

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Dipartimento di Ingegneria dell'Informazione

LAUREA MAGISTRALE IN BIOINGEGNERIA

A.A. 2011/2012

CONFRONTO DI METODI STATISTICI PER LA MISURA DELL'ESPRESSIONE DIFFERENZIALE IN DATI DI RNA SEQUENCING

Relatore: Prof. Barbara Di Camillo

Correlatore: Dr. Francesca Finotello

Laureando: Andrea Apolloni

Abstract

L'RNA sequencing (RNA-Seq) è una tecnica per l'analisi del trascrittoma e la sua quantificazione, basata sulle recenti tecnologie Next-Generation Sequencing (NGS). Lo sviluppo tecnologico ha infatti permesso di ottenere piattaforme di sequenziamento che generano dati ad alto throughput e con costi molto inferiori rispetto ai sequenziatori tradizionali.
La prima parte di questa tesi offre una panoramica sulle più diffuse piattaforme commerciali di sequenziamento NGS (454 Genome Sequencer di Roche, Genome Analyzer di Illumina, SOLiD di Applied Biosystems), valutandone le caratteristiche tecniche e le prestazioni.

I dati grezzi che i sequenziatori permettono di ottenere sono le *read*, cioè sequenze che identificano l'ordine in cui si susseguono le basi azotate che compongono il gene. In un esperimento RNA-Seq, l'espressione genica viene misurata in termini di count, cioè del numero di read mappate sui geni di un genoma o trascrittoma di riferimento. I count sono dunque somme di variabili aleatorie (l'assegnazione delle read a ciascun gene) e sono descrivibili tramite modelli statistici. I principali modelli utilizzati in letteratura sono il modello binomiale, il modello di Poisson e il modello Binomiale Negativo. La descrizione statistica dei dati di RNA-Seq è oggetto di studio molto recente e non esiste ancora una descrizione comune. In questa tesi si è quindi cercato di organizzare le informazioni in un modello generale della distribuzione dei dati, stabilendo una notazione comune nella descrizione dei lavori dei diversi autori.

Una delle più interessanti applicazioni di RNA-Seq è l'analisi dell'espressione differenziale, cioè l'identificazione dei geni che presentano significative differenze del loro livello di espressione fra due o più condizioni sperimentali (interne o esterne alla cellula). In esperimenti RNA-Seq, ciò significa valutare se le differenze osservate nei count delle diverse condizioni sperimentali siano o meno significative in senso statistico.
Sono molti gli autori e i gruppi di ricerca che hanno sviluppato proposte di metodi di analisi differenziale, che implementano i modelli di distribuzione dei dati sopra citati. In questa tesi sono stati considerati i metodi DEGSeq, PoissonSeq (che implementano il modello di Poisson), edgeR e DESeq (che implementano il modello Binomiale Negativo). Ciascun metodo è stato testato su due data set pubblici valutandone le prestazioni in termini di precisione e sensitività. EdgeR è risultato il migliore, anche se tutti i diversi metodi hanno ottenuto risultati molto simili fra loro. Studi futuri con presenza di repliche biologiche potranno fornire indicazioni statisticamente più significative sulla bontà dei metodi.

Indice

Abstract ... **3**
Glossario ... **7**
Notazione ... **10**
1 Introduzione ... **11**
2 Il sequenziamento del DNA ... **13**
2.1 Il sequenziamento Sanger ... 13
2.2 I sequenziatori NGS ... 15
2.3 Il sequenziatore 454 Genome Sequencer di Roche ... 16
2.4 Il sequenziatore Genome Analyzer di Illumina ... 17
2.5 Il sequenziatore SOLiD di Applied Biosystems ... 18
2.6 I sequenziatori di terza generazione ... 20
3 RNA Sequencing ... **23**
3.1 Modello di Poisson ... 23
3.2 Modello Binomiale Negativo ... 24
4 Metodi per l'analisi dell'espressione genica differenziale ... **27**
4.1 DEGSeq ... 27
4.1.1 Metodo MA-plot based con modello a campionamento casuale (MARS) ... 27
4.1.2 Metodo MA-plot based con repliche tecniche (MATR) ... 29
4.1.3 Correzione per test multipli ... 29
4.1.4 Metodo di verifica fra repliche tecniche (CTR) ... 29
4.2 edgeR ... 30
4.2.1 Modello Binomiale Negativo dei count ... 30
4.2.2 Normalizzazione ... 30
4.2.3 Analisi differenziale con edgeR ... 32
4.3 DESeq ... 33
4.3.1 Modello ... 33
4.3.2 Fit dei dati ... 34
4.3.3 Test di analisi differenziale ... 35
4.3.4 Calcolo di P(a,b) ... 35
4.4 PoissonSeq ... 36
4.4.1 Modello ... 36
4.4.2 Fit del modello ... 36
4.4.3 Normalizzazione ... 36
4.4.4 Test statistico per l'analisi differenziale ... 37
4.4.5 Stima della False Discovery Rate ... 37
5 Dati ... **41**

5.1 Data set di Marioni *et al.* 41
5.2 Data set di Cumbie *et al.* 42
5.3 Data set MAQC 42
6 Analisi data set di Marioni *et al.* 45
6.1 DEGSeq 45
6.1.1 Metodo MARS 45
6.1.2 Metodo MATR 46
6.1.3 Metodo CTR 50
6.2 edgeR 52
6.2.1 Analisi differenziale con stima della dispersione comune 53
6.2.2 Analisi con stima della dispersione genewise 55
6.3 DESeq 58
6.4 PoissonSeq 60
6.5 Confronto con microarray 62
7 Analisi data set MAQC 65
7.1 DEGSeq MARS 65
7.2 edgeR 65
7.2.1 Analisi differenziale con stima della dispersione comune 66
7.2.2 Analisi con stima della dispersione genewise 67
7.3 DESeq 69
7.4 PoissonSeq 70
7.5 Confronto con qRT-PCR 70
8 Analisi data set di Cumbie *et al.* 73
8.1 DEGSeq MARS 73
8.2 edgeR 73
8.2.1 Analisi differenziale con stima della dispersione comune 74
8.2.2 Analisi con stima della dispersione genewise 75
8.3 DESeq 76
8.4 PoissonSeq 77
9 Riassunto analisi data set 79
10 Conclusioni 81
11 Bibliografia 83

Glossario

Adattatore Un oligonucleotide sintetico a doppio filamento usato per permettere di legare fra loro due molecole.

Annealing Attacco di un oligonucleotide a una molecola di DNA a singoli filamento mediante ibridazione

bp Unità di misura della lunghezza di un frammento di acido nucleico a doppio filamento. Indica il numero di basi appaiate (base pair) che si susseguono.

cDNA Il DNA complementare è un DNA a doppia elica sintetizzato a partire da una molecola di mRNA, ad opera dell'enzima trascrittasi inversa. I sequenziatori attualmente disponibili richiedono in input DNA, quindi nelle applicazioni RNA-Seq si retrotrascrive l'RNA in cDNA.

Complementare Si dice di due nucleotidi che possono formare coppie di basi per produrre una molecola a doppio filamento. Sono complementari guanina con citosina e adenina con timina (nel DNA) o uracile (nell'RNA).

Count Numero di read mappate su una feature biologica di interesse (gene, trascritto, esone). Si può pensare come proporzionale all'abbondanza della feature, quindi del suo livello di espressione. Sono i dati di esperimenti RNA-Seq.

Denaturazione termica Rottura dei legami idrogeno coinvoli nell'appaiamento delle basi, tramite somministrazione di calore.

DNA L'acido nucleico che contiene le informazioni genetiche necessarie alla sintesi di RNA e di proteine. É un polimero i cui monomeri sono i nucleotici (A, G, C e T), l'ordine cui cui i monomeri sono disposti in sequenza costituisce l'informazione genica. É strutturato in forma di coppia di filamenti complementari fra loro, detta doppia elica. Negli eucarioti, il DNA si trova all'interno del nucleo cellulare, in strutture dette cromosomi.

Flow-cell *(piastra di sequenziamento)* Il substrato solido su cui sono posti i campioni di DNA da sequenziare. I campioni vengono preparati e posti sulla piastra di sequenziamento, che viene poi inserita all'interno del sequenziatore, dove avviene il processo di sequenziamento vero e proprio. Può essere divisa in lane,

indipendenti fra loro, permettendo il sequenziamento di campioni indipendenti in un unica run. Alcuni sequenziatori sono in grado di sequenziare in parallelo due flow-cell per run.

Lane — Regioni di una flow-cell indipendenti fra loro. Possono essere utilizzare per sequenziare in parallelo più campioni diversi o uno stesso campione ma sequenziato in più volte indipendenti. Generalmente, una flow-cell contiene otto lane.

Library size — Numero di read mappate in un campione o libreria.

Ligasi — Enzima che ripara nella cellula discontinuità a singolo filamento in molecole di DNA a doppio filamento. Viene usata nel clonaggio dei geni per unire molecole di DNA.

nt — Unità di misura della lunghezza di un frammento di acido nucleico a singolo filamento. Indica il numero di nucleotidi che si susseguono.

Nucleotidi (dNTP) — Sono i monomeri degli acidi nucleici (DNA e RNA), composti da una base azotata, uno zuccchero a cinque atomi di carbonio e un gruppo fosfato. I nomi dei nucleotidi sono abbreviati in codici standard di tre lettere, precedute da una "d" nel caso dei nucleotidi del DNA. I nucleotidi sono adenina (dATP), guanina (dGTP), citosina (dCTP) e timina (dTTP). Nell'RNA la timina è sostituita dall'uracile (UTP).

PCR — (*Polymerase Chain Reaction*, reazione a catena della polimerasi) É una tecnica di amplificazione che sfrutta l'enzima DNA polimerasi per sistetizzare copie identiche di un singolo filamento di DNA

Plasmide — Un tratto di DNA solitamente circolare, nella maggior parte dei casi indipendente dal cromosoma dell'ospite, presente nei batteri e in alcuni altri tipi di cellule.

Primer — Breve oligonucleotide a singolo filamento che, quando legato (annealing) ad una molecola stampo a singolo filamento, agisce da punto di partenza per la sintesi del filamento complementare diretta da DNA polimerasi.

Read — Sequenza fornita in output dal sequenziatore, che identifica l'ordine in cui si susseguono le basi nei frammenti di cDNA. Il numero di basi che la compongono rappresenta la sua lunghezza (misurata in nt o bp).

RPKM — (*Reads Per Kilobase per Million mapped reads*) Indice del livello di espressione

definito da Mortazavi *et al.* [1]. $RPKM = \frac{totaleread}{readmappate[milioni] * lunghezzagene[Kb]}$

RT-PCR (trascrizione inversa-PCR)
: Una tecnica di PCR in cui il materiale di partenza è RNA. Il primo passaggio del procedimento è la conversione dell'RNA in DNA con la trascrittasi inversa.

Run del sequenziatore
: Un ciclo di funzionamento del sequenziatore.

Sequencing depth
: Numero totale di read sequenziate in un campione. A volte calcolata come il numero di read mappate in un campione o libreria e coincide con la *library size*.

Throughput
: La quantità di dati che la strumentazione può fornire in un periodo di tempo. Nelle tecnologie di sequenziamento è espresso in termini di numero di basi (o milioni di basi, Mb) sequenziate per run dello strumento, che dipende quindi del numero e dalla lunghezza delle read date in output dal sequenziatore.

Notazione

f	indice del trascritto, $f = 1..F$
g	indice del gene, $g = 1..G$
j	indice del campione, $j = 1..J$
$\rho(j)$	indice del gruppo sperimentale a cui appartiene il campione j, $\rho = 1..N$
X_{gj}	count del gene g nel campione j
s_j	library size del campione j
$p_{g\rho(j)}$	probabilità che una read della condizione $\rho(j)$ sia associata al gene g
ϕ_g	dispersione del gene g
T_j	RNA totale di un campione

1 Introduzione

Il trascrittoma è l'insieme delle molecole di RNA (o trascritti) presenti in una cellula e la loro quantità. L'espressione dei trascritti varia in dipendenza delle condizioni ambientali e della cellula, quindi la quantificazione del trascrittoma è utile a comprendere quali geni siano attivati nelle diverse fasi del ciclo cellulare o nello sviluppo di condizioni patologiche. Un'interessante applicazione della trascrittomica quantitativa è l'analisi differenziale dell'espressione genica, ottenuta confrontando i profili trascrizionali di due o più individui, tessuti o tipi cellulari. Ad esempio, informazioni sui trascritti in soggetti sani e soggetti malati possono permettere di rilevare quali geni sono espressi in maniera significativamente diversa fra i due gruppi e quindi di evidenziare differenze che la condizione patologica comporta. Questa stessa metodologia di analisi può essere applicata anche a studi su un unico soggetto valutato nel tempo (ad esempio per valutare l'effetto di un trattamento medico).

L'analisi del trascrittoma può avvenire secondo due tecnologie: quelle basate su ibridazione e quelle, più recenti, basate su sequenziamento. L'ibridazione si basa sulla proprietà dei nucelotidi di appaiarsi con i loro complementari fissate su un supporto. Di questa categoria fanno parte i microarray, da anni largamente utilizzati per ottenere informazioni sull'espressione genica. Sono costituiti da un supporto solido a cui sono ancorate delle sonde di DNA, dette *probe*, in numero molto elevato per ogni gene e disposte in posizioni note. L'RNA estratto dalla cellula viene retrotrascritto, marcato con fluorescente e ibridato con il microarray. L'intensità della fluorescenza è una misura di quante molecole hanno ibridato la probe, ovvero di quanto il gene associato alla probe è espresso nella cellula. Questa tecnica presenta numerosi limiti: la necessità di conoscere a priori le sequenze geniche per la progettazione delle probe e il limitato range dinamico (cioè il rapporto fra i livelli di massima e minima espressione genica misurabili), dovuto al rumore di fondo e al fenomeno di saturazione del segnale.

Per sequenziamento si intende invece l'identificazione della sequenza di un acido nucleico fornito in input alla strumentazione. Una tecnica recente per la misura del trascrittoma basata sul sequenziamento è l'RNA sequencing (RNA-seq) [2], basata sulle tecnologie di sequenziamento NGS (*Next Generation Sequencing*). Il protocollo di un esperimento di RNA-seq varia in base alla tecnologia utilizzata ma è comunque possibile descriverne in linea generale i passaggi principali. I campioni di RNA sono estratti dalle cellule in analisi e retrotrascritti in DNA. Il DNA ottenuto viene sequenziato ad alto throughput per ottenere frammenti a sequenza nota, detti *read*, di lunghezza variabile in base alla tecnologia del sequenziatore utilizzato (capitolo 2, "Il sequenziamento del DNA"). Le read sono quindi le sequenze, ottenute dal sequenziatore, che identificano l'ordine in cui si susseguono le basi nei frammenti di DNA. Le read vengono mappate su un genoma o un trascrittoma di riferimento per identificare quali geni siano espressi nel campione. Infatti, più numerose sono le molecole di un trascritto, più probabilità avrà quel trascritto di essere sequenziato e di generare read. Il totale delle read allineate su un gene (o su un trascritto o un esone), detto *count*, può essere quindi visto come una misura del livello di espressione del gene stesso.

Rispetto ai microarray, RNA-Seq presenta una serie di vantaggi [3]: è applicabile anche ad organismi di cui non sia disponibile il genoma/trascrittoma di riferimento, permette elevata risoluzione delle misure e un più ampio range dinamico, oltre a offrire buona riproducibilità e bassa differenza sistematica tra repliche tecniche [4] e costi relativamente contenuti.

I dati di un esperimento di RNA-Seq sono i count. Mentre i microarray misurano l'espressione genica come intensità luminosa, cioè un valore continuo ("analogico"), i count sono una misura discreta, che può essere intesa come "digitale". Le descrizioni statistiche dei dati da microarray non trovano quindi buona applicabilità ai dati RNA-Seq e sono dunque necessari nuovi modelli. Fra quelli attualmente allo studio il maggior interesse è rivolto al modello di Poisson (capitolo 3.1) e al modello Binomiale Negativo (capitolo 3.2). Il primo è un modello

che descrive la distribuzione dei dati tramite un unico paramentro, che ne rappresenta sia la media sia la varianza. Si tratta quindi di un modello semplice, che richiede la stima di un solo parametro. Tuttavia è proprio questa sua semplicità a rappresentarne il principale limite: generalmente i dati di RNA-Seq presentano una dispersione che non è rappresentabile con la relazione-media varianza prevista dal modello di Poisson. Il modello Binomiale Negativo estende la descrizione anche a dati dispersi, tramite la stima di un secondo parametro che rappresenta proprio il coefficiente di dispersione.

