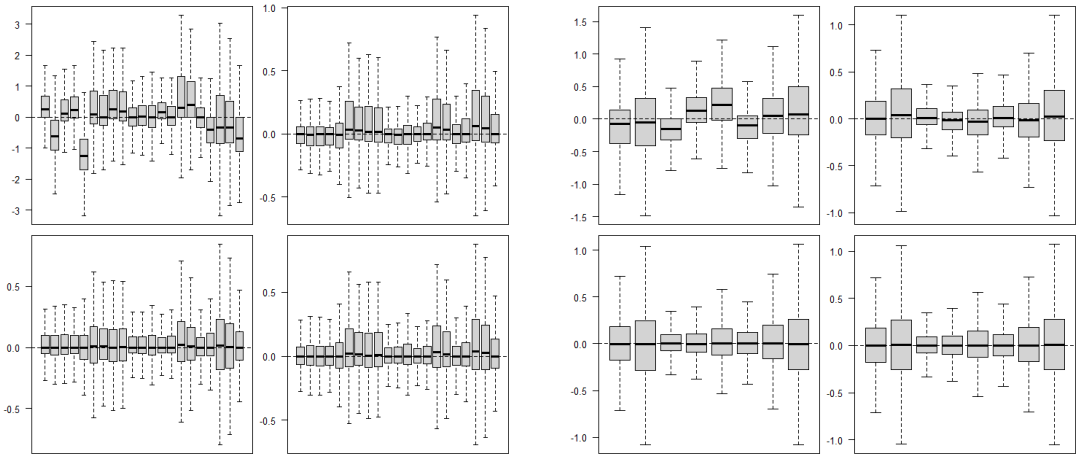


(a) Meningioma (b) Schwannoma

Figura 2.1: PCA-*plots* dei conteggi grezzi (in alto a sinistra) e dei dati normalizzati secondo le tecniche UQ (in alto a destra), RLE (in basso a sinistra) e TMM (in basso a destra), nei dataset meningioma e schwannoma.

compito di riduzione della variabilità in entrambi i dataset in quanto si possono distinguere, in tutti e tre i PCA-*plots*, 4 gruppi ben distinti tra loro in entrambi i dataset. Inoltre, la percentuale di varianza spiegata dalle prime due componenti principali è compresa tra il 69% e il 71% in tutti i grafici. In questo caso non c'è una normalizzazione che prevale sulle altre in termini di marcata divisione in gruppi, quindi si procede guardando anche agli RLE-*plots*.

Gli RLE-*plots* ottenuti prima e dopo le normalizzazioni sui dati sulla NF2 sono riportati in Figura 2.2a per il dataset sul meningioma e in Figura 2.2b per quello sullo schwannoma. Anche da essi si nota che tutte e tre le normalizzazioni hanno svolto il loro compito di riduzione della variabilità in entrambi i dataset in quanto le mediane di ciascun *boxplot* risultano quasi perfettamente allineate sullo zero e in posizione centrale rispetto a ciascuna scatola. In questo caso le normalizzazioni TMM e RLE prevalgono sulla UQ in quanto le mediane sono meglio allineate sullo zero e i *boxplot* son più bilanciati. Per scelta, si decide di procedere con la normalizzazione TMM e i fattori di scala così ottenuti nei dati sulla NF2 sono riportati in Tabella 2.2.



(a) Meningioma (b) Schwannoma

Figura 2.2: RLE-*plots* dei conteggi grezzi (in alto a sinistra) e dei dati normalizzati secondo le tecniche UQ (in alto a destra), RLE (in basso a sinistra) e TMM (in basso a destra), nei dataset meningioma e schwannoma.

2.3 Analisi esplorative

Prima di procedere con la ricerca dei geni differenzialmente espressi, è bene analizzare preliminarmente i dati a disposizione tramite alcuni grafici.

In Figura 2.3 sono riportati i *boxplot* dell'espressione genica logaritmica riferiti al dataset meningioma, divisi per linea cellulare e per trattamento. Essendo il *DMSO*, come detto in precedenza, un composto organico che viene inserito quando si vuol misurare l'espressione genica in cellule contenenti particolari farmaci e, quindi, una condizione di riferimento di non trattamento, si è interessati a vedere se la sua presenza cambia o meno il valore dell'espressione genica. Per fare ciò occorre confrontare i *boxplot* riferiti ad *untreated* e a *DMSO*. Per quanto riguarda la linea cellulare *Syn1*, non si hanno valori riferiti ad *untreated* ma si possono usare quelli di *Syn2* in quanto *Syn1* e *Syn2* hanno le stesse caratteristiche in merito a tipo di tumore (meningioma) e stato di merlina (*wildtype*). Dai suddetti *boxplot* le distribuzioni empiriche di *Syn1-DMSO* e *Syn2-untreated* non presentano differenze. Con lo stesso ragionamento si confrontano tra loro *Syn10-DMSO*, *Syn3-untreated*, *Syn4-untreated*, *Syn5-DMSO*, *Syn5-untreated*, *Syn6-DMSO*, *Syn6-untreated*, in quanto linee cellulari riferite allo stesso tipo di tumore (meningioma) e aventi lo stesso stato di merlina (assente); anche in questo caso le distribuzioni empiriche non presentano differenze. Un

	$S_i^{(TMM)}$
Syn4-untreated	0.94
Syn5-untreated	0.89
Syn2-untreated	0.92
Syn3-untreated	0.92
Syn6-untreated	1.02
Syn1-CUDC.907	1.04
Syn5-CUDC.907	1.01
Syn1-Panobinostat	1.12
Syn5-Panobinostat	1.07
Syn1-DMSO	0.91
Syn5-DMSO	0.96
Syn6-DMSO	1.00
Syn5-GSK2126458	0.90
Syn6-GSK2126458	1.00
Syn6-CUDC.907	1.16
Syn6-Panobinostat	1.13
Syn1-GSK2126458	0.90
Syn10-DMSO	0.96
Syn10-CUDC.907	1.13
Syn10-Panobinostat	1.17
Syn10-GSK2126458	0.91

	$S_i^{(TMM)}$
HS11-GSK2126458	0.98
HS11-Panobinostat	1.02
HS11-CUDC.907	0.98
HS11-DMSO	0.95
HS01-DMSO	0.98
HS01-CUDC.907	1.02
HS01-GSK2126458	1.00
HS01-Panobinostat	1.07

Tabella 2.2: Fattori di normalizzazione per ciascuna libreria nei dataset meningioma (a sinistra) e schwannoma (a destra).

commento analogo può essere fatto per i livelli di espressione genica relativi ad ogni trattamento a seconda dello stato di merlina. Ciò viene confermato anche dalla Figura 2.4.

Analogamente, in Figura 2.5 sono riportati i *boxplot* dell'espressione genica logaritmica riferiti al dataset schwannoma, divisi per linea cellulare e per trattamento. In questo caso non è possibile eseguire un confronto tra *untreated* e *DMSO* perché la prima condizione non è presente per nessuna delle due linee cellulari disponibili. Nuovamente sembrerebbe che i livelli di espressione genica per ogni trattamento non siano influenzati dalla tipologia di linea cellulare e, quindi, dallo stato di merlina (*wildtype* per la *HS01* e assente per la *HS11*).

Per confrontare tra loro i diversi trattamenti, indipendentemente dalla linea cellulare, conviene spostare l'attenzione verso la Figura 2.6a, in cui i *boxplot* riferiti al meningioma non sono più divisi per linea cellulare

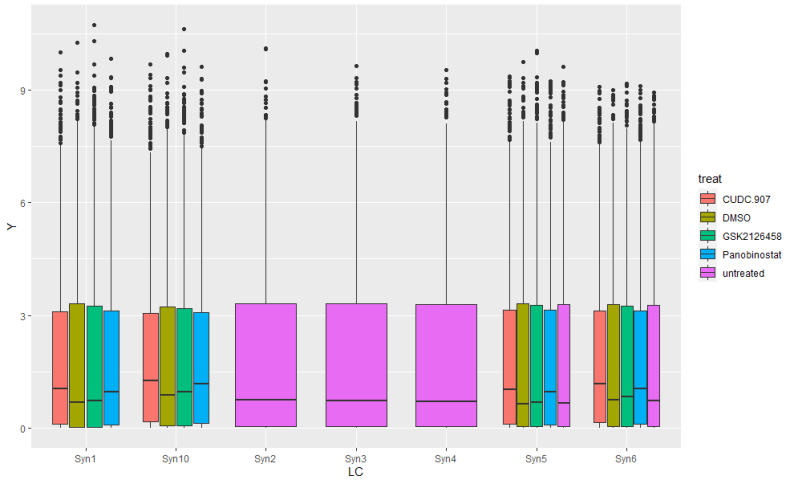


Figura 2.3: *Boxplot* dell'espressione genica logaritmica del dataset meningioma, divisi per linea cellulare e per trattamento.

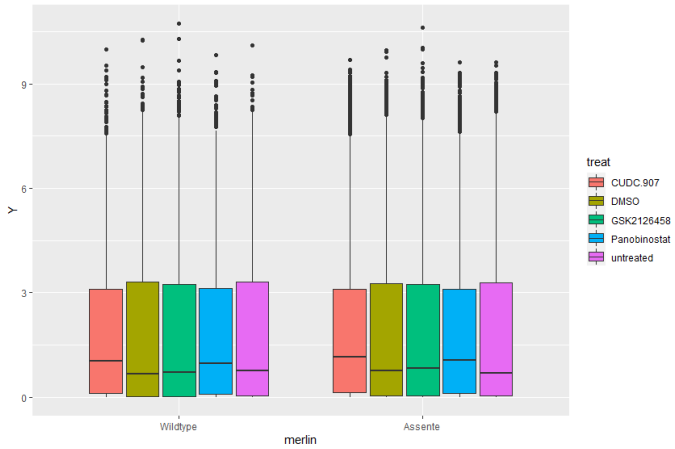


Figura 2.4: *Boxplot* dell'espressione genica logaritmica del dataset meningioma, divisi per stato di merlina e per trattamento.

ma solo per trattamento. Da essa i livelli mediani di espressione genica risultano piuttosto simili tra i vari trattamenti, con una mediana leggermente maggiore in corrispondenza di *CUDC.907* e *Panobinostat*, ma una variabilità che non suggerisce differenze potenzialmente significative.

Un discorso analogo può essere fatto per le distribuzioni riferite allo schwannoma (Figura 2.6b).