I due modelli statistici sono stati implementati in molti nuovi metodi per l'analisi dell'espressione genica differenziale, che permettono di determinare quali geni presentino differenze nel livello di espressione (rappresentato dai count) statisticamente significative fra due o più condizioni sperimentali. In questa tesi sono stati considerati i metodi DEGSeq (capitolo 4.1), edgeR (capitolo 4.2), DESeq (capitolo 4.3) e PoissonSeq (capitolo 4.4).

2 Il sequenziamento del DNA

Il sequenziamento del DNA è il processo che permette di determinare l'esatta sequenza delle basi nucleotidiche, cioè di conoscere l'ordine in cui le basi si susseguono nelle molecole di DNA contenute nel campione in esame. I moderni sviluppi della tecnologia hanno permesso di ottenere strumentazioni (i sequenziatori) in grado di sequenziare in automatico anche genomi complessi, riducendo il carico di lavoro per l'operatore. Generalmente, i sequenziatori non sono in grado di sequenziale il DNA in input per intero, ma "leggono" delle porzioni di sequenza, la cui lunghezza dipende dalla particolare tecnologia utilizzata. Queste stringhe, chiamate appunto "read", costituiscono l'output del sequenziamento e rappresentano le stringhe campionate dal DNA in input. Indipendentemente dalla tecnologia utilizzata, è possibile riconoscere tre fasi comuni a tutte le metodiche di sequenziamento:

1. preparazione del campione di DNA
2. amplificazione dei frammenti
3. processo di sequenziamento vero e proprio

Il metodo Sanger (detto anche "sequenziamento a terminazione della catena") [5], che prende il nome dal suo ideatore, Frederick Sanger, ha rappresentato una svolta epocale nella biologia molecolare ed è tutt'oggi utilizzato in alcuni ambiti, come i test di paternità e la biologia forense. Tuttavia, si tratta di un metodo costoso, che richiede molto tempo per sequenziare genomi complessi, al quale si stanno sostituendo le cosiddette tecnologie NGS (Next-Generation Sequencing) o ultra-high-throughput. Le diverse tecniche NGS offrono simili miglioramenti rispetto al metodo Sanger: maggiore mole di dati in output (si parla di centinaia di milioni di basi per singola run dello strumento), costi e tempi ridotti. Le principali limitazioni legate a queste nuove tecnologie rispetto al metodo Sanger sono invece: riduzione della lunghezza delle read (50-400 nt rispetto 1000 nt delle read Sanger) e una minore accuratezza nella lettura delle basi (il numero e le caratteristiche degli errori di sequenziamento variano in relazione alla specifica piattaforma). Recenti miglioramenti tecnologici hanno permesso di sviluppare nuove piattaforme per il sequenziamento che permettono di ottenere read lunghe più di 1000 nt, a costi ancora più ridotti. Si parla in questo caso di sequenziatori di terza generazione (TGS, Third Generation Sequencing), definendo invece sequenziatori di seconda (SGS, Second Generation Sequencing) quelli descritti sopra e caratterizzati da read medio-corte [6]. Tra questi, i più famosi sono il 454 Genome Analizer di Roche [7], il SOLiD di Applied Biosystems [8] e il Genome Analizer di Illumina [9]; quest'ultimo è la piattaforma ad oggi più utilizzata per studi di RNA sequencing.

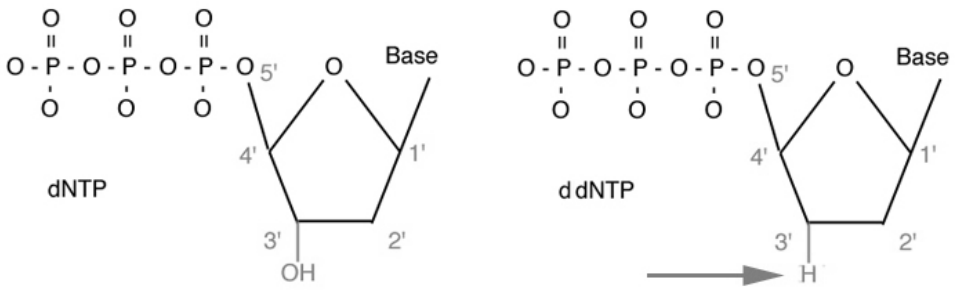
2.1 Il sequenziamento Sanger

Il sequenziamento Sanger è stato sviluppato nel 1977 da Frederick Sanger [5] ed è stato ampiamente utilizzato in passato per caratterizzare il genoma di molti organismi.

Il protocollo prevede inizialmente l'amplificazione delle molecole di DNA da sequenziare. L'amplificazione è l'operazione con cui si ottengono molte copie identiche delle molecole originali e può essere eseguita in due modi: con l'uso di plasmidi (in caso di sequenziamento *de novo*) o tramite la PCR (in caso di ri-sequenziamento). I plasmidi sono piccole porzioni circolari di DNA presenti nei batteri, sono molecole che possiedono almeno una sequenza di DNA che può agire da origine della replicazione, riuscendo a replicarsi all'interno della cellula in modo indipendente dal cromosoma batterico principale. I batteri utilizzano i plasmici come vettore per il trasferimento di DNA e, una volta integrati in un organismo ospite, sono duplicati a ogni sua divisione. Il plasmide viene tagliato ad opera di enzimi di restrizione e la sequenza da amplificare vi viene inserita e legata tramite degli adattatori. Una volta modificato, il plasmide è inserito in un batterio ospite (tipicamente *E. coli*) che si replicherà dando origine a una colonia batterica. Ogni batterio della colonia ottenuta contiene così una

copia della molecola di DNA da amplificare, che può essere isolata dal rimanente DNA. La PCR (o reazione a catena della polmerasi) è invece una tecnica con cui una breve regione di una molecola di DNA è copiata molte volte da una DNA polimerasi, un enzima presente nelle cellule. Ad una estremità della sequenza da amplificare viene ibridata una breve sequenza di oligonucleotidi, detta primer, che funge da innesco per DNA polimerasi, definendo il punto di inizio della regione da amplificare. La fase di adesione del primer alla molecola è detta primer annealing. Per iniziare un'amplificazione mediante PCR, si aggiunge DNA polimerasi allo stampo di DNA legato ai primer e si incuba la miscela in modo che l'enzima sintetizzi nuovi filamenti complementari. La miscela viene quindi scaldata per separare i nuovi filamenti sintetizzati (denaturazione termica) e raffreddata per permettere l'ibridazione di ulteriori primer nelle loro rispettive posizioni, fra cui posizioni sui filamenti appena sintetizzati. La DNA polimerasi esegue quindi un secondo ciclo di sintesi di DNA. L'intero ciclo di denaturazione-ibridazione-sintesi viene ripetuto, di solito 25-30 volte, portando alla sintesi finale di parecchi milioni di copie del frammento di DNA originale.

La successiva fase di sequenziamento inizia con la denaturazione delle molecole in input, che porta alla separazione dei due singoli filamenti di DNA. Il successivo passaggio è il primer annealing, che crea un punto di partenza per la sintesi del filamento complementare, catalizzata da una DNA polimerasi. La reazione di sintesi del filamento è iniziata aggiungendo alla soluzione contenente i filamenti di DNA l'enzima DNA polimerasi e i quattro nucleotidi dNTP (dATP, dTTP, dCTP, dGTP). Alla miscela vengono inoltre aggiunti dei dideossinucleotidi (ddNTP) in quantità inferiore rispetto ai dNTP e marcati con un fluorescente. I ddNTP possono essere incorporati nel filamento sintetizzato dalla DNA polimerasi esattamente come un nucleotide normale, ma comportano la terminazione dell'elongazione del filamento generato. Ciò accade perché il ddNTP è privo del gruppo ossidrilico nella posizione 3' dello zucchero che lo compone (Figura 1) e non fornisce quindi il sito di legame per il dNTP successivo: ogni volta che un ddNTP viene incorporato da DNA polimerasi si ha terminazione della catena.



a b

Figura 1. Struttura di un dNTP (a) e di un ddNTP (b). La freccia indica la posizione in cui il gruppo –OH viene sostituito da –H.

Il rapporto tra dNTP e ddNTP è tale da consentire la polimerizzazione di un tratto notevole del filamento prima che venga aggiunta una molecola di ddNTP. Si ottengono così quattro famiglie distinte di nuovi filamenti: una che contiene filamenti che terminano tutti in dATP, una di filamenti che terminano in dTTP e così via. Ogni famiglia comprende filamenti di lunghezza diversa, secondo la posizione in cui è avvenuto l'evento di terminazione.

I ddNTP sono marcati per fluorescenza, cioè legati covalentemente a un fluorocromo il cui colore identifica univocamente lo specifico ddNTP. Le quattro famiglie di filamenti saranno perciò distinguibili fra loro, anche se unite in un'unica miscela. Il tipo di ddNTP che costituisce l'ultima base del filamento è quindi noto, rimane incognita la lunghezza del filamento, cioè la posizione del ddNTP. Questa informazione si ottiene applicando l'elettroforesi capillare, un'evoluzione della tecnica originaria che utilizzava invece l'elettroforesi su gel. Si fanno passare i filamenti di DNA all'interno di un capillare, ai capi del

quale è applicata una differenza di potenziale; il DNA è carico negativamente e i frammenti migrano quindi verso il polo positivo del capillare. Questo è riempito di un gel che ostacola la migrazione dei frammenti: quelli di lunghezza inferiore saranno meno ostacolati nel percorso e percorreranno il capillare in un tempo minore permettendo di separare i filamenti in base alla loro lunghezza. All'uscita del capillare, un laser eccita i marcatori fluorescenti provocando emissioni elettromagnetiche, captate da un rilevatore che ne riconosce la lunghezza d'onda ottenendo la "traccia" Sanger, poi elaborata dal software in picchi di colore diverso (elettroferogramma). Note le lunghezze dei frammenti e il ddNTP con cui terminano è immediato leggere la sequenza: se un frammento è lungo 20 nt e il marcatore è del colore che identifica una G, in posizione 20 della read ci sarà una G.

La tecnologia Sanger permette di ottenere sequenze di lunghezza tra i 600 e i 1000 nucleotid sequenziando comunemente fino a 96 sequenze, con un throughput giornaliero di 6 Mb. Gli errori di sequenziamento sono imputabili principalmente all'amplificazione, a contaminazioni del campione e a errato riconoscimento dei ddNTP e della lunghezza delle sequenze, a causa di errori nella separazione elettroforetica. Mediamente, il tasso di errore medio (media valutata su tutte le basi della sequenza) è comunque piuttosto basso, compreso tra 10^{-5} e 10^{-6}.

2.2 I sequenziatori NGS

La metodologia RNA-Seq utilizza le piattaforme di sequenziamento NGS per quantificare l'abbondanza delle molecole di RNA (trascritti) e quindi per ottenere una misura del livello di espressione dei geni corrispondenti, in un campione biologico.

Idealmente, la pipeline di un esperimento RNA-Seq dovrebbe prevedere la sola estrazione dell'RNA dalla cellula di interesse e il suo diretto sequenziamento, senza manipolazioni che possano introdurre artefatti. In realtà, gli esperimenti di RNA-sequencing basati sui sequenziatori di seconda generazione prevedono un protocollo che include diverse fasi di manipolazione. Per prima cosa, l'RNA estratto dalla cellula viene frammentato in sequenze di dimensione inferiore, compresa tra i 200 e i 500 bp, per ottenere frammenti di dimensione compatibile con i sequenziatori in uso. Il processo di frammentazione è realizzato tramite idrolisi o nebulizzazione.

Ottenuti i frammenti di RNA, questi vengono retrotrascritti in DNA. La retrotrascrizione (o trascrizione inversa) è il processo di sintesi di un filamento di DNA complementare a partire da un RNA. Nel processo di retrotrascrizione, dopo l'annealing di un primer al filamento di RNA, l'enzima trascrittasi inversa sintetizza un filamento di DNA complementare all'RNA stampo. Terminata la sintesi, si ottiene una molecola ibrida (un filamento di DNA e uno di RNA) di cui viene degradato parzialmente il filamento di RNA. I frammenti rimanenti di RNA servono allora da primer per sintetizzare il secondo filamento, producendo una molecola di DNA a doppio filamento, detta DNA complementare (cDNA). La retrotrascrizione dell'RNA in cDNA è resa necessaria dal fatto che i sequenziatori oggi disponibili sono in grado di sequenziare DNA e non direttamente RNA. Nei protocolli di RNA-seq oggi in uso, la retrotrascrizione viene effettuata dopo la frammentazione e non viceversa poiché in questo modo il bias introdotto è minore e il coverage del trascritto risulta più uniforme [10].

Le molecole di cDNA vengono quindi amplificate, aumentando il numero di copie di ciascun frammento, e fornite in input ai sequenziatori. Il sequenziatore ad oggi più utilizzato per applicazioni RNA-seq è il Genome Analizer di Illumina (capitolo 2.4). Il sequenziatore "legge" i frammenti in input, identificando l'ordine in cui le basi nucleotidiche si susseguono. Le sequenze lette dai sequenziatori sono dette "read". Le read NGS sono molto più corte di quelle prodotte tramite tecnologia Sanger e presenti in numero molto maggiore (si parla di milioni di read per singola run dello strumento).

2.3 Il sequenziatore 454 Genome Sequencer di Roche

Il sequenziatore Genome Sequencer, commercializzato nel 2005 da 454 Life Sciences (ora proprietà di Roche Diagnostics), ha rappresentato la prima piattaforma di sequenziamento NGS disponibile sul mercato.

Ai frammenti, denaturati per renderli a singolo filamento, sono attaccati due adattatori alle estremità. La fase di amplificazione utilizza delle sfere dal diametro di 28 μm la cui superficie è ricoperta da oligonucleotidi complementari ad uno degli adattatori dei frammenti. Frammenti e sfere sono posti in contatto permettendo la formazione di un legame fra loro. Le sfere legate al filamento sono quindi messe, assieme ad enzimi amplificatori, nella fase acquosa di un'emulsione di acqua in olio per far avvenire l'amplificazione clonale tramite PCR (PCR a emulsione, em-PCR). Si tratta di un processo ciclico in cui DNA polimerasi sintetizza il filamento complementare, si denatura la doppia elica e si sintetizzano due nuovi filamenti complementari, fino ad ottenere decine di milioni di copie del frammento per ogni sfera. In Figura 2 è rappresentato graficamente il processo.

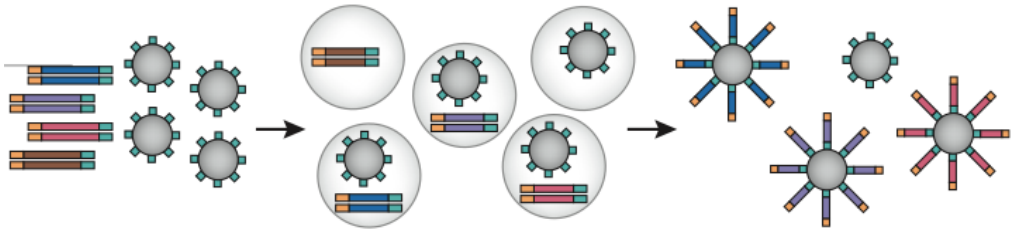


Figura 2. Amplificazione clonale dei frammenti da sequenziare. Ai frammenti di DNA sono aggiunti i due adattatori (rappresentati in colore giallo e verde). Le sfere possiedono sulla loro superficie degli oligonucleotidi complementari a uno degli adattatori. Sfere e frammenti sono posti in gocce d'acqua di un'emulsione di acqua in olio per l'amplificazione (em-PCR). Si ottengono sfere sulla cui superficie sono legate molte copie dei frammenti di partenza. Immagine presa da [11]

Terminata l'amplificazione, le sfere ricoperte dai nuovi filamenti sono poste sulla piastra di sequenziamento. Nel sequenziatore 454/Roche, la piastra (detta PicoTiterPlate) ha circa 1,5 milioni di pozzetti, il cui diametro (44 μm) permette di accogliere una sola sfera. Le sfere vi si depositano per gravità e sono ricoperte da numerose sfere più piccole legate a sulfurilasi, luciferasi, apirasi e luciferina, enzimi e substrati necessari per il processo di sequenziamento. Genome Sequencer utilizza infatti il metodo di pirosequenziamento [12], che prevede il completamento iterativo del filamento e la simultanea lettura del segnale emesso dai nucleotidi incorporati. É un processo ciclico, in cui un singolo dNTP viene aggiunto di volta in volta alla reazione. Se il dNTP è complementare al filamento stampo, la sua incorporazione causa il rilascio di un pirofosfato che viene prelevato dalla sulfurilasi e converito in ATP. L'ATP ottenuto innesca la reazione luminosa della luciferasi (vedi didascalia Figura 3) e il segnale luminoso emesso viene captato e misurato da un sensore CCD. In base al numero di nucleotidi incorporati sono rilasciati più pirofosfati e quindi più molecole di ATP, generando una bioluminescenza più intensa. Misurata l'intensità luminosa è dunque possibile determinare il numero di nucleotidi di uno stesso tipo che si susseguono nella sequenza. Ad esempio, l'incorporazione della sequenza AAA produrrà un'emissione di luce tre volte più intensa rispetto all'incorporazione di un singolo nucleotide A. Se il dNTP non è complementare non verrà incorporato e non verrà generata nessuna emissione luminosa; quindi l'assenza di picchi indica che nella posizione considerata non è presente quel dNTP. Dopo la registrazione dell'emissione luminosa, i rimanenti nucleotidi liberi vengono degradati da apirasi e viene fornito il dNTP successivo, proseguendo la reazione di polimerizzazione. La Figura 3 riassume graficamente un singolo ciclo di pirosequenziamento.