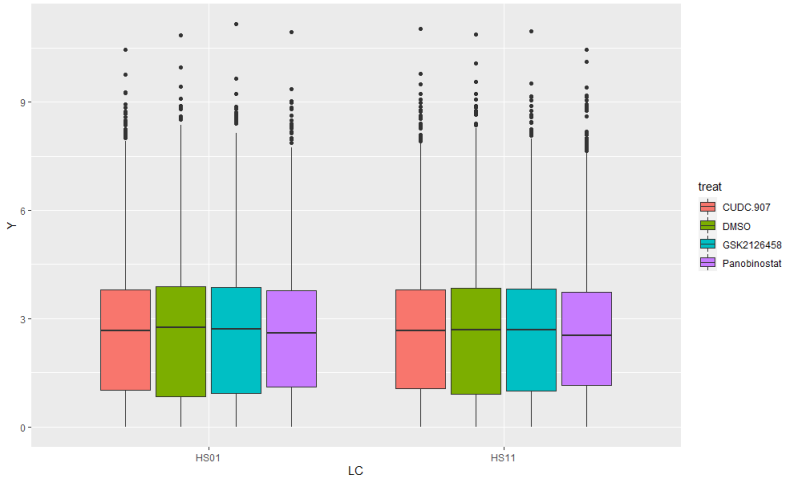
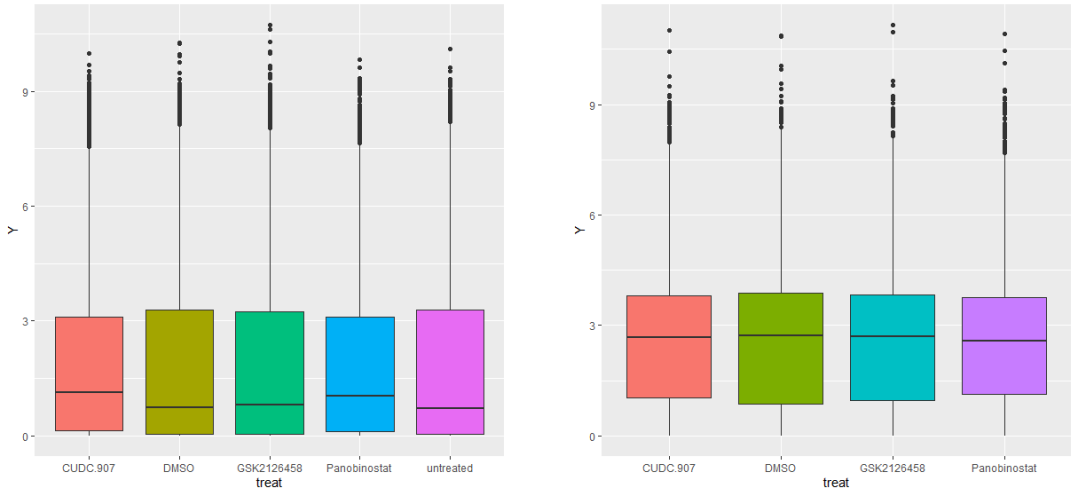


Figura 2.5: *Boxplot* dell'espressione genica logaritmica del dataset schwannoma, divisi per linea cellulare e per trattamento.



(a) Meningioma (b) Schwannoma

Figura 2.6: *Boxplot* dell'espressione genica logaritmica nei dataset meningioma e schwannoma, divisi per trattamento.

Capitolo 3

Ricerca geni differenzialmente espressi

3.1 Specificazione del modello

Una volta eseguite le operazioni di filtraggio e normalizzazione si tratta di individuare quali siano i geni che risultano differenzialmente espressi nel confronto tra una specifica coppia di trattamenti. A questo scopo la pratica comune, nell'analisi dei dati genomici, è quella di stimare, per ogni gene, un modello di regressione che mette in relazione l'espressione genica media con l'applicazione dei diversi trattamenti al variare dei campioni biologici e valutare la significatività dei coefficienti associati stimati.

È utile in questa fase modificare la notazione fin qui usata ai fini di maggior chiarezza e coerenza rispetto ad un tradizionale contesto di regressione. In luogo di Y_{ij}, espressione del gene j nel campione biologico i (ottenuto come combinazione tra linee cellulari e trattamenti, $i = 1, ..., n$), l'osservazione di riferimento sarà data dalle coppie

$$(Y_{lj}, T_{lj}) \quad l = 1, ..., L$$

dove T_{lj} rappresenta il trattamento applicato al gene j sulla linea cellulare l, con $T_{lj} \in \{T_1, ..., T_k, ...T_{K_l}\}$ e Y_{lj} l'espressione genica associata. Si avrà dunque che $LK_l = n$.

Poiché nel seguito si farà riferimento ad un modello per ogni singolo gene, per semplicità di notazione si ometterà l'indice j. Indicato allora con Y_l l'espressione genica e con μ_l il suo valore atteso, un'espressione generale per i modelli di riferimento sarà la seguente:

$$\mu_l = h(\tilde{\mathbf{x}}_l; \beta)$$

dove $\tilde{\mathbf{x}}_l$ è un vettore di contrasti associato al trattamento T_l, come approfondito in seguito. Per garantire una certa semplicità interpretativa, la specificazione della precedente ha tipicamente privilegiato formulazioni lineare e una funzione legame logaritmica:

$$\log \mu_l = \beta_0 + \tilde{\mathbf{x}}_l{}^\top \beta. \tag{3.1}$$

In merito alla scelta della distribuzione di Y, mentre in passato si è fatto ampio uso di modelli normali o, data la natura discreta di Y, Poisson (Marioni *et al.*, 2008), più di recente l'attenzione si è concentrata (Robinson e Smyth, 2007) su formulazioni che consentano di gestire la sovradispersione dei dati. Un modello distributivo per dati discreti che risponde a questa esigenza è quello binomiale negativo, la cui varianza cresce con il quadrato della media:

$$V(Y_l) = \mu_l^2 \phi + \mu_l$$

dove ϕ rappresenta il parametro di dispersione. Per ogni fissato valore di ϕ si può dimostrare che se $Y_l \sim NB(\mu_l, \phi)$, allora Y può essere espresso nella forma richiesta dalla famiglia esponenziale, e dunque la 3.1 è la formulazione di un modello lineare generalizzato.

Come evidenziato nel paragrafo 2.2, tuttavia, gli errori sistematici nel rilevamento dei conteggi e la conseguente necessità di produrre più repliche della medesima rilevazione, rendono i conteggi Y_i difficilmente confrontabili e richiedono una normalizzazione dei dati mediante la determinazione dei fattori di scala S_l.

Questa caratteristica può essere gestita in fase di modellazione mediante l'introduzione di un *offset*, vale a dire una variabile esplicativa dimensionale che entra nel modello con coefficiente unitario. La 3.1 può allora essere aggiornata come segue:

$$\log \mu_l = \beta_0 + \tilde{\mathbf{x}}_l{}^\top \beta + \log S_l \tag{3.2}$$

da cui, indicato con λ_l l'espressione genica media normalizzata, si ha:

$$\log \lambda_l = \log \left(\frac{\mu_l}{S_l} \right) = \beta_0 + \tilde{\mathbf{x}}_l{}^\top \beta. \tag{3.3}$$

Allo scopo di agevolare l'interpretazione dei parametri β, consentendo di valutare la differenza media di espressione genica al variare dei trattamenti, è prassi esprimere questi ultimi mediante un vettore di contrasti additivo, vale a dire

$$\log \lambda_l = \log \left(\frac{\mu_l}{S_l} \right) = \beta_0 + \beta_1 x_{l1} + ... + \beta_k x_{lk} + ... + \beta_{K-1} x_{lK-1}$$

con

$$x_{lk} = \begin{cases} 1 & se \quad T_l = T_{lk} \quad k = 1, ..., K-1 \\ -1 & se \quad T_l = T_{lk} \\ 0 & \text{altrimenti} \end{cases}$$

e $\beta_K = \sum_{k=1}^{K-1} \beta_k$.

Pertanto, quando viene applicato un generico trattamento k ($k = 1, ..., K-1$) il modello diventa

$$\log(\lambda_l | T_l = T_{lk}) = \beta_0 + \beta_k;$$

quando viene applicato il trattamento K invece

$$\log(\lambda_l | T_l = T_{lK}) = \beta_0 + \beta_K = \beta_0 - \sum_{k=1}^{K-1} \beta_k$$

mentre l'intercetta si interpreta in funzione dell'espressione genica al variare dei trattamenti

$$\log(\lambda_l) = \beta_0.$$

In altre parole, ciascun coefficiente β_k è da interpretarsi in funzione del differenziale di espressione genica media normalizzato rispetto al valore medio normalizzato al variare dei trattamenti

$$\begin{aligned} \beta_k &= \log(\lambda_l | T_l = T_{lk}) - \beta_0 \\ &= \log(\lambda_l | T_l = T_{lk}) - \log(\lambda_l) \\ &= \log\left[\frac{(\lambda_l | T_l = T_{lk})}{\lambda_l}\right] \end{aligned}$$

Questo modo di procedere consente di ricavare i differenziali di espressione genica media normalizzata relativi a due trattamenti k_1 e k_2, noti nell'analisi dei dati genomici come *log-fold-change* (*log*FC)

$$\begin{aligned} \log FC(k_1, k_2) &= \beta_{k_1} - \beta_{k_2} \\ &= \log\left[\frac{(\lambda_l | T_l = T_{lk_1})}{\lambda_l}\right] - \log\left[\frac{(\lambda_l | T_l = T_{lk_2})}{\lambda_l}\right] \\ &= \log\left[\frac{(\lambda_l | T_l = T_{lk_1})}{(\lambda_l | T_l = T_{lk_2})}\right] \end{aligned}$$

Nella pratica, trattandosi di confronti a coppie, è in realtà uso comune calcolare i *log*FC usando il logaritmo in base 2, in luogo del logaritmo naturale.

3.2 Stima dei parametri

3.2.1 Stima dei coefficienti di regressione

Come anticipato nel paragrafo precedente, per ogni fissato valore di ϕ, i modelli 3.1 - 3.2 - 3.3 rappresentano dei modelli lineari generalizzati con funzione legame logaritmica e variabile risposta distribuita secondo una legge binomiale negativa.

Pertanto, la stima dei parametri β viene agevolmente determinata mediante il metodo della massima verosimiglianza e l'applicazione dell'algortmo dei minimi quadrati pesati iterati (McCullagh e Nelder, 2019). In questo modo, si perviene alla determinazione di p stime dei vettori $\beta = \beta_j$, una per ogni gene.

3.2.2 Stima del parametro di dispersione

Meno agevole è, invece, la stima dei parametri di dispersione ϕ_j. Per farlo, sono stati proposti diverse modalità, tra cui la stima di massima verosimiglianza profilo aggiustata e la stima Bayesiana empirica basata sulla regressione sulla media.