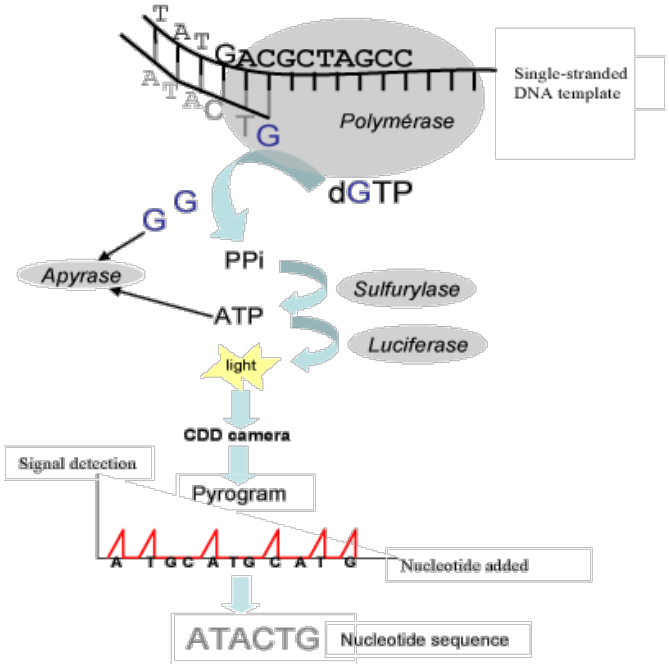


Figura 3. Rappresentazione del processo di pirosequenziamento. L'incorporazione del dGTP provoca il rilascio di un pirofosfato (PPi) che viene convertito in ATP dalla sulfurilasi. L'ATP è coinvolto nella reazione di bioluminescenza ad opera della luciferasi che converte l'energia chmica in energia luminosa. L'emissione di luce è captata da un sensore CCD e elaborata per ottenere la sequenza nucleotidica. I dNTP non incorporati e la molecola di ATP vengono degradati da apirasi. Immagine presa da [13].

A ogni ciclo è aggiunta una singola specie di nucleotidi (A, C, G o T); al ciclo successivo sarà poi introdotta una delle specie restanti, e così via ripetendo la sequenza di introduzione per centinaia di cicli.

Il tasso di errore medio è compreso tra 10^{-3} e 10^{-4}, superiore al tasso di errore del metodo Sanger ma è comunque la migliore prestazione ottenuta nei metodi SGS confrontati in questo capitolo. Una rilevante fonte di errore, oltre alle fasi di amplificazione dei frammenti, è la preparazione delle sfere (in particolare la fase con emulsione PCR) che possono trasportare copie di sequenze diverse producendo read che non rappresentano la sequenza reale.

Genome Sequencer permette di sequenziare 1,5 milioni di sfere per run, determinando sequenze di lunghezza variabile in base al numero di cicli di sequenziamento (limitati per mantenimento dell'efficienza di polimerasi e luciferasi). Le più recenti versioni del sequenziatore (GS FLX Titanium XL+ [14]) ottengono read lunghe fino a 1000 nt con throughput di 700 Mb.

2.4 Il sequenziatore Genome Analyzer di Illumina

Introdotta nel mercato del 2006 da Solexa e in seguito rinominata Genome Analyzer da Illumina, questa piattaforma di sequenziamento ha origine dal lavoro di Turcatti *et al.* [15] sul sequenziamento a terminazione reversibile.

La preparazione dei campioni prevede la denaturazione della doppia elica e l'aggiunta degli adattatori, ottenendo un filamento stampo della sequenza da sequenziare.

Le piastre di sequenziamento o flow-cell di Genome Analyzer sono composte di otto lane indipendenti, sulle cui superfici sono immobilizzati due diversi oligonucleotidi. I frammenti sono immessi sulla piastra di sequenziamento, permettendo l'ibridazione tra i loro adattatori e i complementari oligonucleotidi della piastra. Il legame si forma in entrambe le estremità dei frammenti, che sono così immobilizzati sulla superficie della flow-cell, assumendo una forma "a ponte" (vedi Figura 4). Dopo l'immobilizzazione ha inizio il processo di

amplificazione: DNA polimerasi sintetizza il filamento complementare, la molecola ottenuta viene denaturata e i due filamenti rimangono legati alla piastra a una sola estremità. L'operazione si ripete ciclicamente: i due frammenti si "piegano" fino a ibridare la loro estremità rimasta libera con un complementare oligonucleotide della flow-cell, riassumendo nuovamente la forma a ponte e permettendo la sintesi di un nuovo filamento complementare. I passaggi di immobilizzazione-sintesi-denaturazione continuano fino a ottenere un cluster di migliaia di frammenti, legati a una estremità al substrato e raccolti in uno spazio molto limitato. Il processo (rappresentato in Figura 4) è definito amplificazione bridge-PCR, proprio in virtù della forma "a ponte" che i filamenti assumono quando entrambi gli adattatori sono legati al substrato.

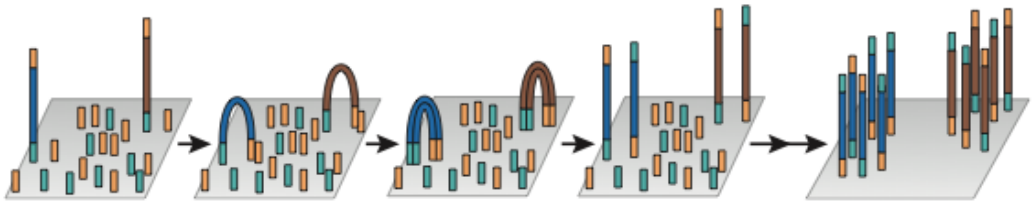


Figura 4. Processo di amplificazione clonale nella tecnologia Illumina (bridge PCR). I frammenti arricchiti con adattatori (in giallo e verde) sono posti a contatto con la flow-cell e ibridano gli oligonucleotidi presenti sulla superficie, assumendo una forma "a ponte". Le fasi di amplificazione producono il filamento complementare, quindi la separazione dei filamenti e la ripetizione ciclica, fino a ottenere i cluster di copie identiche dello stesso filamento. Immagine presa da [11]

I cluster ottenuti sono composti sia di filamenti uguali al filamento originale sia di filamenti inversi; questi ultimi vengono rimossi, ottenendo cluster di circa 1000 copie identiche della stessa sequenza. Ai frammenti di ogni cluster viene eseguito l'annealing del primer di sequenziamento che permette l'avvio della reazione di sequenziamento vera e propria. Ogni ciclo di sequenziamento coinvolge una DNA polimerasi e i quattro dNTP a cui sono state apportate due modificazioni. La doppia modifica consiste nell'incorporazione di un marcatore fluorescente che ne permette l'identificazione e nell'aggiunta di un terminatore reversibile. Il terminatore è una molecola che blocca il gruppo ossidrile impedendo l'ulteriore sintesi, in modo da garantire l'incorporazione di una sola base. Questa limitazione è necessaria poiché in ogni ciclo sono forniti tutti i dNTP e quindi la sintesi potrebbe riguardare più basi in contemporanea; il terminatore permette invece di arrestare la sintesi dopo ogni singola incorporazione. Il terminatore è detto reversibile in quanto può essere dissociato chimicamente, riattivando la sintesi. Dopo ogni incorporazione, un laser eccita il fluorescente del dNTP generando un'emissione luminosa che ne permette l'identificazione. Quindi il terminatore viene rimosso, continuando il sequenziamento della base successiva.

Le prime versioni di Illumina Genome Analyzer erano caratterizzate de un tasso di errore non trascurabile e da read di lunghezza molto inferiore rispetto a 454/Roche (35-50 nt) ma erano in grado di ottenere dati con un throughput di oltre 10000 Mb al giorno (oltre 10 volte più di 454/Roche) a costi molto contenuti. Le modifiche apportate alla tecnologia e ai protocolli sperimentali hanno sensibilmente ridotto gli errori di sequenziamento, seppure siano ancora presenti dei bias [16], e aumentato la lunghezza delle read a 100 nt e oltre, facendola diventare la tecnologia più utilizzata nei progetti di ri-sequenziamento e RNA sequencing.

2.5 Il sequenziatore SOLiD di Applied Biosystems

Sviluppata sul metodo *sequencing-by-ligation* (a differenza dei due metodi sopra descritti che possono essere definiti *sequencing-by-synthesis*) proposto dal laboratorio George Church nel 2005 [17], la piattaforma SOLiD viene introdotta nel mercato a partire dal 2007 da Applied Biosystems.

Dopo la denaturazione della doppia elica e l'aggiunta di adattatori, i frammenti sono ibridati a delle sfere la cui superficie è ricoperta da oligonucleotidi complementari a uno degli

adattatori, come in 454/Roche. Nella tecnologia SOLiD le sfere sono però molto più piccole (diametro 0,75 µm invece di 28 µm), permettendo una maggiore densità di sfere sequenziabili per run. Le sfere sono messe in emulsione assieme ai reagenti e ai primer della PCR per ottenere l'amplificazione clonale dei filamenti (em-PCR), in modo del tutto simile a quanto visto per 454/Roche.

A differenza di quest'ultima tecnologia, la piastra di sequenziamento SOLiD non dispone di pozzetti; le sfere sono invece legate alla superficie opportunamente trattata della piastra di sequenziamento tramite la formazione di un legame covalente con le estremità dei filamenti amplificati, modificate chimicamente [18]. Mentre in 454/Roche esiste un numero predefinito di pozzetti e quindi di sfere sequenziabili, con il sequenziatore SOLiD il limite al numero di sfere che possono essere deposte sulla superficie è dato solo dal loro diametro (inferiore a 1 µm).

La reazione di sequenziamento non avviene grazie a DNA polimerasi, ma tramite DNA ligasi, un enzima in grado di legare covalentemente due frammenti di DNA. L'acronimo SOLiD significa infatti "Sequencing by Oligonucleotide Ligation and Detection". Ogni ciclo di sequenziamento coinvolge la sfera, un primer di sequenziamento, la ligasi e quattro sonde di dNTP. Le sonde sono composte da 8 basi più un marcatore fluorescente e presentano un sito di rottura fra i nucleotidi 5 e 6. Le prime due basi della sonda sono complementari a due nucleotidi da sequenziare, le basi da 3 a 5 sono degeneri, cioè capaci di appaiarsi a qualsiasi base della sequenza stampo, le basi 6, 7 e 8 sono anch'esse degeneri e possono essere rimosse, assieme al marcatore fluorescente, tramite la rottura del legame con la base in posizione 5. Una volta che una sonda si è appaiata alla sequenza stampo, avviene l'eccitazione del marcatore legato alla sonda e la conseguente emissione di fluorescenza. Contemporaneamente, viene rotto il legame tra le posizioni 5 e 6, lasciando libero l'estremità 5' della quinta base della sonda e permettendo il legame con una successiva sonda. L'emissione luminosa è rilevata dalla strumentazione in modo da identificare il colore associato alla coppia delle prime due basi. Ogni marcatore produce una specifica emissione luminosa che identifica 4 di 16 possibili coppie di nucleotidi (4 possibili nucleotidi su sue posizioni, si veda Figura 5).

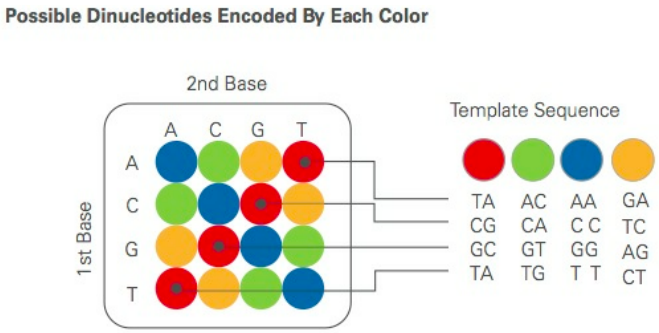


Figura 5. Le possibili combinazioni di nucleotidi che i quattro colori dei marcatori fluorescenti rappresentano. Ciascun colore non rappresenta univocamente una singola coppia di basi, ma 4 combinazioni sulle 16 possibili. Figura presa da [19].

Detta *n* la posizione della base del primer (appaiato all'adattatore che lega la sequenza stampo alla sfera), le basi *n+1* e *n+2* saranno quindi complementari alla sequenza stampo, così come le successive basi *n+6* e *n+7* e così via. Per ciascuna di queste coppie di basi, lo strumento rileva un segnale che può corrispondere a una delle sedici combinazioni possibili (Figura 3). Le basi *n+3*, *n+4* e *n+5* sono anch'esse appaiate (in quanto degeneri) ma indeterminate. La loro identificazione sarà resa possibile in un successivo ciclo di sequenziamento, spostando il primer in una nuova posizione della sequenza stampo. Dopo l'appaiamento della prima sonda, il processo si ripete con il legame di nuove sonde, l'identificazione e la separazione del marcatore e degli ultimi tre nucleotidi; il numero delle iterazioni (solitamente 7) determina la lunghezza di lettura. Terminato il processo, il

filamento ottenuto è rimosso e un nuovo primer viene appaiato al filamento stampo in posizione *n-1*. Nel nuovo ciclo saranno univocamente appaiate le basi *n+2* e *n+3*, nel ciclo successivo le basi *n+3* e *n+4* e così via. Nel complesso sono eseguiti cinque cicli di sequenziamento, ciascuno con il primer spostato in una posizione di volta in volta arretrata. I cinque cicli permettono di interrogare ogni base della sequenza per due volte (in due indipendenti reazioni con diverse posizioni dei primer). Si veda ad esempio la Figura 6: la base nella posizione di lettura 5 viene coinvolta dal primer numero 2 nel secondo ciclo (indicato in azzurro) e dal primer numero 3 nel primo ciclo (indicato in blu). Questa doppia interrogazione permette di migliorare la bontà del sequenziamento.

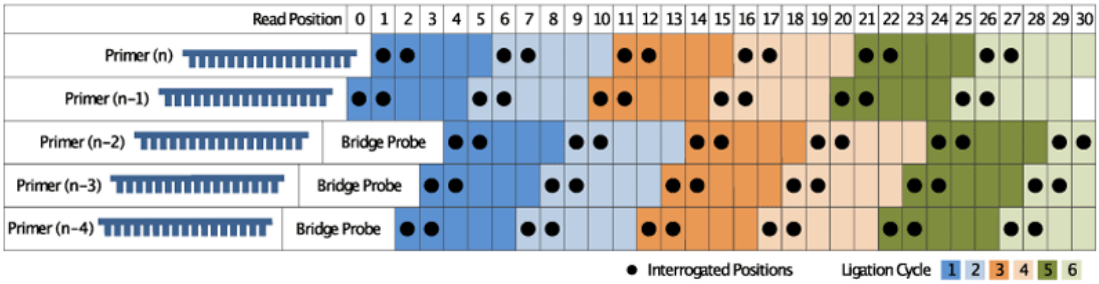


Figura 6. Design del sequenziamento con tecnologia SOLiD. Sono rappresentati i cinque cicli (uno per riga) di sequenziamento, ciascuno dei quali composto da sei passaggi di aggiunta di nuove sonde ad opera della ligasi. I punti neri indicano le coppie di basi interrogate in ogni passaggio. Figura presa da [19].