Verosimiglianza profilo aggiustata di Cox-Reid

Dato che un set di dati RNA-seq ha spesso un esiguo numero di ripetizioni per campione, gli stimatori tradizionali di ϕ tendono a funzionare male. In particolare, lo stimatore di massima verosimiglianza è distorto negativamente in quanto tende a sottostimare i parametri di dispersione, non apportando alcun aggiustamento per il fatto che la media sia stimata dagli stessi dati.

La stima della dispersione per esperimenti di RNA-seq che coinvolgono molteplici condizioni di trattamento è stata studiata da Chen *et al.* (2014). Il loro metodo si basa sull'idea di verosimiglianza profilo aggiustata proposta da Cox e Reid (1987).

Ai fini della stima della dispersione, ϕ_j è il parametro di interesse mentre i coefficienti di regressione β_j e le medie μ_{lj} delle espressioni Y_{lj} sono parametri di disturbo. Nel metodo Cox-Reid si presuppone che gli stimatori dei parametri di disturbo siano ortogonali alla stima del parametro di interesse, cioè che la matrice informativa di Fisher sia diagonale a blocchi. Si può dimostrare che qui l'ortogonalità tra $\hat{\beta}_j$ e $\hat{\phi}_j$ deriva dal fatto che $\hat{\phi}_j$ compare solo nella funzione della varianza e non

anche in quella della media dei modelli lineari generalizzati binomiali negativi.

Per correggere la distorsione dello stimatore, la verosimiglianza profilo di Cox-Reid per ϕ_j viene penalizzata come segue:

$$l_C(\phi_j; \mathbf{Y}_{*j}, \hat{\beta}_j) = l(\phi_j; \mathbf{Y}_{*j}, \hat{\beta}_j) - \frac{1}{2} \log |I_j|,$$

dove $\mathbf{Y}_{*j}$ è il vettore L-dimensionale dei conteggi per il gene j, $\hat{\beta}_j$ è il vettore dei coefficienti di regressione stimato e $|I_j|$ indica il determinante dell'informazione attesa di Fisher di β_j valutato in $\hat{\beta}_j$ e ϕ_j.

Si noti che $\hat{\beta}_j$ è la stima di massima verosimiglianza di β_j dato ϕ_j. Quindi, $\hat{\beta}_j$ è a sua volta una funzione di ϕ_j. Ciò vuol dire che la log-verosimiglianza l può essere considerata come una verosimiglianza profilo l_p che dipende solo da ϕ_j, ovvero $l(\phi_j; \mathbf{Y}_{*j}, \hat{\beta}_j) = l_p(\phi_j; \mathbf{Y}_{*j})$, e lo stimatore di ϕ_j è dato dalla quantità che massimizza tale verosimiglianza profilo approssimata.

Shrinkage empirico di Bayes

L'approccio più semplice per condividere le informazioni tra tutti i geni consiste nell'assumere che essi abbiano tutti lo stesso valore di dispersione ϕ, chiamato dispersione comune (Robinson e Smyth, 2008), che può essere stimato massimizzando

$$l_C(\phi) = \frac{1}{p} \sum_{j=1}^{p} l_j(\phi), \tag{3.4}$$

dove p è il totale di geni disponibili nel dataset dopo l'operazione di filtraggio.

La dispersione comune è certamente la più semplice ma ovviamente non è realistica in quanto alcuni geni hanno maggior o minor valore di dispersione rispetto ad altri. È stato trovato in molti dataset di RNA-Seq (tra cui anche in quello relativo ai dati NF2, come si potrà notare in seguito) che geni con un basso livello di espressione tendono ad avere maggior dispersione, e viceversa. Quindi, usando una dispersione comune si finirebbe per sottostimare la dispersione dei geni con basso livello di espressione e a sovrastimare quelli con alto livello. Dunque, sembrerebbe ragionevole presumere che i valori di ϕ_j dipendano dai livelli di espressione genica e che possano essere modellati da un trend di dispersione media (Anders e Huber, 2010), dato dalla massimizzazione di $l_S(\phi_j)$, una verosimiglianza profilo localmente condivisa per il gene j.

Quest'ultimo approccio sarebbe sufficiente se le vere dispersioni seguissero il trend di quella media e se i geni con lo stesso livello di espressione avessero una dispersione identica. Tuttavia ciò accade molto raramente nei dataset reali in quanto le dispersioni sono gene-specifiche. Pertanto, si dovrebbe stimare una dispersione individuale per ogni singolo gene ma subentra un ulteriore problema: i dati di un singolo gene sono spesso insufficienti per una stima affidabile di questa dispersione. Quindi, si ha bisogno di un metodo che consenta a ciascun gene di avere la propria stima della dispersione pur ottenendo informazioni dagli altri geni. Ciò può essere ottenuto mediante un approccio empirico di Bayes che combina informazioni individuali e comuni per ottenere stimatori di dispersione stabili.

Il metodo empirico di Bayes ha l'obiettivo di stimare la distribuzione a priori dai dati e applicare l'approccio bayesiano standard per ottenere stime a posteriori. Tale approccio empirico diretto, però, non può essere applicato ai dati RNA-Seq perché non esiste un coniugato a priori per il parametro di dispersione della binomiale negativa.

Si può dimostrare, però, che uno stimatore empirico di Bayes è equivalente a una stima ottenuta massimizzando una funzione di verosimiglianza ponderata su un insieme di osservazioni (Wang, 2006). Tale risultato offre l'opportunità di implementare un'approssimazione del metodo empirico di Bayes per i dati RNA-seq (Robinson e Smyth, 2007).

Pertanto, la stima di ϕ_j è data dalla quantità che massimizza

$$l_W(\phi_j) = l_j(\phi_j) + \alpha_0 l_S(\phi_j) \tag{3.5}$$

dove $l_j(\phi_j)$ è la verosimiglianza profilo che usa informazioni provenienti solo dal gene j, $l_S(\phi_j)$ è la verosimiglianza profilo localmente condivisa per il gene j e α_0 è il peso associato alla $l_S(\phi_j)$.

In termini Bayesiani, $l_S(\phi_j)$ può essere interpretato come la distribuzione a priori per ϕ_j e $l_j(\phi_j)$ come la verosimiglianza proveniente direttamente dai dati osservati. Ciò significa che $l_W(\phi_j)$ può essere interpretato come la distribuzione a posteriori di ϕ_j, visto come un compromesso tra le due quantità.

La scelta ottimale di α_0 dipende dalla variabilità della dispersione. Si scelgono valori grandi di α_0 quando le dispersioni sono costanti tra loro e si vogliono "schiacciare" di più verso la dispersione media. Valori piccoli di α_0, invece, sono più adatti quando le dispersioni differiscono molto tra i diversi geni. Se $\alpha_0 = 0$ allora nessuna informazione viene presa in prestito dagli altri geni, ossia la dispersione gene-specifica per un particolare gene è puramente stimata dalla sua verosimiglianza profilo

l_j. Per maggiori approfondimenti sul calcolo di α_0 si vedano Robinson e Smyth (2007) e Chen *et al.* (2014).

3.3 Valutazione della significatività

Una volta stimati i parametri dei modelli considerati, è necessario valutare, per ciascun gene, se e quali trattamenti sono associati ad un'espressione significativamente diversa. Successivamente, saranno selezionati i soli geni in cui almeno un confronto risulta significativo.

L'ipotesi da sottoporre a verifica, per confrontare l'effetto del trattamento k_1 rispetto all'effetto k_2 sarà dunque la seguente:

$$H0 : \beta_{0j} + \beta_{k_1 j} = \beta_{0j} + \beta_{k_2 j}$$

contro l'alternativa bilaterale.

Tuttavia, in tale contesto, il numero di test effettuati, ciascuno ad un livello nominale α, è pari al numero p di geni e dunque, la probabilità di rifiutare almeno una ipotesi nulla quando tutte sono vere è data da $1-(1-\alpha)^p \approx 1$.. In altre parole, seguendo la prassi comune di rifiutare l'ipotesi nulla di assenza di espressione differenziale in corrispondenza di un *p-value* minore di α, non si controllerebbe la probabilità di commettere un errore di primo tipo

Questo problema può essere affrontato correggendo opportunamente i *p-value* per tenere conto della molteplicità. I metodi classici di correzione della molteplicità controllano, al posto del *p-value*, il *False Discovery Rate* (FDR), definito come la frazione attesa di falsi positivi tra le ipotesi che sono state dichiarate significative.

La procedura più semplice per il controllo del FDR è quella di Benjamini e Hochberg (1995), che, supposto avere p ipotesi nulle $H_1, ..., H_p$ e i corrispettivi *p-value* $\alpha_{01}, ..., \alpha_{0p}$, consiste in:

1. ordinare i *p-value* dal più basso al più alto: $\alpha_{0(1)}, ..., \alpha_{0(p)}$;
2. per un livello di significatività α, considerare il più grande k tale che
$$\alpha_{0(k)} \leq \frac{k}{p}\alpha;$$
3. rifiutare le ipotesi $H_{(1)}, ..., H_{(k)}$ e non rifiutare le altre.

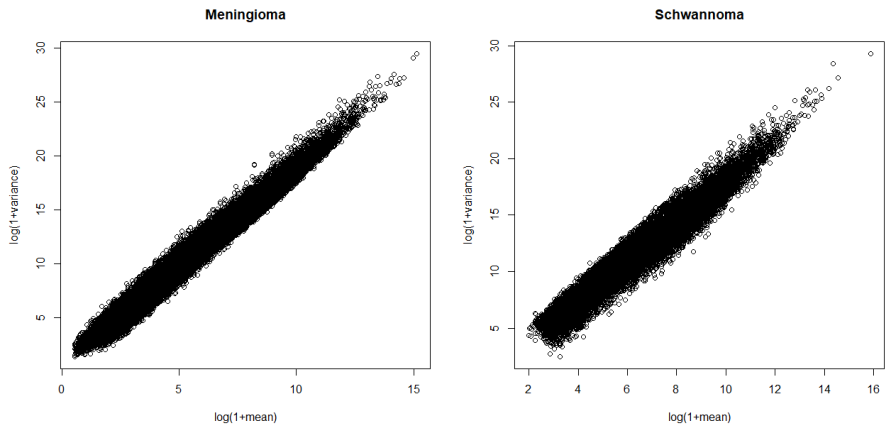


Figura 3.1: Media e varianza delle espressioni geniche riferite a ciascun gene presente nei dataset meningioma e schwannoma.

3.4 Ricerca geni differenzialmente espressi nei dati NF2

Per dare giustificazione all'utilizzo del modello binomiale negativo nei dati NF2, occorre guardare ai grafici media-varianza in Figura 3.1. Ciascun gene j viene rappresentato nel grafico tramite un punto di ascissa pari al logaritmo naturale della media delle espressioni geniche Y_{ij} $\forall i = 1, ..., n$ ad esso riferite e di ordinata pari al logaritmo naturale della varianza degli stessi valori; questo, per entrambi i dataset meningioma e schwannoma. Per ciascun gene si nota che il valore della varianza è sempre superiore alla media e la presenza di tale sovradispersione dà giustificazione valida all'utilizzo del modello binomiale negativo descritto nel primo paragrafo del corrente capitolo.