Il colore dell'emissione da parte del fluorescente permette di identificare la coppia delle prime basi, ma non è sufficiente per distinguere un nucleotide dall'alto. Oltre alla specifica emissione per ogni coppia è infatti necessario che una delle basi della sequenzia sia nota. La base nota è incorporata nell'ultimo (il quinto) ciclo di sequenziamento e corrisponde all'ultimo nucleotide del primer. Quindi, dato che ciascun colore rappresenta 4 possibili coppie di dNTP (Figura 5) e per ognuna di esse il secondo nucleotide coincide con la prima base della coppia successiva, conoscere una base della sequenza permette di interpretare tutta la sequenza nella sua interezza. La Figura 7 propone un esempio di decodifica dei colori rilevati, nota la prima base della sequenza.

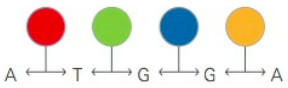


Figura 7. Esempio di decodifica dei colori per determinare la sequenza. Si suppongono note la prima base (A) e l'ordine dei colori rilevati. Dallo schema di codifica (Figura 5) è evidente che se la coppia è identificata dal colore rosso e la prima base è una A allora sicuramente la seconda base sarà un T. Nella seconda coppia, la prima base è la T appena identificata e il colore della coppia è verde: la seconda base sicuramente sarà una G. Figura presa da [19].

Le fonti possibili di errore del sequenziatore SOLiD sono diverse: errore di background nella fase di amplificazione, letture di bassa qualità a causa di presenze di sequenze diverse su una stessa sfera, oltre ad artefatti introdotti nel processo ciclico di ligasi e rimozione del marcatore. Il tasso medio di errore vari da 10^{-2} a 10^{-4}. La tecnologoia SOLID sequenzia oltre 300 milioni di sfere in parallelo, generando read di lunghezza pari a 50 nt, quindi molto corte. Le prestazioni in termini di throughput (20 Gb/giorno) e di costo di sequenziamento sono superiori a quelle degli altri sequenziatori NGS [20].

2.6 I sequenziatori di terza generazione

Le due principali limitazioni delle tecnologie finora descritte sono rappresentate dalla lunghezza delle read, generalmente molto limitata, e dal bias introdotto dalla fase di amplificazione. Il primo aspetto risulta critico in caso di sequenziamento *de novo*, in cui l'assemblaggio delle read per ricostruire il genoma d'interesse risulta piuttosto complicato se

si hanno a disposizione read molto corte che non permettono di distinguere zone ripetute del genoma sequenziato. Il secondo aspetto penalizza invece tutte le applicazioni di trascritto mica quantitativa basate su sequenziamento come l'RNA-seq, poiché modificano l'abbondanza relativa delle trascritti contenuti nel campione d'interesse.

Negli ultimi anni, molti sforzi sono stati dedicati allo sviluppo di una nuova generazione di tecnologie, denominate sequenziamento a singola molecola (SMS, *Single Molecule Sequencing*). Le tecnologie di questa nuova generazione si possono suddividere in tre gruppi: (i) metodi basati sulla fluorescenza come true single-molecule sequencing (tSMS) e single-molecule real-time sequencing (SMRT); (ii) metodi non basati su fluorescenza come Nano-edges di Nanopore e sequenziamento tramite microscopia elettronica; (iii) metodi basati sulla spettroscopia Raman [21]. Le tecnologie SMS sono anche note come TGS (Third Generarion Sequencing), proprio a rappresentare come queste siano un'evoluzione della generazione precedente offrendo read più lunghe, più alto throughput e maggior accuratezza oltre a una riduzione di tempo impegnato e costi.

3 RNA Sequencing

In un esperimento RNA-Seq, le read rappresentano i dati grezzi dai quali ricavare l'informazione sul livello di espressione dei geni nel campione. Più numerose sono le copie di un trascritto in un campione, più probabilità avrà quel trascritto di essere sequenziato e di generare read. Per quantificare il numero di read riferite a ciascun gene, le read di ogni campione sono mappate su un genoma o un trascrittoma di riferimento. La fase di allineamento delle read sul riferimento presenta un aspetto critico: idealmente si vorrebbe trovare l'univoca posizione del genoma in cui il riferimento sia identico alla read. In realtà il riferimento non sarà mai una rappresentazione perfettamente identica del campione biologico, a causa di errori di sequenziamento e/o di imperfetta similarità tra le sequenze del campione d'interessa e quelle del riferimento. Lo scopo dell'allineamento diventa quindi l'identificazione della posizione del genoma in cui ogni read ottiene il miglior match con il riferimento. Una volta determinate le posizioni delle read sul riferimento, è possibile contare il numero di read allineate su un gene (o trascritto o esone). Il totale delle read allineate su un gene è detto appunto "count" e può essere inteso come una misura del livello di espressione del gene stesso.

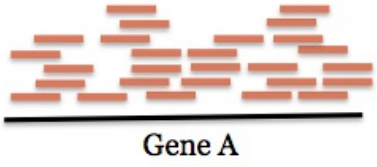


Figura 8. Rappresentazione grafica di un count. Le read (indicate con le barrette arancioni) vengono allineate sul gene di riferimeno. Il numero di read mappate è detto count. Nella rappresentazione grafica, il count del gene A è 27, cioè il numero di barrette arancioni associate al gene A.

I count sono i dati finali di un esperimento di RNA-seq per la quantificazione del trascrittoma e sono memorizzati in matrici in cui le righe rappresentano i diversi geni e le colonne i diversi campioni sequenziati. Per loro stessa definizione, dal punto di vista statistico i count rappresentano una somma di eventi aleatori indipendenti (la mappatura di ogni read sui geni). Possono quindi essere descritti da una variabile aleatoria che segue una determinata distribuzione statistica. I due modelli di distribuzione più utilizzati per descrivere i count sono il modello di Poisson e il modello Binomiale Negativo.

3.1 Modello di Poisson

Jiang *et al.* [22] hanno proposto un modello di Poisson dei count calcolati sui trascritti presenti in un campione sottoposto a RNA-seq. Si descrive di seguito il modello proposto, che può essere esteso ai geni semplicemente assegnando a ciascun gene la somma dei count sui trascritti corrispondenti.

Sia F l'insieme dei trascritti presenti nel campione sequenziato. Per ogni trascritto $f \in F$, siano l_f la sua lunghezza (il numero di basi che lo compone) e k_f il numero di copie presenti nel campione. Utilizzando questa notazione, i nucleotidi totali dei trascritti nel campione sono $\sum_{f \in F} k_f l_f$. Sotto l'ipotesi che il sequenziamento possa essere modellizzato come un processo di campionamento casuale, in cui ogni read è campionata indipendentemente ed uniformemente da ogni possibile nucleotide nel campione, la probabilità che una read provenga da uno specifico trascritto è $p_f = k_f l_f / \sum_{f \in F} k_f l_f$. Definendo l'abbondanza relativa del trascritto f come $\beta_f = k_f / \sum_{f \in F} k_f l_f$, è possibile riscrivere $p_f = l_f \beta_f$.

Sia ora s la library size del campione, cioè il numero totale di read nel campione. Sotto l'ipotesi di campionamento uniforme, il numero di read mappate sul trascritto f, descritto dalla variabile aleatoria X_f, segue una distribuzione binomiale $\mathcal{B}(s, p_f)$ con parametri s e

$p_f = l_f \beta_f$. Negli esperimenti RNA-Seq generalmente s è molto grande e p_f molto piccolo, quindi la distribuzione binomiale può essere ben approssimata da una distribuzione di Poisson di parametro $\mu_f = sp_f = sl_f\beta_f$. La distribuzione di Poisson $\mathcal{P}(\mu) = e^{-\mu}\frac{\mu^s}{s!}$ è caratterizzata dall'unico parametro μ, che ne identifica sia la media che la varianza. Infatti, per una variabile aleatoria X di distribuzione di Poisson si ha:

$$E[X] = \sum se^{-\mu}\frac{\mu^s}{s!} = \mu e^{-\mu}\sum\frac{\mu^{s-1}}{(s-1)!} = \mu$$

$$Var(X) = E[X^2] - E[X]^2 = \sum s^2 e^{-\mu}\frac{\mu^s}{s!} - \mu^2 = \mu^2 + \mu - \mu^2 = \mu = E[X]$$

La notazione fin qui utilizzata fa riferimento ai trascritti valutati su un singolo campione. In realtà, è corretto considerare più campioni, i quali possono a loro volta appartenere a diversi gruppi sperimentali (ad esempio campioni provenienti da diversi tessuti, o da colture cellulari di tipo controllo/trattato). Indicando dunque con *j* il campione e con *ρ(j)* il gruppo sperimentale a cui esso appartiene, la distribuzione dei count per il trascritto *f* nel campione *j* può essere descritta dalla variabile aleatoria $X_{fj} \sim \mathcal{P}(\mu_{fj})$, con $\mu_{fj} = s_j l_f \beta_{f\rho(j)}$ in cui s_j è la library size del campione *j*-esimo e $\beta_{f\rho(j)}$ il livello di espressione del trascritto f nella condizione *ρ(j)*.

3.2 Modello Binomiale Negativo

Il modello di Poisson sopra descritto fornisce una prima approssimazione dei dati. I dati di RNA-Seq esprimono però solitamente una variabilità non trascurabile, dovuta al contributo di due termini: variabilità tecnica e biologica. La variabilità tecnica comprende tutte quelle variazioni di misura imputabili a errori o imprecisioni introdotte dalla strumentazione, dalla manipolazione dei campioni o dall'elaborazione dei dati. La variabilità biologica è invece la variazione che si manifesta nella misura dell'espressione genica in condizioni o soggetti diversi. Si supponga ad esempio di voler determinare il livello di espressione di un gene in un particolare tessuto umano e si immagini di ripetere la misura su cellule di individui diversi, su cellule diverse di uno stesso individuo o anche sulla stessa cellula ma in istanti diversi. Le misure ottenute, pur riferite allo stesso tipo di cellula (lo stesso tessuto umano), saranno diverse fra loro a causa dalla diversa espressione del gene nell'individuo/istante considerato. La variabilità biologica è dunque espressione della variabilità genica e non è eliminabile. La variabilità complessiva può essere ridotta migliorando le procedure sperimentali e le fasi di analisi, ma non può essere completamente eliminata; è per questo importante considerarla e poterla stimare. È già stato dimostrato che la condizione stringente di uguaglianza tra media e varianza prevista dal modello di Poisson non è sufficiente a descrivere adeguatamente la variabilità dei dati di RNA-Seq [23] [24] [25]. Se la varianza è maggiore della media, i dati si dicono *sovradispersi* e l'assunzione di distribuzione Poissoniana è inappropriata.

Con l'obiettivo di descrivere i dati di RNA-seq anche in presenza di sovra dispersione, Robinson *et al.* [26] hanno proposto un modello dei count basato su distribuzione binomiale negativa $X_{gj} \sim \mathcal{NB}(s_j p_{g\rho(j)}, \phi_g)$ per il gene g ed il campione *j*; s_j è la library size del campione *j*, ϕ_g la dispersione del gene g e $p_{g\rho(j)}$ l'abbondanza relativa del gene g rispetto al totale dei geni misurata nella condizione sperimentale $\rho(j)$ a cui il campione appartiene. Sotto questa parametrizzazione, la media è $\mu_{gj} = s_j p_{g\rho(j)}$ e la varianza è $\mu_{gj}(1 + \mu_{gj}\phi_g)$. Il fattore ϕ_g rappresenta il coefficiente di variazione della variabilità biologica che si manifesta tra i diversi soggetti e permette di separare gli effetti della variabilità biologica da quelli dovuti a variabilità tecnica. Evidentemente, la distribuzione binomiale negativa si riduce ad una distribuzione di Poisson quando $\phi_g = 0$.

Una delle applicazioni più interessanti dell'RNA-seq è l'analisi differenziale dell'espressione genica. Scopo dell'analisi differenziale è identificare quali geni mostrano significative differenze nel livello di espressione (quindi nell'abbondanza dei loro trascritti) fra due o più

condizioni sperimentali. Il termine "significativa" esprime un concetto fondamentale: infatti non è semplice stabilire se la differenza esistente tra i count di diversi gruppi sperimentali sia significativa in senso statistico. La verifica dell'espressione differenziale avviene utilizzando i test di ipotesi, metodi statistici che valutano, sotto opportune ipotesi e livelli di "affidabilità", se la differenza osservata fra due campioni possa o meno essere considerata significativa. Nelle applicazioni RNA-Seq, ciò significa utilizzare come dati le tabelle dei count dei gruppi sperimentali e applicarvi dei test di ipotesi. I diversi metodi di analisi differenziale differiscono per il modello adottato nella descrizione dei dati, per le assunzioni poste e per gli approcci con cui l'analisi viene condotta. Il risultato dei metodi di analisi differenziale è una lista ordinata di geni, per ciascuno dei quali sono determinati un indice della differenza osservata (fold-change) e il livello di significatività osservato (p-value). Un gene si dirà differenzialmente espresso se il relativo p-value osservato è inferiore a un fissato valore di soglia. Il valore di soglia del test è stabilito tenendo conto della probabilità di errore che si è disposti ad accettare nell'identificazione dell'espressione differenziale; tipicamente si sceglie di controllare l'errore di tipo I (falsi positivi). È evidente come la scelta del valore di soglia sia un aspetto critico: a differenti livelli di soglia corrisponderanno diverse identificazioni dei geni differenzialmente espressi.

4 Metodi per l'analisi dell'espressione genica differenziale

4.1 DEGSeq

DEGseq è un pacchetto R [27] per l'identificazione dei geni differenzialmente espressi a partire da dati di RNA-Seq, sviluppato da Wang *et al.* [10].

L'analisi differenziale implementata da DEGseq è basata sui grafici MA-plot, strumenti di diagnostica della qualità e della necessità di normalizzare i dati, molto utilizzati per l'analisi dei dati da microarray. Gli MA-plot sono utilizzati per studiare il rapporto tra la differenza di due variabili (considerate su scala logaritmica) e la media di queste, indicate rispettivamente con M (da "*minus*") e A (da "*average*"). Gli MA-plot, o MvA-plot, sono degli scatterplot dei valori di M (asse y) rispetto ad A (asse x).

In un esperimento RNA-Seq, detti X_1 e X_2 i count di un gene ottenuti in due diverse condizioni sperimentali, si definiscono $M = \log_2(X_1) - \log_2(X_2)$ e $A = [\log_2(X_1) + \log_2(X_2)]/2$. Per ogni gene, le coppie (A, M) determinate vengono riportate sull'MA-plot. Dato che solo una parte minoritaria dei geni considerati sarà differenzialmente espressa, la maggior parte dei punti dell'MA-plot assumerà valore $M \cong 0$ mentre i geni con diversa espressione diffenziale avranno valori di M molto diversi da 0.

DEGSeq offre due diversi approcci per l'analisi dell'espressione differenziale: I metodi MARS e MATR. Il metodo MARS (*MA-plot based with Random Sampling model*) si basa sull'ipotesi di campionamento casuale e uniforme delle read dai trascritti [22] e ne descrive la distribuzione con un modello binomiale (approssimato in realtà da un modello di Poisson). Il metodo MATR (*MA-plot based with Technical Replicates*), si serve invece delle repliche tecniche per ottenere una stima della variabilità tecnica, che viene utilizzata in seguito per identificare i geni differenzialmente espressi in maniera più robusta.

4.1.1 Metodo MA-plot based con modello a campionamento casuale (MARS)

I count di uno specifico gene su due campioni appartenenti a diverse condizioni sperimentali possono essere descritti da una distribuzione binomiale, ovvero: $X_j \sim \mathcal{B}(s_j, p_j)$, j = 1,2 in cui s_j e p_i rappresentano, rispettivamente, la library size del campione (il numero totale di read nel campione) e la probabilità che una read provenga da quel gene. X_1 e X_2 sono assunti indipendenti. Ponendo $Z = \log_2(X_1)$ e $Y = \log_2(X_2)$ si ridefiniscono $M = Z - Y$ e $A = (Z + Y)/2$.

Per s_1 e s_2 sufficientemente elevati, è possibile ottenere la distribuzione asintitotica di X_1/s_1 e X_2/s_2:

$$\sqrt{s_1}\left(\frac{X_1}{s_1} - p_1\right) \to \mathcal{N}(0, p_1(1-p_1))$$

$$\sqrt{s_2}\left(\frac{X_2}{s_2} - p_2\right) \to \mathcal{N}(0, p_2(1-p_2))$$

Allora $Z = h(X_1/s_1) = \log_2(s_1 X_1/s_1)$ dove $h(x) = \log_2(s_1 x)$. É possibile ottenere la distribuzione asintotica di Z quando $s_1 \to \infty$, cioè

$$\sqrt{s_1}(Z - \log_2(s_1 p_1)) = \sqrt{s_1}\left(h\left(\frac{X_1}{s_1}\right) - h(p_1)\right) \to \mathcal{N}(0, p_1(1-p_1)[h'(p_1)]^2)$$
$$= \mathcal{N}\left(0, \left(\frac{1-p_1}{p_1}\right)(\log_2 e)^2\right)$$

Quindi,

$$Z \to \mathcal{N}\left(\log_2(s_1 p_{1,}), \left(\frac{1-p_1}{s_1 p_1}\right)(\log_2 e)^2\right)$$

Analogalmente, la distribuzione asintotica di *Y* può essere ottenuta come

$$Y \to \mathcal{N}\left(\log_2(s_2 p_{2,}), \left(\frac{1-p_2}{s_2 p_2}\right)(\log_2 e)^2\right)$$

Dimostrato che Z e Y seguono approssimativamente una distribuzione normale (per s_j sufficientemente elevati), è possibile utilizzare la notazione

$$Z \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$$
$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

Sotto l'assunzione di X_1 e X_2 indipendenti (quindi anche Z e Y sono indipendenti), le distribuzioni di M e A possono essere ottenute come:

$$M \sim N(\mu_Z - \mu_Y, \sigma_Z^2 + \sigma_Y^2) = N(\mu_M, \sigma_M^2)$$
$$A \sim N\left(\frac{1}{2}(\mu_Z + \mu_Y), \frac{1}{4}(\sigma_Z^2 + \sigma_Y^2)\right) = N(\mu_A, \sigma_A^2)$$

La distribuzione condizionale di M dato $A = a$ può quindi essere ottenuta per ogni valore di a come:

$$M|(A=a) \sim N\left(\mu_M + \gamma \frac{\sigma_M}{\sigma_A}(a - \mu_A), \sigma_M^2(1-\gamma^2)\right)$$

in cui γ rappresenta il coeeficiente di correlazione tra M e A:

$$\gamma = \frac{Cov(M,A)}{\sigma_M \sigma_A} = \frac{\sigma_Z^2 - \sigma_Y^2}{\sigma_Z^2 + \sigma_Y^2}$$

Allora, il valore atteso e la varianza di $M|A$ sono dati da:

$$E(M|A=a) = \mu_M + \gamma \frac{\sigma_M}{\sigma_A}(a - \mu_A) = \mu_Z - \mu_Y + 2\frac{\sigma_Z^2 - \sigma_Y^2}{\sigma_Z^2 + \sigma_Y^2}\left(a - \frac{1}{2}(\mu_Z + \mu_Y)\right)$$

$$Var(M|A=a) = \sigma_M^2(1-\rho) = 4\frac{\sigma_Z^2 \sigma_Y^2}{\sigma_Z^2 + \sigma_Y^2}$$

Per ciascun gene con $(A = a, M = m)$ sull'MA-plot, viene valutato il test di ipotesi $H_0: p_1 = p_2 = p$ contro $H_1: p_1 \neq p_2$. Sotto l'ipotesi nulla si ha $\mu_A = \frac{1}{2}(\mu_Z + \mu_Y) = \frac{1}{2}\log_2(s_1 s_2 p^2)$ e quindi $p = \sqrt{2^{2\mu_A}/(s_1 s_2)}$. Usando il valore a come stima di μ_A si ottiene lo stimatore $\hat{p} = \sqrt{2^{2a}/(s_1 s_2)}$.

Di conseguenza gli stimatori $\hat{E}(M|A=a)$ e $\widehat{Var}(M|A=a)$ possono essere riscritti come

$$\hat{\hat{E}}(M|A=a) = \log_2(s_1 \hat{p}) - \log_2(s_2 \hat{p}) = \log_2(s_1) - \log_2(s_2)$$

$$\widehat{Var}(M|A=a) = \frac{4\left(1 - \sqrt{2^{2a}/(s_1 s_2)}\right)(\log_2 e)^2}{(s_1 + s_2)\sqrt{2^{2a}/(s_1 s_2)}}$$

Dagli stimatori di media e varianza è possibile standardizzare la distribuzione di $M|(A=a)$ determinando uno *Z-score*, per ogni gene con $(A = a, M = m)$:

$$Z - score = \frac{\left|m - \hat{\hat{E}}(M|A=a)\right|}{\sqrt{\widehat{Var}(M|A=a)}}$$

Dalle tavole della normale standardizzata, in corrispondenza dello *Z-score* calcolato si ricava il *p-value* da confrontare con il livello di significatività desiderato per determinare se il gene

sia differenzialmente espresso o meno. Fissata una soglia per lo *Z-score*, ad esempio 4 (valore di default di DEGseq, pari ad un livello di significatività di 0,0032%), vengono tracciate sull'MA-plot le curve che descrivono 4 volte la deviazione standard di M in dipendenza da A (*4sd-curves*). Le 4sd-curves vengono calcolate grazie all'equazione:

$$m_{1,2}^{MARS} = \hat{\hat{E}}(M|A=a) \pm 4\sqrt{\widehat{Var}(M|A=a)}$$

$$= \log_2(s_1) - \log_2(s_2) \pm 4\sqrt{\frac{4\left(1-\sqrt{2^{2a}/(s_1 s_2)}\right)(\log_2 e)^2}{(s_1+s_2)\sqrt{2^{2a}/(s_1 s_2)}}}$$

Le 4sd-curves vengono definite *teoriche* dato che sono calcolate in accordo con il modello a campionamento casuale.

4.1.2 Metodo MA-plot based con repliche tecniche (MATR)

DEGseq prevede la possibilità di stimare il livello di rumore di geni con differente intensità o identificare le differenze di espressione genica in diversi gruppi, tramite il metodo MATR, anch'esso basato su MA-plot. La differenza rispetto al metodo precedente consiste nel fatto che, mentre MARS ottiene una stima teorica della dispersione, MATR stima la dispersione direttamente dalle repliche tecniche.

Viene dapprima tracciato un MA-plot usando due repliche tecniche (i count di uno stesso campione biologico, sequenziate però in due diverse lane/run). Al grafico è applicata una finestra scorrevole, comprendente l'1% dei punti, che viene fatta scorrere lungo l'asse A. Per ogni finestra centrata in $A = a$ si può stimare la variazione locale di $M|(A = a)$ su tutti i valori di M compresi nella finestra. Sulle stime delle variazioni locali di ciascuna finestra viene fatta una regressione *loess* per ottenere una stima della deviazione standard locale di $M|(A = a)$. La deviazione standard locale σ_a di $M|(A = a)$ viene quindi usata per confrontare le differenze osservate tra due diverse librerie.

Viene tracciato un secondo MA-plot dei dati di due diversi gruppi. Per ogni gene g avente $(A = a_g, M = m_g)$ è calcolato uno $Z - score = |m_g - \mu_g|/\sigma_g$ in cui μ_g è la media locale di M e σ_g è la deviazione standard locale di $M|(A = a_g)$ stimata al passo precedente. Come per il metodo precedente, dallo Z-score si ricava un p-value che viene confrontato con un livello di significatività fissato dall'utente.

Sull'MA-plot sono infine riportate due curve che rappresentano le *4sd-curves* di M teoriche, ovvero stimate in ipotesi di campionamento uniforme.

4.1.3 Correzione per test multipli

In entrambi i metodi è applicata una correzione per test multipli che determina per ogni p-value un corrispondente *q-value* secondo FDR[1]. DEGseq accetta in input sia p-value che q-value come valore di soglia per selezionare i geni differenzialmente espressi.

4.1.4 Metodo di verifica fra repliche tecniche (CTR)

DEGseq dispone inoltre del metodo CTR (*Check between Technical Replicates*) che non implementa un metodo per la selezione dei geni differenzialmente espressi ma consente di

[1] *FDR (False Discovery Rate)*: metodo statistico utilizzato nei test di ipotesi per correggere l'errore di tipo I (falsi positivi) nei confronti multipli. In una lista di test multipli che rifiutano l'ipotesi nulla, FDR controlla la proporzione attesa di falsi positivi (cioè rifiuti errati dell'ipotesi nulla). Il q-value è definito come l'analogo, secondo FDR, del p-value.

valutare se la variazione tra due repliche tecniche di uno stesso campione può essere descritta dal modello di campionamento uniforme di Jiang *et al.* [22]

4.2 edgeR

edgeR è un pacchetto R che implementa il metodo di Robinson *et al.* [24] per l'analisi e la quantificazione dell'espressione differenziale in dati "digitali" di espressione genica (*digital gene expression* o DGE). edgeR, il cui acronimo significa proprio "empirical analysis of DGE in R", è applicabile a dati RNA-seq, Tag-Seq, SAGE, CAGE e ChIP-Seq, in cui l'abbondanza delle feature di interesse, come ad esempio i trascritti, è misurata in termini di count.

Il pacchetto implementa metodi statistici esatti sviluppati da Robinson e Smyth [28] [29] per la valutazione dell'espressione differenziale tra più gruppi. La nuova versione di edgeR implementa inoltre metodi basati su modello lineare generalizzato (GLM) recentemente proposti da McCarty *et al*. [30] e utili in caso di design sperimentali più complessi. In questo lavoro considereremo solo la prima versione di edgeR, ovvero quella basato sul metodo proposto da Robinson e Smyth [28] [29]. Il modello dei dati proposto per descrivere i count è di tipo Binomiale Negativo, più adatto di quello di Poisson nel descrivere la variabilità tecnica e biologica delle repliche (sovradispersione).

Il parametro dispersione del modello viene stimato dai dati con l'utilizzo di metodi empirici Bayesiani, che permettono la stima della variazione biologica per singolo gene, o *genewise*, in esperimenti che comprendano repliche biologiche. Il pacchetto comprende inoltre un metodo per lo scaling dei dati basato su "Trimmed Mean of M values" (TMM) [31].

4.2.1 Modello Binomiale Negativo dei count

In edgeR, i count X_{gj} per ciascun gene g nel campione j (ovvero la somma delle read ottenute nel campione j e mappate sul gene g) sono descritti con un modello negativo binomiale in grado di catturare anche la dispersione dei dati tramite il parametro ϕ_g:

$$X_{gj} \sim \mathcal{NB}(s_j p_{g\rho(j)}, \phi_g)$$

in cui s_j rappresenta la library size del campione j, ϕ_g la dispersione del gene g e $p_{g\rho(j)}$ l'abbondanza relativa del gene g nel gruppo sperimentale $\rho(j)$. Parametri della distribuzione NB sono la media $\mu_{gj} = s_j p_{g\rho(j)}$ e la varianza $\mu_{gj}(1 + \mu_{gj}\phi_g)$. Evidentemente, quando la dispersione è nulla ($\phi_g = 0$), la distribuzione NB si riduce alla distribuzione di Poisson. Il parametro di interesse per l'analisi differenziale è l'abbondanza relativa p_{gj}. Quando si esegue il fit sui dati di un modello NB, è necessario stimare la dispersione ϕ_g; con edgeR la stima può avvenire secondo due diverse strategie: se l'abbondanza di ciascun gene varia fra i campioni in modo tale che la deviazione standard di ogni gene sia proporzionale alla sua media, è ragionevole supporre che ϕ_g sia *comune* a tutti i geni; in alternativa, è possibile considerare ϕ_g variabile fra geni (dispersione *genewise*, si veda il paragrafo 4.2.3 "Analisi differenziale con edgeR").

4.2.2 Normalizzazione

Il metodo implementato in edgeR considera esclusivamente i count DGE "grezzi", non trasformati attraverso procedure di normalizzazione/standardizzazione come quelle utilizzate per il calcolo degli RPKM[2] [1]. Esistono tuttavia situazioni in cui normalizzare i dati è fondamentale per garantire che geni con lo stesso livello di espressione in due diversi

[2] *Reads Per Kilobase per Million mapped reads*: indice del livello di espressione definito da Mortazavi *et al.* [1]. $RPKM = \frac{totaleread}{readmappate[milioni] * lunghezzagene[Kb]}$

campioni non vengano selezionati come differenzialmente espressi a causa di bias nei dati. L'esempio più comune di tali bias è dato dalla library size, ovvero dal numero di read totali sequenziate in una libreria. É intuitivo pensare, infatti, che lo stesso campione sequenziato con due diverse library size, generi count diversi per una stessa feature. Altri bias possono presentarsi nel caso in cui alcuni geni siano geni fortemente espressi in un campione ma non in un altro. Robinson e Oshlack [31], analizzando il dataset di Marioni *et al.* [4], hanno dimostrato che un insieme di geni ad alta espressione presenti nei campioni di RNA epatico ma non in quelli di RNA renale "consumano" gran parte delle profondità di sequenziamento disponile, facendo sì che i restati geni risultino sottocampionati. Robinson e Oshlack [31] hanno messo a punto un metodo per lo scaling dei count che tiene conto non solo della profondità di sequenziamento delle librerie, ma dei livelli di espressione dell'intero set di geni del campione. Questo metodo, detto TMM, è implementato nel pacchetto edgeR.

Siano β_{gj} il vero livello di espressione relativo (incognito), l_g la lunghezza del gene g e s_j la library size del campione *j*; allora il valore atteso dei count X_{gj} per il gene g nel campione *j* è dato da:

$$E[X_{gj}] = \frac{\beta_{gj} l_g}{T_j} s_j$$

con

$$T_j = \sum_{g=1}^{G} \beta_{gj} l_g$$

Si noti come la formulazione di $E[X_{gj}]$ coincide sostanzialmente con $E[X_f] = \mu_f = p_f s = l_f \beta_f s$ vista nel capitolo 3.1, "Modello di Poisson". T_j rappresenta l'RNA totale di un campione, dipende dalla composizione dell'RNA e non può quindi essere stimato direttamente, dato che non sono noti i livelli di espressione e le lunghezze effettive dei geni del campione. Tuttavia è possibile stimare la quantità relativa di RNA fra due campioni $f_j = T_j/T_{j'}$, che può essere pensata come ad una sorta di fold change globale. Gli autori propongono di stimare questo rapporto di produzione dell'RNA usando la media TMM pesata degli "M value" o "log-fold-change" per ogni gene definiti come:

$$M_g = \log_2 \frac{X_{gj}/s_j}{X_{gj'}/s_{j'}}$$

e i valori di espressione:

$$A_g = \frac{1}{2} \log_2 (X_{gj}/s_j * X_{gj'}/s_{j'})$$

per $X_{gj}, X_{gj'} \neq 0$

La media "trimmed" è la media calcolata sui dati dopo che sono stati rimossi i valori estremi. Di default, il metodo prevede di eliminare il 30% dei valori più alti/bassi degli M_{gj}^r (ovvero delle log-fold change per il gene *g* nel campione *j* confrontato con il campione di riferimento *r*) e successivamente il 5% dei valori più alti/bassi degli A_g. Dopo il trimming, viene calcolata la media pesata degli *G* valori M_g, in cui i pesi sono dati dall'inverso della varianza asintotica calcolata con il metodo "delta" [32]. Precisamente, il fattore di normalizzazione per il campoione *j* usando la *r* come campione di riferimento è dato da

$$\log_2(TMM_j^{(r)}) = \frac{\sum_{g \in G^*} w_{gj}^r M_{gj}^r}{\sum_{g \in G^*} w_{gj}^r}$$

dove

$$M_{gj}^r = \frac{\log_2(X_{gj}/s_j)}{\log_2(X_{gr}/s_r)} \quad \text{e} \quad w_{gj}^r = \frac{s_j - X_{gj}}{s_j X_{gj}} + \frac{s_r - X_{gr}}{s_r X_{gr}} \quad \text{con } X_{gj}, X_{gr} > 0$$

G^* rappresenta l'insieme di geni per cui possono essere calcolati i log-fold change, ovvero i geni per cui $X_{gj} \neq 0,\ X_{gr} \neq 0$. Per un confronto fra due campioni la normalizzazione richiede un unico fattore di scala f_j. Tale rapporto può infatti essere usato per regolare entrambe le library size, dividendo i count del campione di riferimento per $\sqrt{f_j}$ e moltiplicando per $\sqrt{f_j}$ gli altri. La normalizzazione su più campioni può essere applicata selezionando un campione come reference e calcolando f_j per ogni altro campione che non sia il riferimento.