Per quanto riguarda la stima dei parametri di dispersione, nei dati NF2 viene utilizzato lo *shrinkage* empirico di Bayes. Un modo utile per dare giustificazione al fatto che l'adozione di una dispersione comune o di un trend di dispersione media non è realistica è l'utilizzo del grafico del coefficiente di variazione biologico, ossia della radice quadrata del parametro di dispersione stimato. In esso vengono riportati su uno stesso piano i valori della dispersione calcolata individualmente per ciascun gene massimizzando la 3.5, della dispersione comune data dalla massimizzazione della 3.4 e del trend di dispersione media dato dalla massimizzazione della verosimiglianza profilo $l_S(\phi_j)$ localmente condivisa per ciascun gene. Tali quantità vengono rappresentate rispettivamente tramite punti (*tagwise*), una linea rossa (*common*) e una linea blu (*trend*). In ascissa vengono riportati i logaritmi dei conteggi per

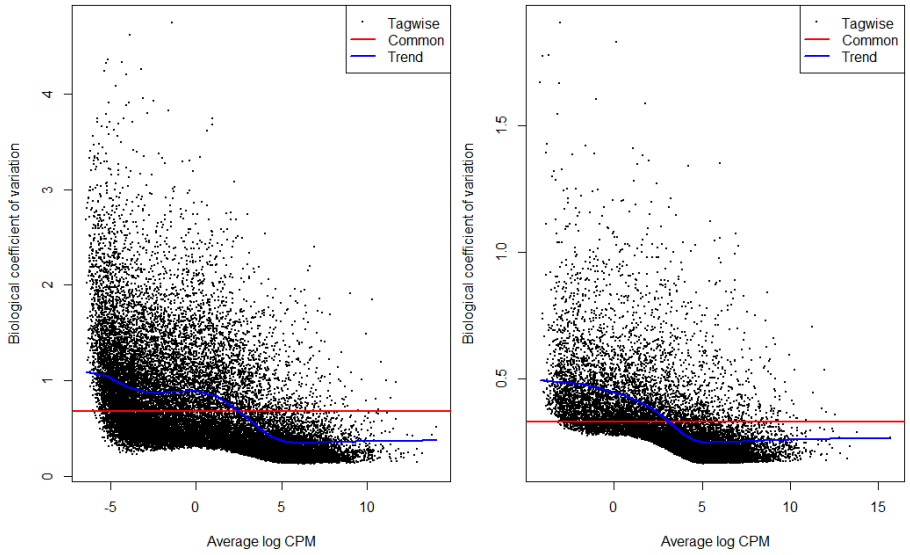


Figura 3.2: Grafico del coefficiente di variazione biologico riferito al dataset meningioma (a sinistra) e schwannoma (a destra). In rosso la stima della dispersione comune, in blu quella del trend di dispersione media e in nero le dispersioni gene-specifiche.

milione riferiti a ciascun gene: nella parte sinistra (risp. destra) del grafico ci si riferisce, quindi, ai geni con bassa (risp. alta) espressione genica. I grafici in Figura 3.2, sia del meningioma che dello schwannoma, confermano quanto detto in precedenza in merito alla dispersione comune, ossia che essa finirebbe per sottostimare la dispersione dei geni con basso livello di espressione e per sovrastimare quelli con alto livello. Inoltre, si osserva che, come innanzi spiegato, l'approccio del trend di dispersione media non risulta sufficiente in quanto le vere dispersioni non seguono tale trend e i geni con lo stesso livello di espressione non hanno dispersione identica. Dunque, dai grafici del coefficiente di variazione biologico si conferma che il metodo più indicato è la stima delle dispersioni gene-specifiche ricavate dalla massimizzazione della 3.5.

Date, quindi, le stime dei precedenti parametri di dispersione, i corrispettivi parametri di regressione stimati con il metodo di massima verosimiglianza sono riportati in Figura 3.3 per il tumore meningioma e in Figura 3.4 per lo schwannoma.

I geni differenzialmente espressi tra due librerie sono quei geni con valore di *False Discovery Rate* ≤ 0.01, dove la soglia 0.01 viene scelta pari al valore comune usato in letteratura. In Tabella 3.1 per il tumore meningioma e in Tabella 3.2 per lo schwannoma vengono riportati il numero di geni differenzialmente espressi per ciascun confronto possibile tra i vari trattamenti e il corrispondente valore percentuale rispetto al

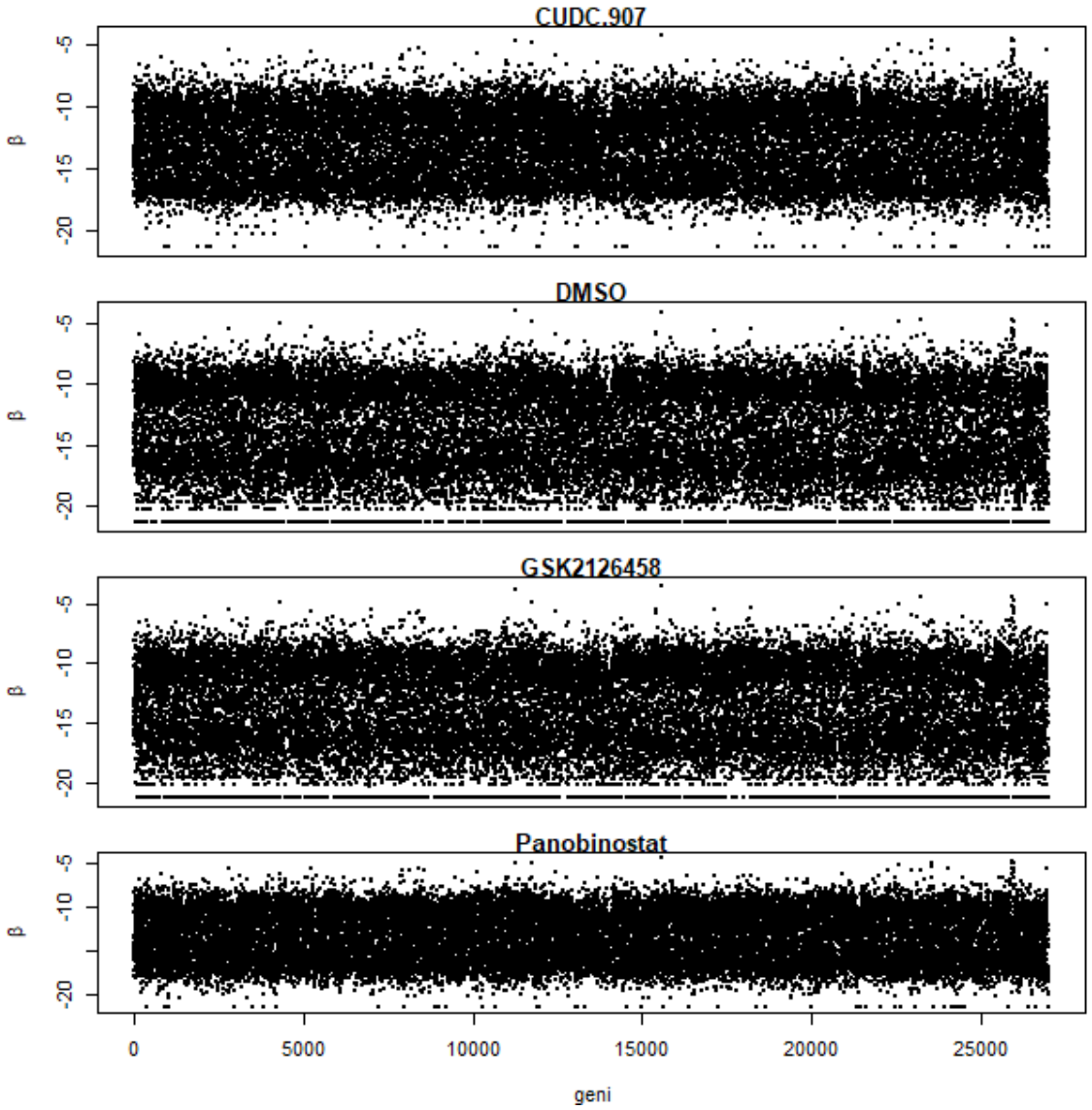


Figura 3.3: Coefficienti di regressione relativi a ciascun trattamento stimati per ogni gene contenuto nel dataset relativo al tumore meningioma.

numero di geni rimasti dopo l'operazione di filtraggio precedentemente descritta.

Una volta svolti i confronti tra tutte le librerie in entrambi i dataset meningioma e schwannoma, sono stati tenuti solamente i geni che sono risultati essere differenzialmente espressi in almeno uno dei confronti presenti nelle Tabelle 3.1 e 3.2. Per il dataset relativo al meningioma sono stati tenuti 14.916 geni, pari a circa il 55,3% di quelli rimasti dopo il filtraggio, mentre per quello riferito allo schwannoma ne sono stati tenuti 7.173, pari a circa il 44,8%.