La funzione *calcNormFactors()* di edgeR utilizza il metodo TMM scegliendo come riferimento il campione con 75-esimo percentile più vicino alla media dei 75-esimi percentili. In alternativa, il riferimento può essere impostato dall'utente (parametro *refColumn)*.

4.2.3 Analisi differenziale con edgeR

Prima di procedere all'analisi differenziale, è necessario stimare la dispersione ϕ_g del modello negativo binomiale. È noto in letteratura che la stima della dispersione tramite metodi a maximum likelihood (MLE, *Maximum Likelihood Estimator*) tende ad essere poco robusta, a causa del fatto che media e varianza vengono stimate dagli stessi dati. Cox e Reid [33] hanno proposto un metodo (CR) in grado di ridurre il bias introdotto dagli stimatori MLE. Nel caso di librerie con la stessa dimensione $s_j \equiv s$, la dispersione può essere stimata in forma chiusa tramite metodo *conditional maximum likelihood* (CML) [34]. Dato che però nei dati RNA-seq le library size sono generalmente diverse, in edgeR è stato implementato un metodo proposto da Robinson e Smyth [29] che permette di creare degli "pseudo-dati" in cui le library size sono corrette considerando i quantili dei dati originali e ai quali si può quindi applicare la strategia CML (*quantile-adjusted* CML, qCML).

Si considerino ad esempio i count per un dato gene g in J campioni, con distribuzioni $X_{g1}, \ldots, X_{gJ}$ modellabili come variabili indipendenti $\mathcal{NB}(s_j p_{gj}, \phi_g)$; la stima della dispersione ϕ_g viene determinata come segue. Per semplicità di notazione, viene omesso dalle formule seguenti il pedice 'g', ma tutte le grandezze indicate fanno riferimento implicitamente al singolo gene g definito sopra.

Sia $s^* = \sqrt[m]{\prod_{j=1}^{J} s_j}$ la media geometrica delle diverse library size. L'algoritmo qCML "corregge" i dati osservati rendendoli campioni provenienti da una stessa distribuzione $NB(s^* p, \phi)$, secondo questi passi :

1. inizializza ϕ, tramite la stima CML standard
2. stima p dal valore attuale di ϕ
3. assumendo ogni osservazione x_j campionata da una distribuzione $NB(s_j p, \phi)$, calcola i quantili
$$q_j = P(X < x_j; s_j p, \phi) + \frac{1}{2} P(X = x_j; s_j p, \phi)$$
4. tramite un'interpolazione lineare della funzione quantile, calcola gli pseudodati da una distribuzione teorica $\mathcal{NB}(s^* p, \phi)$ caratterizzata dai quantili q_j. Gli pseudodati (per ogni gene) sono approssimativamente identicamente distribuiti
5. aggiorna ϕ calcolandola tramite CML sugli pseudodati
6. ripete i passi dal 2 al 5 fino alla convergenza di ϕ

L'aggiustamento quantile interviene aumentando i count osservati nel caso di library size inferiore alla media geometrica s^* e diminuendo i count in caso di library size maggiore di s^*.

Pensare di poter stimare correttamente la dispersione genewise è poco realistico, poiché un singolo gene non offre sufficienti dati per ottenere una stima robusta, soprattutto in dataset RNA-seq, caratterizzati da un limitato numero di repliche. edgeR utilizza quindi una strategia empirica Bayesiana, simile a quella dei modelli gerarchici utilizzati con i dati di microarray [35], per far convergere le dispersioni calcolate per i singoli geni Φ_g verso una dispersione comune Φ. Precisamente, per ogni gene viene calcolata una likelihood pesata (WL, *weighted*

conditional likelihood) per Φ_g, data dalla combinazione della likelihood genewise L_g e di quella comune L_c:

$$WL(\Phi) = L_g(\Phi_g) + \alpha L_c(\Phi_g)$$

α è il peso assegnato alla likelihood comune e corrisponde al prior della strategia Bayesiana. Nel caso di $\alpha = 0$, ci si riconduce alla stima qCML genewise. All'opposto, con α sufficientemente grande il contributo della likelihood individuale risulta trascurabile rispetto alla likelihood comune e ci si riconduce alla condizione di dispersione comune. La stima finale della dispersione sarà quindi tanto più "smooth", quanto più saranno basse le precisioni delle stime genewise e quanto più sarà grande α.

4.3 DESeq

DESeq è un pacchetto R che implementa il metodo di Anders e Huber [36] per l'analisi e la quantificazione dell'espressione differenziale in dati RNA-Seq. DESeq, il cui acronimo significa "Differential Expression analysis for Sequence count data", è applicabile alle metodologie di quantificazione basate su sequenziamento che forniscono dati "digitali", in cui l'abbondanza delle feature viene espressa tramite count, come RNA-Seq.

Nel pacchetto edgeR [24], Robinson e Smyth utilizzano un modello in cui media e varianza sono collegate dalla relazione $\sigma^2 = \mu + \alpha\mu^2$ con coefficiente di proporzionalità α mantenuto costante su tutto l'esperimento e stimato dai dati RNA-Seq. Per ciascun gene è dunque sufficiente la stima di un solo parametro, permettendo l'analisi anche di dati con un basso numero di repliche. DESeq [36] estende tale modello rendendolo più generale, con relazione fra media e varianza più flessibile e *data-driven*, permettendo un'analisi differenziale più bilanciata e robusta.

Il pacchetto è sviluppato per essere applicato a dataset contenenti repliche biologiche, da cui stimare la varianza. Tuttavia, gli autori forniscono delle indicazioni [37] e metodi che lo rendono applicabile, seppur con alcune limitazioni, anche ai casi in cui non siano disponibili repliche biologiche.

4.3.1 Modello

I count del gene g nel campione j appartenente alla condizione sperimentale $\rho(j)$ possono essere modellizzati con una distribuzione binomiale negativa $X_{gj} \sim \mathcal{NB}(\mu_{gj}, \sigma^2_{gj})$ avente media μ_{gi} e varianza σ^2_{gi}. I due parametri sono incogniti e vanno di conseguenza stimati dai dati. Generalmente, il numero di repliche a disposizione è limitato, rendendo necessarie delle assunzioni sul modello dei dati:

- la media μ_{gj} è il prodotto della library size s_j e di un fattore $p_{g\rho(j)}$, proporzionale al valore atteso della reale abbondanza (incognita) del gene g nella condizione sperimentale $\rho(j)$
  $$\mu_{gj} = s_j p_{g\rho(j)}$$
  Gli autori introducono inoltre il termine *common scale*, che indica una generica grandezza che è stata normalizzata per la library size s_j
- la varianza σ^2_{gj} è la somma di termine due contributi: *shot noise* μ_{gj} e *raw variance* $s_j^2 v_{g\rho(j)}$:
  $$\sigma^2_{gj} = \mu_{gj} + s_j^2 v_{g\rho(j)}$$
- la *raw variance* per gene $v_{g\rho(j)}$ è funzione smooth di $p_{g\rho(j)}$:
  $$v_{g\rho(j)} = v_\rho(p_{g\rho(j)})$$

Si distinguono quindi due componenti della varianza, una intrinseca ai dati ed espressione della variabilità biologica (*raw variance*) e una dovuta all'incertezza introdotta dalla stima

dell'abbondanza a partire da dati di conteggio (*shot noise*). La componente *shot noise* corrisponde quindi alla varianza prevista dal modello di Poisson (varianza coincidente alla media). Gli autori hanno evidenziato come con geni debolmente espressi (count molto bassi), il termine *shot noise* diventi dominante, mentre già con livelli di espressione moderati (e ancor più con geni fortememente espressi) l'effetto della variabilità biologica tra campioni rende dominante il termine *raw variance*.

L'assunzione di *raw variance* come funzione smooth di $p_{g\rho(j)}$ viene introdotta a conseguenza del fatto che generalmente il numero di repliche a disposizione è troppo basso per permettere una stima accurata della varianza di un gene utilizzando i soli dati a disposizione per quel gene. L'assunzione posta permette quindi di considerare i dati riferiti a più geni con simile livello di espressione, aumentando i dati a disposizione e migliorando la stima della varianza. Tale assunzione permette, anche se con alcune limitazioni, di estendere l'uso del pacchetto a dataset privi di repliche biologiche o in cui le repliche biologiche siano disponibili solo per alcuni campioni.

4.3.2 Fit dei dati

Il modello è identificato da tre parametri:

- J library size s_j: i valori attesi di tutti i count del campione j sono proporzionali a s_j
- per ogni condizione sperimentale $\rho(j)$, G parametri di livello di espressione $p_{g\rho(j)}$, che rappresentano l'abbondanza attesa del gene g nella condizione $\rho(j)$
- la funzione smooth $v_\rho: \mathbb{R}^+ \to \mathbb{R}^+$ per ciascuna condizione ρ, che modella la dipendenza della *raw variance* $v_{g\rho(j)}$ dalla media attesa $p_{g\rho(j)}$

L'uso del fattore s_j permette di rendere confrontabili i count di campioni diversi, ottenuti con diverse library size. La stima di s_j è ottenuta come:

$$\hat{s}_j = mediana_g \frac{X_{gj}}{\sqrt[J]{\prod_{i=1}^{J} X_{gi}}}$$

In questa formulazione, il denominatore $\sqrt[J]{\prod_{i=1}^{J} X_{gi}}$ può essere inteso come uno pseudo-count di riferimento per il gene g, ottenuto come media geometrica dei count sugli J campioni a disposizione per il gene g. Si ottengono così, per il j-esimo campione considerato, G rapporti tra i count di ciascun gene e lo "pseudo-count" di riferimento. s_j viene quindi calcolata come valore mediano dei G rapporti appena definiti.

La stima di $p_{g\rho(j)}$ avviene come media dei count dei campioni corrispondenti alla condizione sperimentale $\rho(j)$, normalizzati a s_j (common scale):

$$\hat{p}_{g\rho(j)} = \frac{1}{J_\rho} \sum_{j \in \rho(j)} \frac{X_{gj}}{\hat{s}_j}$$

in cui J_ρ è numero dei campioni che appartengono alla condizione ρ considerata. Per la stima di v_ρ viene calcolata prima la varianza su *common scale*:

$$w_{g\rho} = \frac{1}{J_\rho - 1} \sum_{j \in \rho(j)} \left(\frac{X_{gj}}{\hat{s}_j} - \hat{p}_{g\rho(j)} \right)^2$$

e si definisce

$$z_{g\rho} = \frac{\hat{p}_{g\rho(j)}}{J_\rho} \sum_{j \in \rho(j)} \frac{1}{\hat{s}_j}$$

È possibile dimostrare [36] che $w_{g\rho} - z_{g\rho}$ è uno stimatore privo di bias di $v_{g\rho(j)}$. Tuttavia, con un numero di repliche J_ρ limitato, come è tipico nelle applicazioni RNA-Seq, $w_{g\rho}$ è fortemente variabile e lo stimatore $w_{g\rho} - z_{g\rho}$ potrebbe non rappresentare una stima robusta della varianza. Per questo motivo, gli autori propongono l'uso di una regressione locale [38] sul grafico $(\hat{p}_{g\rho}, w_{g\rho})$ per ottenere una funzione smooth $w_\rho(p)$ con

$$\hat{v}_\rho(\hat{p}_{g\rho(j)}) = w_\rho(\hat{p}_{g\rho(j)}) - z_{g\rho}$$

come stima della *raw variance*.

4.3.3 Test di analisi differenziale

Supponiamo J_A repliche per la condizione A e J_B repliche per la condizione B. Per ciascun gene g, si vuole testare l'ipotesi nulla $p_{gA} = p_{gB}$, cioè la condizione in cui non vi sia differenza fra i livelli di espressione delle due condizioni.

Vengono definite le statistiche date dai count totali in ciascuna condizione:

$$X_{gA} = \sum_{j:\rho(j)=A} X_{gj} \qquad X_{gB} = \sum_{j:\rho(j)=B} X_{gj}$$

e la loro somma $X_{gS} = X_{gA} + X_{gB}$.

Viene calcolata la probabilità di avere $X_{gA} = a$ e $X_{gB} = b$ per ogni coppia (a, b). Tale probabilità viene denotata con $P(a, b)$. Il p-value di una coppia di somme dei count osservate (x_{gA}, x_{gB}) è ottenuto come somma di tutte le probabilità minori o uguali a $P(x_{gA}, x_{gB})$:

$$P_j = \frac{\sum_{\substack{a+b=x_{gS} \\ P(a,b)\leq P(x_{gA},x_{gB})}} P(a,b)}{\sum_{a+b=x_{gS}} P(a,b)}$$

Le variabili a e b assumono tutti i valori $0, \ldots, x_{gS}$.

4.3.4 Calcolo di P(a,b)

Assumendo che, sotto l'ipotesi nulla, i count di diversi campioni siano indipendenti. Allora

$$P(a, b) = P(X_{gA} = a)P(X_{gB} = b)$$

Il problema si riconduce al calcolo della probabilità degli eventi $X_{gA} = a$ e $X_{gB} = b$. La variabile aleatoria X_{gA} è somma di J_A variabili aleatorie NB; la si può rappresentare con una distribuzione NB di parametri ottenuti da quelli di $X_{gj} \sim \mathcal{NB}(\mu_{gj}, \sigma^2_{gj})$.

Viene calcolata la media *pooled* dai count in entrambe le condizioni

$$\hat{p}_{g0} = \sum_{j:\rho(j)\in\{A,B\}} X_{gj}/s_j$$

che considera in ipotesi nulla $p_{gA} = p_{gB}$.

Media e varianza totali per la condizione A sono ottenute come:

$$\hat{\mu}_{gA} = \sum_{j:\rho(j)\in A} s_j \hat{p}_{g0}$$

$$\sigma^2_{gA} = \sum_{j:\rho(j)=A} \hat{s}_j \hat{p}_{g0} + \hat{s}_j^2 \hat{v}_A(\hat{p}_{g0})$$

In modo del tutto analogo, media e varianza per la condizione B si ottengono come:

$$\hat{\mu}_{gB} = \sum_{j:\rho(j)\in B} s_j \hat{p}_{g0}$$

$$\sigma^2_{gB} = \sum_{j:\rho(j)=B} \hat{s}_j \hat{p}_{g0} + \hat{s}_j^2 \hat{v}_B(\hat{p}_{g0})$$

4.4 PoissonSeq

PoissonSeq è un pacchetto R che implementa il metodo di Li *et al.* [39] per la normalizzazione e l'analisi differenziale di dati RNA-Seq. Il pacchetto è applicabile a dati riferiti a due classi (ad esempio *sano* contro *malato*), ma anche a dati riferiti a più classi (ad esempio *malato tipo 1* contro *malato tipo 2* contro *malato tipo 3*) e a variabili quantitative (ad esempio pazienti con diverse pressioni sanguigne).

4.4.1 Modello

I count del gene g nel campione j sono modellizzati con una variabile di Poisson $X_{gj} \sim Poisson(\mu_{gj})$, con μ_{gj} esprimibile da un modello log-lineare che prende in considerazione se l'outcome è di tipo qualitativo o può essere rappresentato da $N \geq 2$ classi. Di seguito viene riportata la trattazione per il solo caso "qualitativo", a J classi.

Definiamo $X_{g\cdot} = \sum_{j=1}^{J} X_{gj}$, $X_{\cdot j} = \sum_{g=1}^{G} X_{gj}$, $X_{\cdot\cdot} = \sum_{j=1}^{J} \sum_{g=1}^{G} X_{gj}$. Siano inoltre β_g il livello di espressione (incognito) del gene g e s_j la library size del campione j che appartiene alla condizione sperimentale $\rho(j)$. La formulazione sviluppata dagli autori prevede di esprimere l'unico parametro della distribuzione di Poisson μ_{gj} tramite

$$\log \mu_{gj} = \log s_j + \log \beta_g + \sum_{\rho=1}^{N} \gamma_{g\rho} I_{(i\in\rho)}$$

in cui i coefficienti $\gamma_{g\rho}$ indicano se l'espressione del gene g è associata alla classe ρ. In partaciolare, se $\gamma_{g\rho} = 0$ l'espressione del gene g non è associata alla classe sperimentale ρ, il gene è invece associato alla classe ρ se $\gamma_{g\rho} \neq 0$.