Per vedere se i geni differenzialmente espressi hanno espressione genica maggiore per un trattamento piuttosto che per l'altro, è possibile

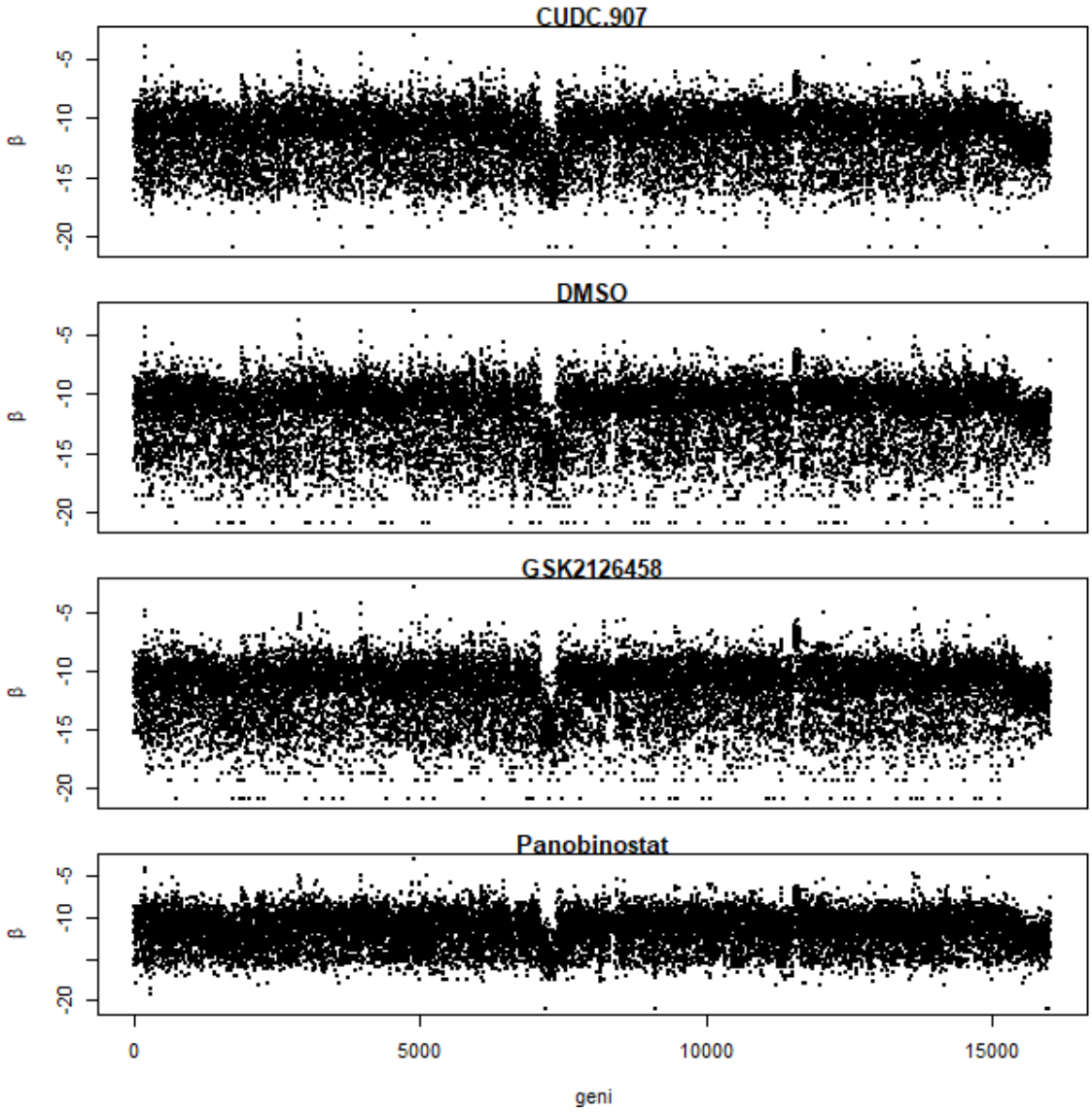


Figura 3.4: Coefficienti di regressione relativi a ciascun trattamento stimati per ogni gene contenuto nel dataset relativo al tumore schwannoma.

usare strumenti grafici come il *Volcano plot* e il *Mean Different plot*. In entrambi i grafici non solo vengono evidenziati quali geni hanno un *p-value* basso, ossia sono significativi dal punto di vista statistico, ma quali, tra essi, hanno anche un effetto alto, ossia che si distinguono rispetto agli altri per le loro caratteristiche biologiche.

Nelle Figure 3.5-3.14 e 3.15-3.20 sono riportati i *Volcano plot* e i *Mean Different plot* per i confronti tra i diversi trattamenti nel dataset relativo rispettivamente al meningioma e allo schwannoma. Concentrandosi solamente sulle Figure 3.5 e 3.15, si nota che per entrambi i tumori nel primo grafico ci sono più punti rossi nella parte destra del grafico rispetto a quella sinistra e nel secondo ci sono più punti rossi rispetto

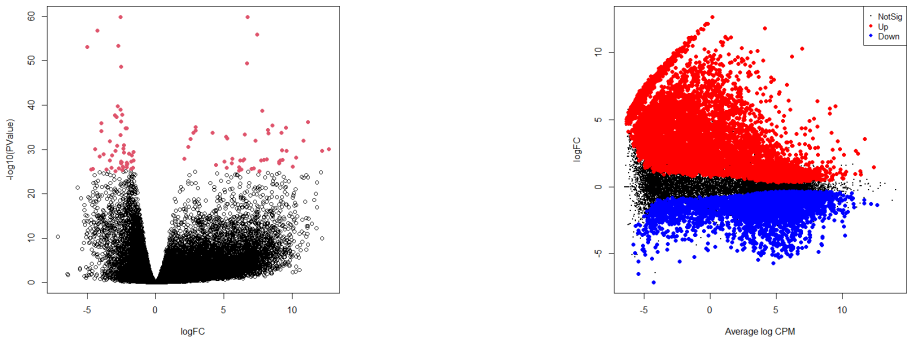
Campioni biologici	Geni DE	% Geni DE
CUDC.907 vs DMSO	9222	31.0%
CUDC.907 vs GSK2126458	8399	28.2%
CUDC.907 vs Panobinostat	4	0.01%
CUDC.907 vs untreated	11530	38.8%
DMSO vs GSK2126458	57	0.2%
DMSO vs Panobinostat	8829	29.7%
DMSO vs untreated	179	0.6%
GSK2126458 vs Panobinostat	8373	28.1%
GSK2126458 vs untreated	1222	4.1%
Panobinostat vs untreated	11127	37.4%

Tabella 3.1: Geni differenzialmente espressi (DE) per ciascun confronto tra campioni biologici nel dataset relativo al meningioma.

Campioni biologici	Geni DE	% Geni DE
CUDC.907 vs DMSO	895	5.6%
CUDC.907 vs GSK2126458	1526	9.5%
CUDC.907 vs Panobinostat	2674	16.7%
DMSO vs GSK2126458	1715	10.7%
DMSO vs Panobinostat	4506	28.1%
GSK2126458 vs Panobinostat	5128	32.0%

Tabella 3.2: Geni differenzialmente espressi (DE) per ciascun confronto tra campioni biologici nel dataset relativo allo schwannoma.

a quelli blu. Quindi, da entrambi i grafici si nota che ci sono più geni differenzialmente espressi con un effetto positivo, ossia nella condizione *CUDC.907* rispetto alla condizione *DMSO*, in entrambi i tipi di tumore. È possibile estendere tali osservazioni per ognuno dei confronti possibili tra trattamenti, sia per il dataset relativo al meningioma che per quello relativo allo schwannoma.

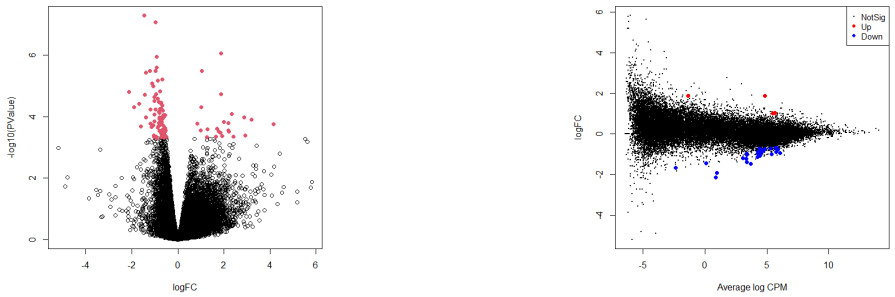


(a) *Volcano plot* (b) *Mean Different plot*

Figura 3.5: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *CUDC.907* e *DMSO* nel dataset relativo al meningioma.

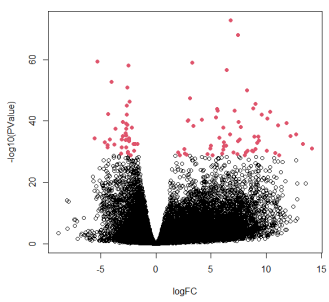
(a) *Volcano plot* (b) *Mean Different plot*

Figura 3.6: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *CUDC.907* e *GSK2126458* nel dataset relativo al meningioma.

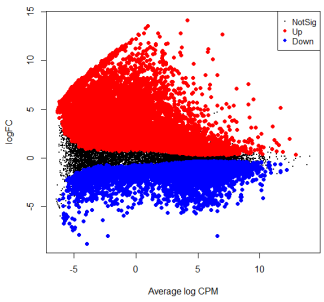


(a) *Volcano plot* (b) *Mean Different plot*

Figura 3.7: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *CUDC.907* e *Panobinostat* nel dataset relativo al meningioma.

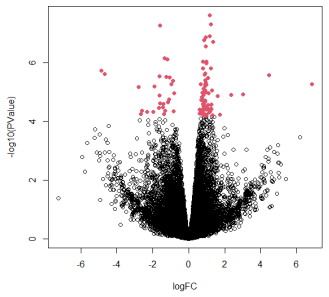


(a) *Volcano plot*

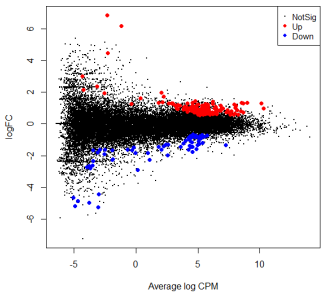


(b) *Mean Different plot*

Figura 3.8: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *CUDC.907* e *untreated* nel dataset relativo al meningioma.

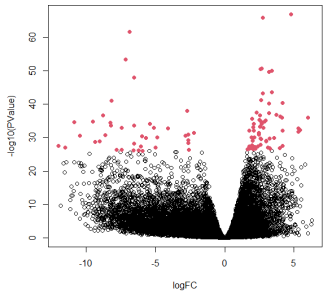


(a) *Volcano plot*

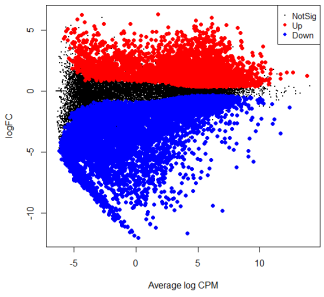


(b) *Mean Different plot*

Figura 3.9: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *DMSO* e *GSK2126458* nel dataset relativo al meningioma.



(a) *Volcano plot*



(b) *Mean Different plot*

Figura 3.10: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *DMSO* e *Panobinostat* nel dataset relativo al meningioma.

(a) *Volcano plot* (b) *Mean Different plot*

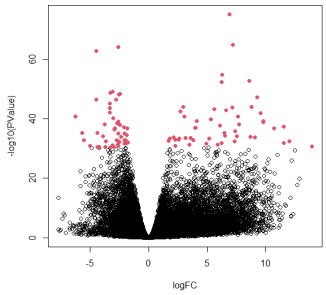
Figura 3.11: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *DMSO* e *untreated* nel dataset relativo al meningioma.