4.4.2 Fit del modello

Il fit del modello sui dati avviene tramite una procedura a due passi:

1. fit del modello sotto l'ipotesi nulla per cui nessun gene sia associato ad una classe (quindi $\gamma_{g\rho} = 0$):
   $$\log \mu_{gj} = \log s_j + \log \beta_g$$
   Il fit di tale modello viene indicato con $\hat{\mu}_{gj} = X_{gj}^{(0)}$
2. fit del modello con un termine aggiuntivo per modellare l'espressione differenziale:
   $$\log \mu_{gj} = \log X_{gj}^{(0)} + \sum_{\rho=1}^{N} \gamma_{g\rho} I_{(i\in\rho)}$$

4.4.3 Normalizzazione

Al passo 1, il parametro β_g viene stimato tramite la maximum likelihood $\hat{\beta}_g = X_{g\cdot}$, ottenendo il livello di espressione β_g come somma dei count su tutti i campioni relativi al gene g. La library size $\hat{s}_j$ viene stimata su un insieme G^* di geni non differenzialmente espressi, secondo la formulazione:

$$\hat{s}_j = \frac{\sum_{g \epsilon G^*} X_{gj}}{\sum_{g \epsilon G^*} X_{g\cdot}}$$

La stima di $\hat{s}_j$ per il j-esimo campione è dunque ottenuta come rapporto fra la somma dei count dei geni non differenzialmente espressi misurati in quel campione, rispetto alla somma dei count degli stessi geni misurati in tutti i campioni. Quando G^* coincide con l'insieme di tutti i geni, la stima $\hat{s}_j$ si riconduce a $\hat{s}_j = X_{\cdot j}/X_{\cdot\cdot}$ (normalizzazione *total-count*). Si descrive di seguito il metodo implementato in PoissonSeq per selezionare il set di geni G^*. Date le stime di $\hat{s}_j$ e $\hat{\beta}_g = X_{g\cdot}$, il valore atteso di X_{gj} è $\hat{s}_j X_{g\cdot}$ ed è dunque possibile definire la statistica *goodness-of-fit*

$$GOF_g = \sum_{j=1}^{J} \frac{\left(X_{gj} - \hat{s}_j X_{g\cdot}\right)^2}{\hat{s}_j X_{g\cdot}}$$

Vengono inclusi in G^* i soli geni aventi valore di GOF_g compreso nel quantile $(\epsilon, 1-\epsilon)$ di tutti i valori GOF_g, con ϵ costante compresa tra 0 e 0,5 (default 0,25). La scelta della costante ϵ influenza le performance dell'algoritmo: valori elevati escludono un maggior numero di geni, riducendo il bias ma aumentando la varianza. Utilizzando l'insieme G^* così definito, è possibile aggiornare la library size $\hat{s}_j$ determinata precedentemente e con questa aggiornare nuovamente il valore di GOF_g. La procedura aggiorna le stime iterativamente fino a raggiungere la convergenza. Gli autori hanno dimostrato che la stima converge rapidamente [39], generalmente in meno di cinque iterazioni.

Dalle stime $\hat{\beta}_g$ e $\hat{s}_j$, si ottiene $X_{gj}^{(0)} = \hat{s}_j \hat{\beta}_g$.

4.4.4 Test statistico per l'analisi differenziale

Al passo 2 viene considerato il termine aggiuntivo che esprime il grado con cui ogni gene è associato alle classi sperimentali. Dato che l'obiettivo dell'analisi dell'espressione genica differenziale non consiste nelle stime puntuali dei parametri $\gamma_{g\rho}$, ma piuttosto nel determinare se tali stimi siano non nulle, gli autori hanno deciso di utilizzare lo *score test*, che non richiede di stimare il valore del parametro testato.

Ottenuto $X_{gj}^{(0)}$ da $\log \mu_{gj} = \log X_{gj}^{(0)} + \sum_{\rho=1}^{N} \gamma_{g\rho} I_{(i \epsilon \rho)}$, la log lokelihood è definita come

$$L = \sum_{g=1}^{G} \sum_{j=1}^{J} \left[X_{gj} \left(\log X_{gj}^{(0)} + \sum_{\rho=1}^{N} \gamma_{g\rho} I_{(i \epsilon \rho)} \right) - X_{gj}^{(0)} e^{\sum_{\rho=1}^{N} \gamma_{g\rho} I_{(i \epsilon \rho)}} \right]$$

Lo score statistico S_g per il gene g è definito, sotto l'ipotesi nulla $\gamma_{g\rho} = 0$ come

$$S_g = \sum_{\rho=1}^{N} \frac{\left(\frac{\eth L}{\eth \gamma_{g\rho}} |_{\gamma_{g\rho}=0} \right)^2}{-\frac{\eth^2 L}{\eth {\gamma_{g\rho}}^2} |_{\gamma_{g\rho}=0}} = \sum_{\rho=1}^{N} \frac{\left[\sum_{j \epsilon \rho} \left(X_{gj} - X_{gj}^{(0)} \right) \right]^2}{\sum_{j \epsilon \rho} X_{gj}^{(0)}}$$

Il valore dello score statistico S_g, determinato per ciascun gene, viene quindi confrontato con un valore di soglia, come criterio di selezione per l'analisi differenziale.

4.4.5 Stima della False Discovery Rate

La stima della false discovery rate (FDR) [40] [41] [42] viene generalmente effettuata tramite i seguenti passaggi:

1. calcolo delle statistiche $S_1, S_2, \dots, S_n$ basate sui dati

2. permutazione B volte dell'assegnazione dei campioni alle classi sperimentali. Nella b-esima permutazione, si calcolano le statistiche $S_1^b, S_2^b, \dots, S_n^b$ riferite ai dati permutati
3. per un range di valori della statistica superiori ad una valore di soglia C, si determinano

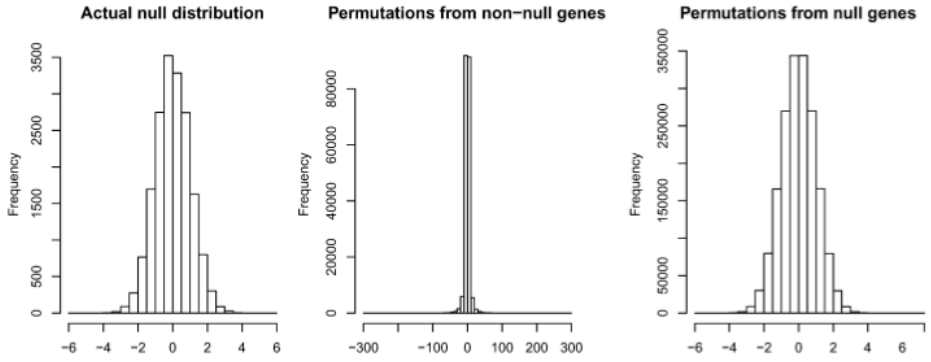
$$\hat{R} = \sum_{g=1}^{G} I_{(S_g > C)} \qquad \hat{V} = \frac{\hat{\pi}_0}{B} \sum_{g=1}^{G} \sum_{b=1}^{B} I_{(S_g^b > C)}$$

   in cui $\hat{R}$ e $\hat{V}$ rappresentano, rispettivamente, la stima del numero di geni con statistica che supera la soglia C (numero di geni selezionati come espressi) e del numero di geni con statistica della b-esima permutazione che supera la soglia ma che non sono realmente espressi (falsi positivi)
4. stima della FDR alla soglia C tramite il raporto $\widehat{FDR}_C = \hat{V}/\hat{R}$

Nel terzo passaggio, $\hat{\pi}_0$ è una stima di π_0, cioè della reale proporzione di geni non espressi nella popolazione. La stima rappresenta il rapporto del numero di geni con statistica inferiore ad una soglia rispetto al numero di geni complessivo. Nel dettaglio, la stima è ottenuta come

$$\hat{\pi}_0 = \sum_{g=1}^{n} I_{(S_g \le q_{2\zeta})} / \big(G(1-2\zeta)\big)$$

in cui la soglia $q_{2\zeta}$ è il quantile di ordine 2ζ della distribuzione dei valori permutati S_g^b, con ζ=0,25. Il numeratore del rapporto è dunque il numero di geni con statistica S_g inferiore al quantile 2ζ della distribuzione delle statistiche permutate S_g^b, mentre il denominatore rappresenta il numero di geni esclusi quelli compresi nel quantile 2ζ della distribuzione di S_g^b. La stima utilizza quindi i valori permutati di tutti i geni per stimare la distribuzione nulla. Questo approccio produce buone stime di FDR quando S_g viene calcolato a partire da dati gaussiani (cioè quando i count X_{gj} sono normalmente distribuiti). In caso di dati con distribuzione di Poisson gli score di geni espressi e non espressi hanno distribuzioni molto diverse. Per dimostrare questa affermazione gli autori hanno generato dei dati coerenti con il modello di Poisson e confrontato la distribuzione degli score di geni espressi e non espressi [39]. Confrontando le due distribuzioni con la distribuzione degli score in ipotesi nulla (Figura 9) si nota come la distribuzione degli score della permutazione dei geni non espressi sia molto simile alla reale distribuzione degli score dei geni non espressi, mentre la distribuzione degli score della permutazione dei geni espressi è molto più larga.



(a) (b) (c)

Figura 9. Istogramma delle statistiche per dati simulati. La distribuzione degli score dei geni espressi (b) è più larga della distribuzione degli score dei geni di geni non espressi (c), che è molto simile alla reale distribuzione degli score di geni non espressi (a)

Come conseguenza, la stima di FDR basata sulla distribuzione della permutazione di tutti i geni generalmente sovrastima la reale FDR. Gli autori hanno quindi scelto di considerare gli score dei soli geni non espressi, cioè i geni aventi score osservato basso. In dettaglio, vengono prima permutate B volte le assegnazioni dei campioni alle classi sperimentali, ottenendo le statistiche S_g^b. Quindi per stimare π_0, sia M l'indice dei geni aventi statistiche S_g che cadono

nei quantili $(\eta, 1-\eta)$ della distribuzione di S_g. M rappresenta cioè gli indici dei geni con statistica compresa fra i quantili η e $1-\eta$ della distribuzione di S_g. Dalla distribuzione delle statistiche S_g^b dei soli geni appartenenti a M si determina il quantile $\acute{q}_{2\zeta}$.

La stima di π_0 si ottiene in modo analogo a quanto visto in precedenza:

$$\hat{\pi}_0' = \sum_{g=1}^{G} I_{(S_g \leq q_{2\zeta}')} / (G(1-2\zeta))$$

con $\zeta = \eta = 0{,}25$. Infine, sia D l'indice dei geni aventi valori di S_g minori del quantile $\acute{\hat{\pi}}_0$ della distribuzione di S_g, in modo che D contenga i geni che sono realisticamente non espressi. Per determinare FDR alla soglia C, si calcolano allora

$$\hat{R} = \sum_{g=1}^{G} I_{(S_g > C)} \qquad \hat{V}' = \frac{1}{B} \sum_{g \in D} \sum_{b=1}^{B} I_{(S_g^b > C)}$$

in cui R stima come in precedenza il numero di geni con statistica che supera la soglia mentre la stima dei falsi positivi $\acute{\hat{V}}$ si riferisce al numero di geni appartenenti a D (realisticamente non espressi) aventi statistica della b-esima permutazione che supera la soglia. La stima della FDR alla soglia C è dunque determinata come

$$\widehat{FDR}_C = \hat{V}'/\hat{R}$$

5 Dati

5.1 Data set di Marioni *et al.*

Il data set di Marioni *et al.* [4] è stato generato a partire da due campioni di tessuto umano di fegato (*liver*) e rene (*kidney*), prelevati da un singolo maschio umano. I campioni sono stati sequenziati con Illumina Genome Analyzer (capitolo 2.4). Ogni campione è stato sequenziato in sette lane, suddivise in due run. Il design sperimentale consente di valutare nelle diverse repliche la variabilità dovuta alle lane (*lane effect*) e alle run del sequenziatore (*run effect*). Gli autori hanno valutato inoltre l'effetto della variazione della quantità dell'RNA di partenza, usando due diverse concentrazioni, 3 pM e 1,5 pM (indicata in Figura 10 con l'asterisco).

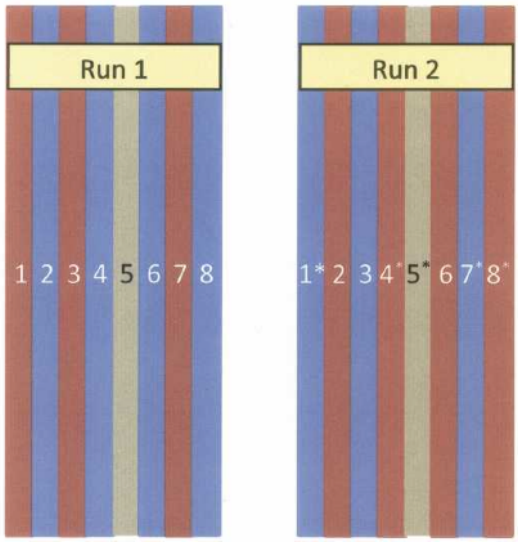


Figura 10. Design dell'esperimento di Marioni *et al*. Sono raffigurate le lane su cui i campioni di *kidney* (rosso) e di *liver* (azzurro) sono stati sequenziati nelle due run. La lane 5 è stata usata per il campione di controllo. Con asterisco sono indicate le repliche a concentrazione 1,5 pM. Immagine presa da [4].

Gli autori hanno reso disponibile, come materiale supplementare, la tabella dei count ottenuta dal mappaggio delle read RNA-seq campioni delle diverse repliche [43]. Il file *SupplementaryTable2.txt* contiene i count relativi a 32000 geni per i campioni di *kidney* e *liver*, su sette repliche tecniche ciascuno. Gli autori hanno dimostrato che i dati ottenuti da repliche a 3 pM presentano una variabilità molto ridotta, mentre l'uso di dati a 3pM mette in luce una maggior variabilità che rende i dati non completamente confrontabili [4]. Sono quindi stati mantenuti i dati relativi alle sole lane a 3 pM (riportate in Figura 11).

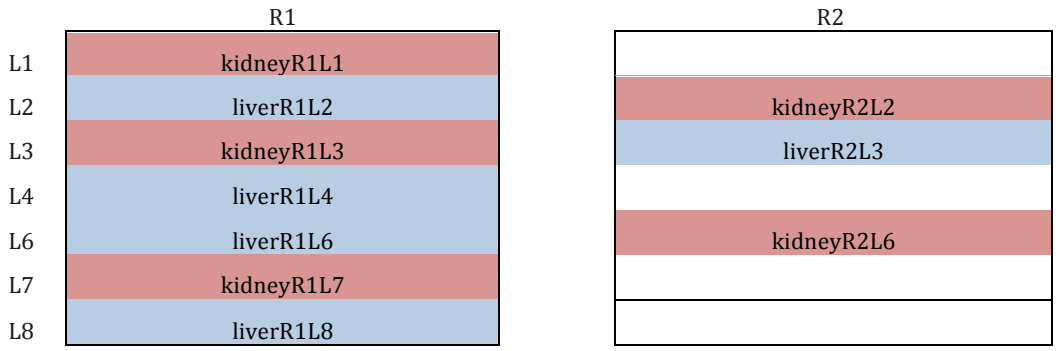


Figura 11. Le sole lane considerate nelle due run (R1 e R2), corrispondenti ai campioni di *kidney* e *liver* sequenziati a partire da una concentrazione di RNA in input di 3 pM.