(a) *Volcano plot* (b) *Mean Different plot*

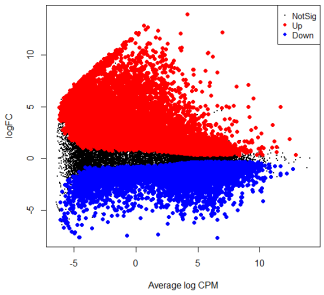
Figura 3.12: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *GSK2126458* e *Panobinostat* nel dataset relativo al meningioma.

(a) *Volcano plot* (b) *Mean Different plot*

Figura 3.13: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *GSK2126458* e *untreated* nel dataset relativo al meningioma.

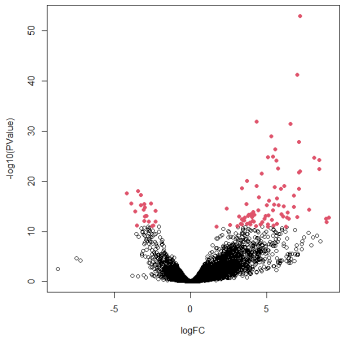


(a) *Volcano plot*

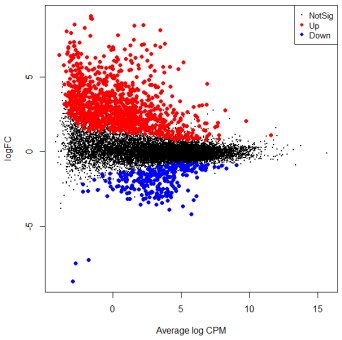


(b) *Mean Different plot*

Figura 3.14: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *Panobinostat* e *untreated* nel dataset relativo al meningioma.

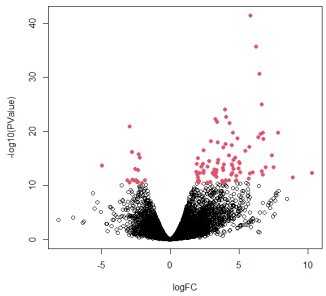


(a) *Volcano plot*

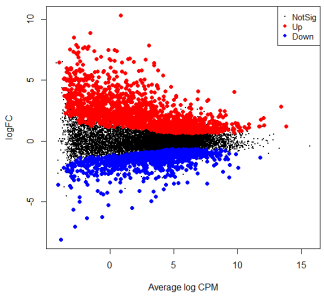


(b) *Mean Different plot*

Figura 3.15: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *CUDC.907* e *DMSO* nel dataset relativo allo schwannoma.

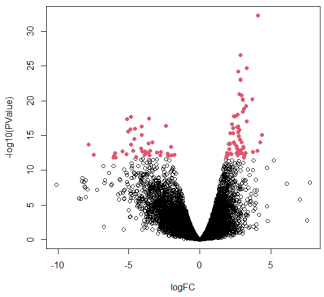


(a) *Volcano plot*

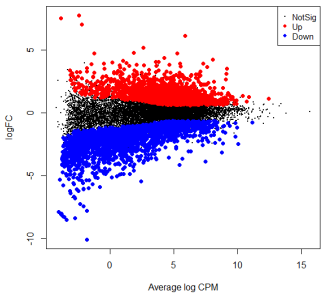


(b) *Mean Different plot*

Figura 3.16: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *CUDC.907* e *GSK2126458* nel dataset relativo allo schwannoma.

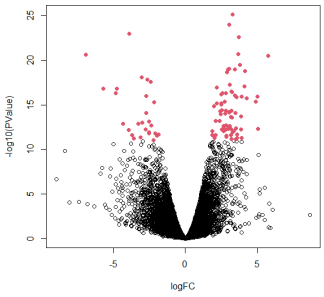


(a) *Volcano plot*

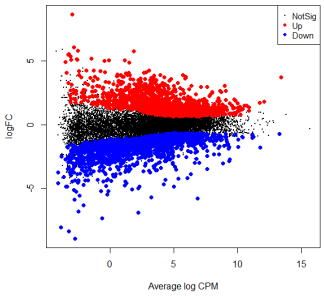


(b) *Mean Different plot*

Figura 3.17: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *CUDC.907* e *Panobinostat* nel dataset relativo allo schwannoma.

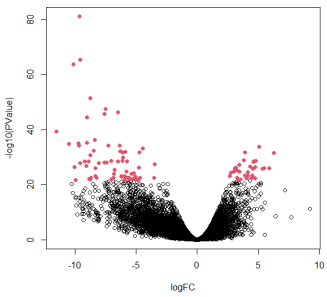


(a) *Volcano plot*

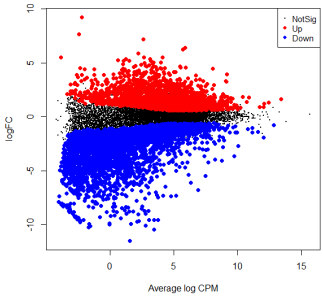


(b) *Mean Different plot*

Figura 3.18: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *DMSO* e *GSK2126458* nel dataset relativo allo schwannoma.



(a) *Volcano plot*



(b) *Mean Different plot*

Figura 3.19: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *DMSO* e *Panobinostat* nel dataset relativo allo schwannoma.

(a) *Volcano plot* (b) *Mean Different plot*

Figura 3.20: *Volcano plot* e *Mean Different plot* per il confronto tra i trattamenti *GSK2126458* e *Panobinostat* nel dataset relativo allo schwannoma.

Capitolo 4

Valutazione dell'effetto dei trattamenti

4.1 Formulazione del modello

4.1.1 Generalità

Una volta selezionati solamente i geni differenzialmente espressi in almeno uno dei confronti, si vuole verificare se e come il tipo di trattamento, lo stato di merlina e il tipo di tumore influenzino i valori dell'espressione genica.

A questo scopo i dati vengono riorganizzati in una matrice che riporta sulle righe le espressioni geniche riferite ai geni selezionati e sulle colonne le variabili esplicative di riferimento, ossia:

- il tipo di trattamento (categoriale con 5 livelli: *untreated*, *DMSO*, *Panobinostat*, *CUDC.907*, *GSK2126458*);
- lo stato di merlina (dicotomica con livelli *Wildtype* e Assente);
- il tipo di tumore (dicotomica con livelli Meningioma e Schwannoma).

Indicato come Y_l l'espressione genica e con μ_l il suo valore atteso, l'obiettivo viene dunque perseguito mediante la formulazione di un modello di regressione avente la seguente espressione generale:

$$\mu_l = h(\tilde{\mathbf{x}}_l; \beta) \tag{4.1}$$

dove $\tilde{\mathbf{x}}_l$ è il vettore di variabili esplicative e β il vettore di parametri. In questo caso le variabili esplicative sono descritte dalle seguenti variabili

indicatrici:

$$x_{l1} = \begin{cases} 1 & trattamento = CUDC.907 \\ 0 & altrimenti \end{cases}$$

$$x_{l2} = \begin{cases} 1 & trattamento = DMSO \\ 0 & altrimenti \end{cases}$$

$$x_{l3} = \begin{cases} 1 & trattamento = GSK2126458 \\ 0 & altrimenti \end{cases}$$

$$x_{l4} = \begin{cases} 1 & trattamento = Panobinostat \\ 0 & altrimenti \end{cases}$$

e dunque la modalità di riferimento, descritta dal vettore ($x_{l1} = 0, x_{l2} = 0, x_{l3} = 0, x_{l4} = 0$) indica che il gene non ha subito alcun trattamento;

$$x_{l5} = \begin{cases} 1 & tumore = Schwannoma \\ 0 & altrimenti \end{cases}$$

$$x_{l6} = \begin{cases} 1 & merlina = Assente \\ 0 & altrimenti \end{cases}$$

Poiché i geni selezionati sono 14.916 per il meningioma e 7.173 per lo schwannoma e, per ciascuno di essi, si hanno a disposizione rispettivamente 21 e 8 valori di espressione genica, la matrice così costituita risulta formata da $m = 14916 \cdot 21 + 7173 \cdot 8 = 370620$ righe e $d = 6$ colonne.

4.1.2 Analisi esplorative

Al fine di orientare la specificazione del modello 4.1 di riferimento, si riportano di seguito alcune analisi esplorative sulla variabile risposta presa sia singolarmente sia in funzione delle esplicative considerate. Occorre tenere presente che, affinché le espressioni geniche riferite ai tumori meningioma e schwannoma possano essere confrontate allo stesso modo, con il termine "espressione genica" si intenderà, per le successive analisi esplorative, il suo valore normalizzato con il metodo TMM e non il conteggio grezzo.

In Figura 4.1 viene riportata la distribuzione empirica della variabile risposta e della sua trasformata logaritmica. Da essa si può notare che,

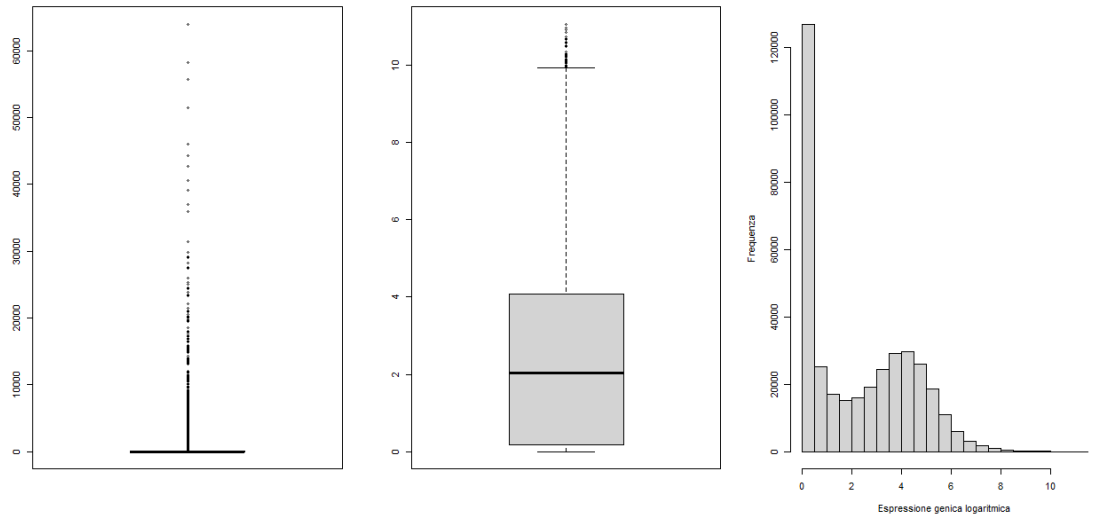


Figura 4.1: Distribuzione empirica dell'espressione genica (a sinistra) e della sua trasformata logaritmica (in centro e a destra).

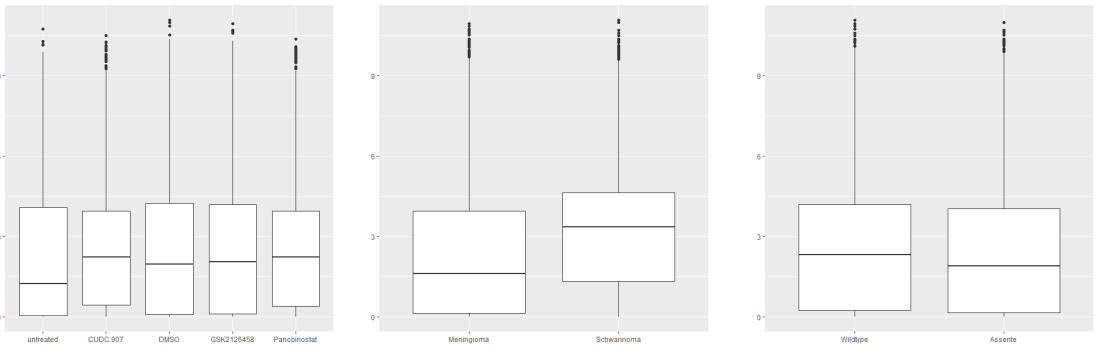


Figura 4.2: *Boxplot* dell'espressione genica logaritmica divisi per tipo di trattamento (a sinistra), per tipo di tumore (in centro) e per stato di merlina (a destra).

mentre il logaritmo consente di ridurre l'asimmetria della distribuzione, anche a seguito della trasformazione permane una struttura bimodale con un elevato numero di espressioni pari a zero, che dovrà essere opportunamente gestita.

In Figura 4.2 sono riportati i *boxplot* dell'espressione genica logaritmica divisi per tipo di trattamento, per tipo di tumore e per stato di merlina. Da notare che rispetto alle figure 2.3-2.6 in questo caso le distribuzioni riportate si riferiscono solo ai geni differenzialmente espressi. Per quanto riguarda il primo grafico, prendendo come livello di riferimento *untreated*, si nota che tutti i trattamenti presentano un'espressione genica mediana leggermente maggiore ma una variabilità tale da sug-

gerire l'assenza di differenze di rilievo. Per quanto riguarda il secondo, l'espressione genica risulta più elevata quando essa è riferita al tumore schwannoma piuttosto che al meningioma. Per quanto riguarda il terzo, infine, non emergono differenze di rilievo.

4.2 Alcune specificazioni alternative

Data l'evidente struttura bimodale che caratterizza la distribuzione del logaritmo dell'espressione genica normalizzata (Figura 4.1), l'adattamento di un modello di regressione lineare normale

$$\log Y = \beta_0 + \beta_1 x_{l1} + \beta_2 x_{l2} + \beta_3 x_{l3} + \beta_4 x_{l4} + \beta_5 x_{l5} + \beta_6 x_{l6} + \epsilon_l$$

risulterebbe del tutto inadeguato.

A causa della presenza di un numero elevato di osservazioni pari a zero e l'originale natura di conteggio dell'espressione genica, una specificazione adeguata potrebbe essere un modello di Poisson con inflazione di zeri (*zero-inflated Poisson model*, ZIP, Lambert, 1992) dove la variabile risposta è costituita dai conteggi grezzi non normalizzati e i fattori di normalizzazione TMM vengono inseriti nel modello come *offset*. Per gestire la presenza della sovradispersione nei dati, si può inoltre considerare un più avanzato modello binomiale negativo con inflazione di zeri (*zero-inflated Negative Binomial model*, ZINB, Greene, 1994). In entrambi i casi, l'eccesso di zeri si suppone essere generato e descritto da un processo bernoulliano indipendente dal processo di conteggio.

Indicata con $1 - \pi_l$ la probabilità dell'evento $Y_l = 0$ secondo il primo processo, e con μ_l il valore atteso del secondo processo, descritto da un modello di Poisson o, rispettivamente, binomiale negativo, una formulazione generale è

$$logit(1-\pi_l) = \gamma_0+\gamma_1 x_{l1}+\gamma_2 x_{l2}+\gamma_3 x_{l3}+\gamma_4 x_{l4}+\gamma_5 x_{l5}+\gamma_6 x_{l6}+\log S_l^{(TMM)}$$

$$\log \mu_l = \beta_0 + \beta_1 x_{l1} + \beta_2 x_{l2} + \beta_3 x_{l3} + \beta_4 x_{l4} + \beta_5 x_{l5} + \beta_6 x_{l6} + \log S_l^{(TMM)}.$$

Un'ulteriore strada che ha senso percorrere è quella di mantenere una formulazione lineare normale dopo aver però operato una trasformazione dei dati in grado di renderli più gaussiani possibili, aggirando il problema probabilmente più impegnativo dello sviluppo di tecniche statistiche che potrebbero ospitare dati non normali.

Oltre al più noto metodo di Box e Cox (1964) esistono in letteratura trasformazioni più adeguate a gestire anche deviazioni dalla normalità

quali quelle riscontrati nei dati NF2. In proposito, Bartlett (1947) ha proposto una tecnica di normalizzazione mirata a rendere alcune distribuzioni, quali la bimodale, quanto più gaussiane possibili: la *Ordered Quantile normalization* (ORQ).

La tecnica di normalizzazione ORQ prende i dati originali riferiti all'espressione genica normalizzata con TMM contenuti nel vettore $\mathbf{Y}$ e applica la seguente trasformazione:

$$g(Y^{(k)}) = \Phi^{-1}\left(\frac{k - 1/2}{m}\right)$$

dove $Y^{(k)}$ è l'elemento di $\mathbf{Y}$ che ha rango k nel vettore ordinato e Φ è la funzione di ripartizione della Normale Standard.

Va tenuto però presente che tale tecnica di normalizzazione non garantisce dati trasformati distribuiti normalmente quando ci si trova in presenza di repliche (ties, Peterson e Cavanaugh, 2020), anche se potrebbe comunque produrre la migliore trasformazione normalizzante rispetto ad altre alternative.

Una volta effettuata la trasformazione, è possibile procedere con la stima di un modello lineare normale

$$g(Y_l) = \beta_0 + \beta_1 x_{l1} + \beta_2 x_{l2} + \beta_3 x_{l3} + \beta_4 x_{l4} + \beta_5 x_{l5} + \beta_6 x_{l6} + \epsilon_l \quad (4.2)$$

4.3 Applicazione dei modelli ai dati NF2

I modelli discussi nel paragrafo precedente sono stati applicati ai dati NF2 e sono di seguito illustrati.

Come atteso, il lineare modello normale applicato al logaritmo dell'espressione genica risulta del tutto inadeguato in quanto la sua analisi dei residui evidenzia chiari allontanamenti dalle ipotesi di normalità (Figura 4.3, a sinistra).

Anche l'analisi dei residui dei modelli con inflazione di zeri, tuttavia, evidenzia chiari allontanamenti dalle ipotesi dei modelli stessi, come è possibile notare dalla Figura 4.3 in centro e a destra. Inoltre, i coefficienti stimati (riportati in Tabella 4.1 con i relativi errori standard, statistiche test e p-value) non sono coerenti con l'analisi grafica del paragrafo precedente: il livello medio di espressione genica per tutti i trattamenti risulta essere superiore, e non inferiore, a quello di *untreated* e quello per il tumore schwannoma aumenta, e non diminuisce, rispetto al meningioma.

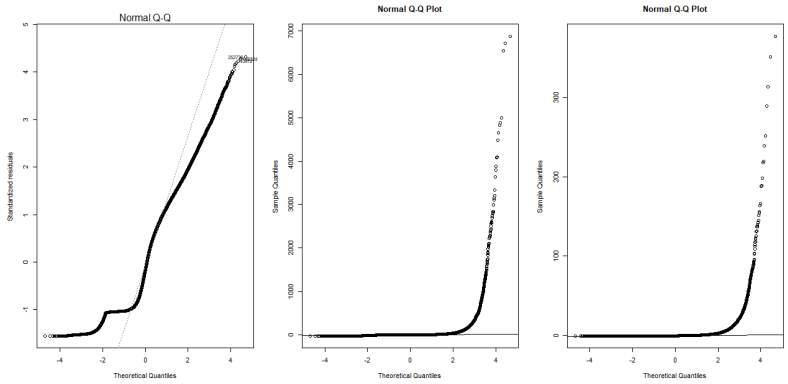


Figura 4.3: Analisi grafica dei residui del modello lineare normale (a sinistra), del modello ZIP (in centro) e del modello ZINB (a destra) applicato all'espressione genica in funzione del tipo di trattamento, dello stato di merlina e del tipo di tumore.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercetta	-2.22545	0.03365	-66.130	< 2e-16***
treatCUDC.907	-10.74044	2.80596	-3.828	0.000129***
treatDMSO	-0.36638	0.03638	-10.072	< 2e-16***
treatGSK2126458	-0.40561	0.03749	-10.821	< 2e-16***
treatPanobinostat	-10.78136	2.98477	-3.612	0.000304***
tumorSchwannoma	-6.93462	0.81295	-8.530	< 2e-16***
merlinAssente	-0.14257	0.03457	-4.124	3.73e-05***

Tabella 4.1: Adattamento modello ZINB all'espressione genica con l'introduzione di *offset* dati dai fattori di normalizzazione TMM.

Dunque, né il modello lineare normale né i modelli ZIP e ZINB adattati come innanzi descritti sono correttamente specificati per il set di dati a disposizione.

Il modello più appropriato risulta essere, invece, il modello lineare normale (4.2), ottenuto dopo aver trasformato i dati mediante il procedimento ORQ, volto a ridurne la bimodalità. Per i dati NF2, che presentano un numero elevato di *ties*, è necessario tuttavia applicare la normalizzazione ORQ con cautela, considerando il limite precedentemente descritto a cui si va incontro.

Come quanto è accaduto nell'analisi esplorativa, per semplicità di espressione da qui in poi con il termine "espressione genica" si intenderà il suo valore normalizzato con il metodo TMM e non il conteggio grezzo.

La distribuzione del logaritmo dell'espressione genica successivamente alla trasformazione ORQ è riportata in Figura 4.4. Come atteso,

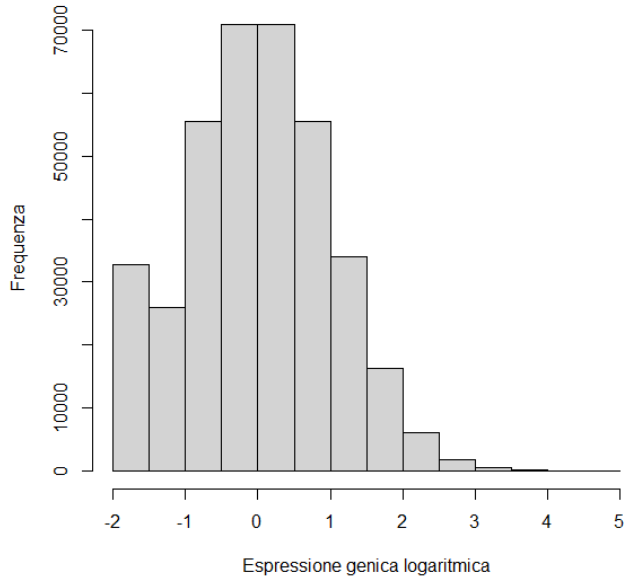


Figura 4.4: Istogramma della distribuzione del logaritmo dell'espressione genica trasformata con ORQ.

essendoci numerosi *ties* la distribuzione non è perfettamente gaussiana ma, in ogni caso, non presenta più la precedente marcata bimodalità.

La Figura 4.5 riporta l'analisi grafica dei residui del modello 4.2 applicato ai dati NF2. Sebbene l'adattamento non possa dirsi ottimale, il grafico quantile-quantile mostra un notevole miglioramento rispetto agli altri modelli. Inoltre, nel valutare l'elevata variabilità dei residui evidente anche dalle altre analisi diagnostiche, bisogna tenere conto della natura categoriale delle variabili esplicative. Nel complesso, e anche data la complessità del problema, il modello risulta soddisfacente.

Le stime dei coefficienti del modello di regressione lineare 4.2 sui dati relativi alla NF2 sono riportate in Tabella 4.2. Osservando i *p-value*, tutti i coefficienti risultano marginalmente statisticamente significativi ad eccezione di quello riferito alla condizione "Assente" dello stato di merlina. Ciò significa che lo stato di merlina è un fattore che non influenza l'espressione genica, al contrario di quanto accade per il tipo di trattamento e il genere di tumore.

Per quanto riguarda quest'ultimi due fattori, i coefficienti risultano coerenti con quanto visto nell'analisi grafica precedente. In particolare, tutti i trattamenti hanno un'espressione genica normalizzata media superiore a quella di *untreated* e il livello medio per il tumore schwannoma è superiore rispetto a quello del meningioma. Inoltre, in ordine crescen-

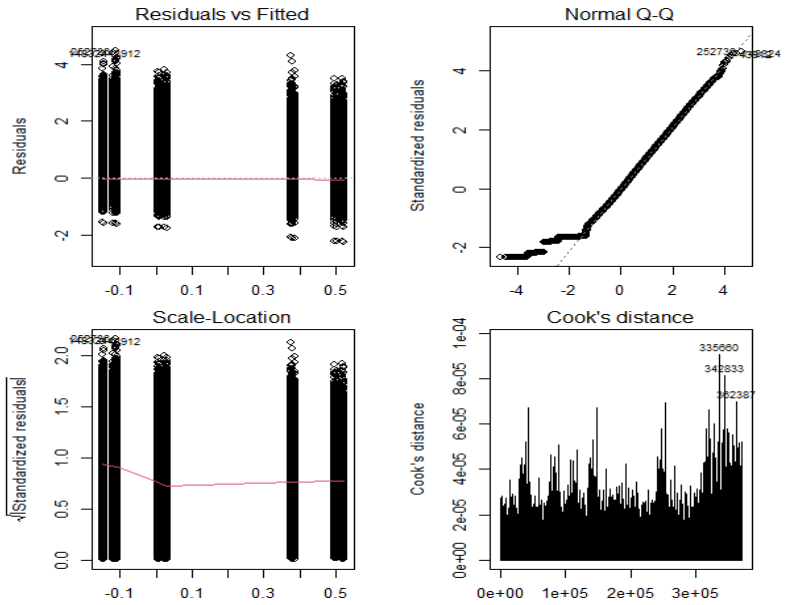


Figura 4.5: Analisi grafica dei residui del modello lineare normale applicato all'espressione genica trasformata con ORQ in funzione del tipo di trattamento, dello stato di merlina e del tipo di tumore.

te di espressione genica media normalizzata si ha: *untreated*, *DMSO*, *GSK2126458* e *Panobinostat*.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercetta	-0.150494	0.004521	-33.285	< 2e-16***
treatCUDC.907	0.174858	0.005032	34.750	< 2e-16***
treatDMSO	0.031165	0.005032	6.193	5.89e-10***
treatGSK2126458	0.037086	0.005032	7.370	1.71e-13***
treatPanobinostat	0.154907	0.005032	30.785	< 2e-16***
tumorSchwannoma	0.493563	0.004529	108.973	< 2e-16***
merlinAssente	0.005863	0.003583	1.636	0.102

Tabella 4.2: Adattamento modello lineare normale al logaritmo dell'espressione genica trasformata con ORQ.

Conclusioni

Con questa relazione si è voluto fornire una panoramica generale sulla neurofibromatosi di tipo 2 e sull'analisi dei dati di espressione genica derivanti dal sequenziamento dell'RNA.

Da un punto di vista statistico, le difficoltà principali che emergono nell'analisi di dati di espressione genica derivano in primo luogo dalla necessità di gestire un'ingente mole di dati, prodotta dal sequenziamento dell'RNA relativo a decine di migliaia di geni; tale difficoltà, oltre che avere un impatto sui tempi e sullo sforzo computazionale, deriva dal fatto che, a seconda degli obiettivi, il gene può fungere da variabile statistica, oltre che da unità, e questo comporta di dover gestire insiemi di dati in cui il numero di colonne supera abbondantemente il numero delle righe. I metodi di filtraggio e di determinazione dei geni differenzialamente espressi hanno dunque lo scopo di ridurre, almeno in parte, tale squilibrio. Una volta eliminati i geni poco importanti può emergere più chiaramente se, e come, diversi trattamenti influenzano l'espressione genica. Questo obiettivo è stato perseguito mediante l'applicazione di modelli di regressione, privilegiando tuttavia procedure in grado di gestire la sovradispersione e il permanere di un eccesso di geni non espressi.

Come già anticipato, la speranza iniziale dello studio *Synodos NF2* da cui sono stati tratti i dati analizzati, era quella di individuare un trattamento capace di agire allo stesso modo per entrambi i tipi di tumore e in modo differente a seconda dello stato di merlina. La significatività dei coefficienti del modello di regressione stimato, tuttavia, suggerisce che l'utilizzo dei trattamenti *CUDC.907*, *GSK2126458* e *Panobinostat* su un paziente affetto da meningioma e schwannoma rischierebbe di agire solamente su uno dei due tumori, con il rischio di causare possibili conseguenze all'altro. Inoltre, i modelli suggeriscono che nessuna terapia farmacologica, tra quelle considerate, è efficace a sconfiggere entrambi i tipi di tumore quando questi sono causati dalla NF2 (ossia quando lo stato di merlina risulta assente).

Dalle analisi affrontate nella relazione, dunque, sono emersi risultati interessanti che, tuttavia, meriterebbero degli approfondimenti futuri sia dal punto di vista statistico che da quello biomedico affinché si possa giungere, in futuro, all sviluppo di nuove terapie farmacologice efficaci contro i due tumori primari correlati alla NF2.

Bibliografia

Abbas-Aghababazadeh F.; Li Q.; Fridley B. L. (2018). Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PloS one*, **13**(10), e0206312.

Allaway R.; Angus S. P.; Beauchamp R. L.; Blakeley J. O.; Bott M.; Burns S. S.; Carlstedt A.; Chang L.-S.; Chen X.; Clapp D. W.; Desouza P. A.; Erdin S.; Fernandez-Valle C.; Guinney J.; Gusella J. F.; Haggarty S. J.; Johnson G. L.; La Rosa S.; Morrison H.; Petrilli A. M.; Plotkin S. R.; Pratap A.; Ramesh V.; Sciaky N.; Stemmer-Rachamimov A.; Stuhlmiller T. J.; Talkowski M. E.; Welling D. B.; Yates C. W.; Zawistowski J. S.; Zhao W.-N. (2018). Traditional and systems biology based drug discovery for the rare tumor syndrome neurofibromatosis type 2. *PLOS ONE*, **13**(6), 1–26.

Anders S.; Huber W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, pp. 1–1.

Asthagiri A. R.; Parry D. M.; Butman J. A.; Kim H. J.; Tsilou E. T.; Zhuang Z.; Lonser R. R. (2009). Neurofibromatosis type 2. *The Lancet*, **373**(9679), 1974–1986.

Bartlett M. S. (1947). The use of transformations. *Biometrics*, **3**(1), 39–52.

Benjamini Y.; Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, **57**(1), 289–300.

Box G. E.; Cox D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, **26**(2), 211–243.

Bullard J. H.; Purdom E.; Hansen K. D.; Dudoit S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, **11**(1), 1–13.

Chen Y.; Lun A. T.; Smyth G. K. (2014). Differential expression analysis of complex rna-seq experiments using edger. *Statistical analysis of next generation sequencing data*, pp. 51–74.

Chen Y.; Lun A. T.; Smyth G. K. (2016). From reads to genes to pathways: differential expression analysis of rna-seq experiments using rsubread and the edger quasi-likelihood pipeline. *F1000Research*, **5**.

Chen Y.; McCarthy D.; Ritchie M.; Robinson M.; Smyth G.; Hall E. (2020). edger: differential analysis of sequence read count data user's guide. *Accessed: Jul*, **8**.

Cox D. R.; Reid N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, **49**(1), 1–18.

Ferner R. E. (2007). Neurofibromatosis 1 and neurofibromatosis 2: a twenty first century perspective. *The Lancet Neurology*, **6**(4), 340–351.

Gandolfo L. C.; Speed T. P. (2018). Rle plots: Visualizing unwanted variation in high dimensional data. *PloS one*, **13**(2), e0191629.

Gerber P.; Antal A.; Neumann N.; Homey B.; Matuschek C.; Peiper M.; Budach W.; Bölke E. (2009). Neurofibromatosis. *European journal of medical research*, **14**(3), 102–105.

Greene W. H. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models.

Lambert D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1), 1–14.

Marioni J. C.; Mason C. E.; Mane S. M.; Stephens M.; Gilad Y. (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, **18**(9), 1509–1517.

McCullagh P.; Nelder J. A. (2019). *Generalized linear models*. Routledge.

Peterson R. A.; Cavanaugh J. E. (2020). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of applied statistics*, **47**(13-15), 2312–2327.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Riccardi V. M. (1981). von recklinghausen neurofibromatosis. *New England Journal of Medicine*, **305**(27), 1617–1627.

Robinson M. D.; Oshlack A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, **11**(3), 1–9.

Robinson M. D.; Smyth G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**(21), 2881–2887.

Robinson M. D.; Smyth G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, **9**(2), 321–332.

Wang X. (2006). Approximating bayesian inference by weighted likelihood. *The Canadian Journal of Statistics/La revue canadienne de statistique*, pp. 279–298.