Per permettere un confronto fra RNA-Seq e microarray, Marioni *et al.* hanno ibridato i campioni di RNA sequenziati con Illumina Genome Analyzer con le probe della piattaforma *Affymetrix Human Genome U133 Plus 2* [44]. Ogni campione di RNA è stato ibridato a tre array Affymetrix (repliche tecniche). I segnali delle probe sono stati normalizzati e corretti per il background di ciascun array, utilizzando l'algoritmo RMA [45]. Da questi dati grezzi, sono poi state considerate le sole probe set che risultavano mappate sui geni annotati nel database Ensembl [46]. Nel caso in cui più probe set mappassero sullo stesso gene, è stato considerato il valore di intensità della probe set più frequentemente rilevata come presente dal software Affymetrix sulle sei ibridazioni. In caso di probe egualmente presenti, è stato selezionato il valore di intensità di una fra queste, scelta casualmente. Per calcolare i p-value e identificare i geni differenzialmente espressi fra i due tessuti, gli autori hanno utilizzato un *t*-test empirico Bayesiano sviluppato da Smyth [35]; sono stati inoltre calcolati i q-value applicando la correzione per test multipli secondo l'approccio di Storey e Tibshirani [41]. Questi dati sono stati resi disponibili dagli autori, come materiale supplementare, nel file *SupplementaryTable3.txt* [47], che contiene, per ciascuno dei 32000 geni, i risultati dei test di analisi differenziale effettuati sui dati di RNA-seq e microarray. Queste le informazioni riportate per i microarray:

- media (sulle repliche tecniche) dei valori di intensità per i campioni di *kidney* e *liver*
- log-fold change fra *kidney* e *liver*
- p-value per l'analisi differenziale
- q-value corretto per test multipli

5.2 Data set di Cumbie *et al.*

Il data set di Cumbie *et al.* [48] deriva dal sequenziamento di RNA estratti da cellule di *Arabidopsis thaliana*, uno dei più studiati organismi modello per le scienze vegetali. Lo studio considerato si pone l'obiettivo di investigare la risposta di *A. thaliana* all'infezione batterica e i dati sequenziati derivano quindi da un esperimento "controllo-trattato". Nel gruppo *trattato*, esemplari indipendenti di *A. thaliana* sono stati infiltrati con il patogeno *Pseudomonas syringae* pathovar *tomato* DC3000. Per entrambi i gruppi è stato estratto l'RNA da tre diverse piante (repliche biologiche). Ciascuna replica biologica è stata sequenziata su una lane della piattaforma Illumina 1G Genome Analyzer (capitolo 2.4). Il design sperimentale consente di valutare la variabilità biologica nella risposta cellulare ai due trattamenti (seppur non separatamente dalla variabilità tecnica) nonché l'espressione differenziale. Questo data set è stato utilizzato in uno studio successivo di Di *et al.* [49] e la matrice dei count è stata resa disponibile dagli autori come materiale supplementare. Il file *arabidopsis.csv* contiene i count misurati per 26222 geni sulle tre lane di controllo e sulle tre lane di infezione batteria.

5.3 Data set MAQC

Il data set MAQC-2 (*MicroArray Quality Control*) è stato ottenuto da Bullard *et al.* [50] dal sequenziamento di due diversi campioni biologici:

- *Ambion Human Brain Reference RNA* (gruppo *Brain*): RNA estratto da cellule cerebrali umane di diverse aree del cervello di vari donatori
- *Stratagene Universal Human Reference RNA* (gruppo *UHR*): una miscela di RNA estratto da dieci diverse colture cellulari umane

I dati RNA-Seq sono stati ottenuti tramite Illumina 1G Genome Analyzer (capitolo 2.4). I campioni di *Brain* e *UHR* sono stati preparati secondo lo stesso protocollo e sequenziati su sette lane in due run del sequenziatore. Il design sperimentale (Figura 12) permette di valutare la variabilità tecnica dovuta alle lane e alle diverse run della strumentazione.

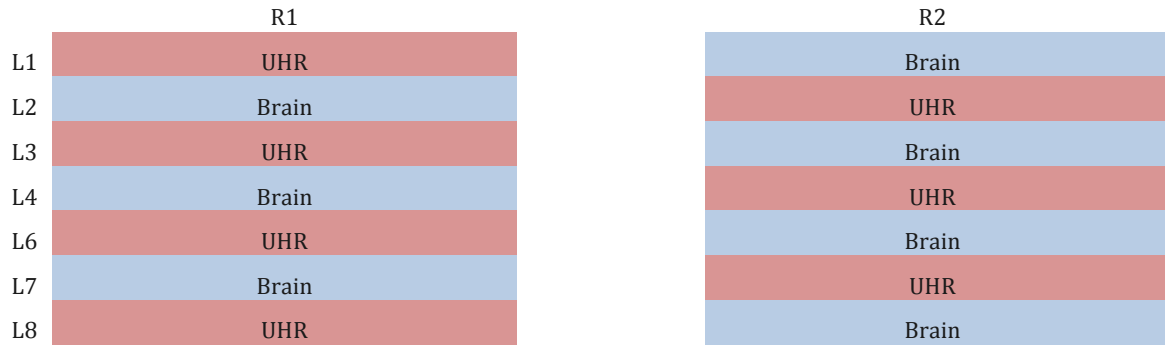


Figura 12. L'assegnazione dei campioni alle lane nei due cicli di sequenziamento (R1 e R2), corrispondenti ai campioni di *UHR* e *Brain*.

I dati grezzi (read) sono disponibili su SRA archive [51] accession number: SRA010153. Per permettere una valutazione della bontà dei dati RNA-Seq, gli stessi campioni sono inoltre stati quantificati tramite qRT-PCR (*quantitative Real-Time Polymerase Chain Reaction*). I dati di qRT-PCR sono disponibili su Gene Expression Omnibus [52] accession number: GSE5350.

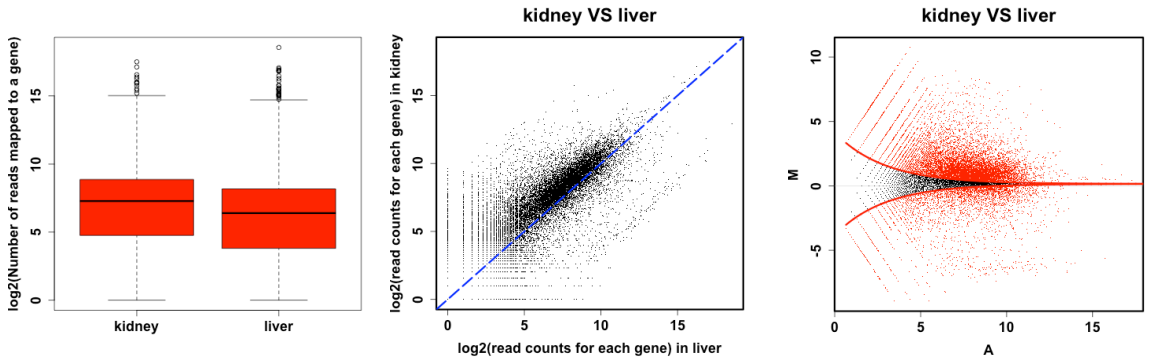
6 Analisi data set di Marioni *et al.*

Prima di elaborare i dati con i pacchetti R considerati, è stato applicato un filtraggio alla matrice dei count per eliminare i geni caratterizzati da un numero basso di count nelle diverse repliche, poichè presentavano elevato rumore. Precisamente, sono stati mantenuti i soli geni aventi media dei count (valutata su tutte 10 le repliche disponibili) pari ad almeno 0,5. Applicando tale correzione si passa dai 32000 geni originali a 18667 geni.

6.1 DEGSeq

6.1.1 Metodo MARS

Con l'approccio MARS, DEGSeq produce in output boxplot e scatterplot di confronto fra le due condizioni sperimentali testate oltre a un MA-plot riassuntivo dell'analisi differenziale.



(a) (b) (c)

Figura 13. L'output grafico prodotto da DEGseq con il metodo MARS: (a) boxplot, (b) scatterplot e (c) MA-plot.

In Figura 13a è riportato il boxplot della distribuzione dei count per gene nelle repliche di *kidney* e di *liver*. Il boxplot è un grafico che permette di visualizzare il primo, secondo (la mediana, rappresentata dalla linea nera orizzontale di Figura 13a) e terzo quantile della distribuzione di un campione; gli spazi tra le componenti del box rappresentano la dispersione e l'asimmetria nei dati. Permette un confronto immediato fra le distribuzioni dei due campioni, senza formulare nessuna assunzione sulla loro distribuzione statistica. Dato che l'ordinata esprime il numero di read mappate, è possibile osservare come il campione di *kidney* presenti un maggior numero di geni ad alta espressione.

In Figura 13b è mostrato lo scatterplot di confronto fra *liver* e *kidney*. Il grafico descrive la relazione tra i count (su scala logaritmica) dei due campioni di ciascun gene in *liver* e in *kidney*. I geni non differenzialmente espressi si disporranno approssimativamente lungo la diagonale, avendo read count uguali (o molto simili) nei due campioni. É evidente come DEGSeq produca grafici in cui le distribuzioni dei due campioni non sono state normalizzate.

In Figura 13c è riportato l'MA-plot, che esprime la relazione tra il livello medio di espressione genica (A, asse x) e la variazione di espressione (M, asse y) osservata tra i due gruppi sperimentali. Precisamente, $M = \log_2(X_1) - \log_2(X_2)$ e $A = [\log_2(X_1) + \log_2(X_2)]/2$ in cui sono stati definiti con X_1 i count nei campioni di *kidney* e X_2 i count nei campioni di *liver*; i geni non differenzialmente espressi si disporranno approssimativamente lungo l'asse *M=0*. I punti evidenziati in rosso corrispondono ai geni che DEGseq ha selezionato come differenzialmente espressi. Sull'MA-plot sono riportate anche le due curve m_1 e m_2 (capitolo

4.1.1, "Metodo MA-plot based con modello a campionamento casuale (MARS)") che identificano la regione teorica avente quattro volte la deviazione standard di $M|A$.

Sull'MA-plot è ben evidente uno shift verso l'alto dei valori M, rispetto a $M = 0$, causato dalla presenza di un set di geni fortemente sovra-espressi in *liver*. La Figura 14 evidenzia la nuvola di punti caratterizzati da valori di M particolarmente bassi, che corrispondono alla condizione $X_1 \ll X_2$ identificando quindi tali geni a forte espressione in *liver*. Per lo stesso motivo, lo scatterplot (Figura 13b) presenta uno shift verso l'alto rispetto alla diagonale.

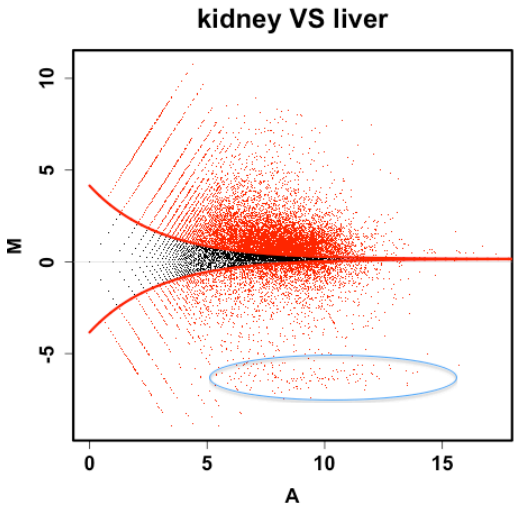


Figura 14. MA-plot dei count in *liver* e *kidney*. In azzurro è evidenziato il particolare set di geni upregolati in *liver* rispetto a *kidney* (M < -5).

Dalle liste dei geni e relativi p-value forniti dal metodo MARS, sono stati selezionati i geni differenzialmente espressi tra le due condizioni, impostando quattro diverse soglie sui p-value non corretti, come mostrato in Tabella 1.

p-value	N° geni DE
5%	13833
1%	12343
0,05%	10575
0,01%	9945

Tabella 1. Numero di geni selezionati come differenzialmente espressi dal metodo MARS, per i quattro valori di soglia di p-value (non corretto per test multipli) considerati.

6.1.2 Metodo MATR

Il metodo MATR considera le repliche tecniche di uno stesso campione biologico per ottenere una stima della varianza tra di esse. Segue poi l'analisi differenziale tra due campioni, usando la varianza stimata dalle repliche tecniche per la determinazione dello Z-score. Quindi, come per MARS, dallo Z-score viene calcolato un p-value da confrontare con il livello di significatività desiderato per la verifica del test di ipotesi. Per questa analisi è stata utilizzata una soglia sul p-value pari al 5%. Il metodo MATR di DEGseq permette di indicare in input due campioni da utilizzare come repliche tecniche per la stima della varianza e due campioni da utilizzare per l'analisi differenziale. Sono state definite cinque combinazioni di repliche e per ciascuna è rilevato il numero di geni selezionati come differenzialmente espressi (Tabella

2). Le combinazioni scelte permettono di evidenziare gli effetti della variabilità biologica (fra *kidney* e *liver*) e della variabilità misura tecnica tra le diverse lane (*lane effect*) e diverse run (*run effect*) dello strumento.

Campioni confrontati		**Fonti di variabilità misurate**		**N° geni DE**
Stima varianza	**Analisi differenziale**	**Stima della varianza**	**Analisi diffrerenziale**	
kidneyR1L3 vs kidneyR1L7	kidneyR1L1 vs liverR1L2	lane	biologica e lane	9461
kidneyR1L3 vs kidneyR1L7	kidneyR1L1 vs liverR2L3	lane	biologica, lane e run	9600
kidneyR1L3 vs kidneyR1L1	kidneyR2L2 vs liverR1L2	lane	biologica e run	9574
kidneyR1L1 vs kidneyR2L2	kidneyR1L3 vs liverR1L2	lane e run	biologica e lane	9448
kidneyR1L1 vs kidneyR2L2	kidneyR1L7 vs liverR2L3	lane e run	biologica, lane e run	9485

Tabella 2. Le cinque combinazioni di repliche testate con il metodo MATR. Per ciascuna prova sono indicate le possibili fonti di variabilità (biologica, lane, run) e il numero di geni selezionati come differenzialmente espressi con p-value≤5%.

I geni selezionati come differenzialmente espressi in tutte le combinazioni considerate sono 9448. Le figure Figura 15-Figura 19 riportano boxplot delle distribuzioni di *kidney* e *liver*, scatterplot fra le repliche e fra i campioni e MA-plot nelle cinque combinazioni testate (Tabella 2, "Analisi differenziale"). Nell'MA-plot, le linee blu rappresentano le curve a quattro volte la deviazione standard di M stimata dalla repliche, i punti rossi identificano i geni differenzialmente espressi.

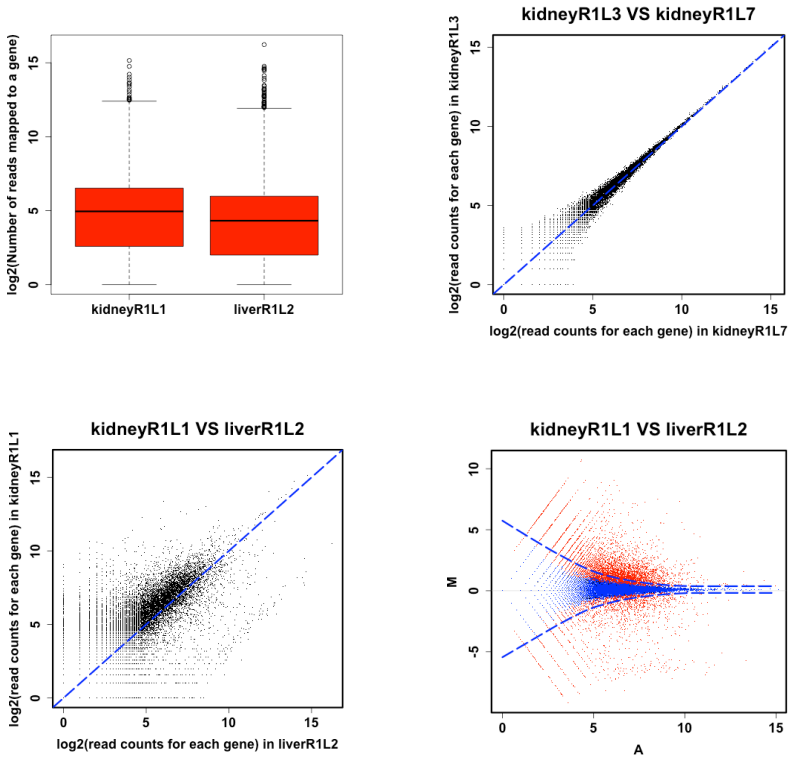


Figura 15. Output grafico prodotto con il metodo MATR di DEGseq per la combinazione 1 (varianza da kidneyR1L3 vs kidneyR1L7, analisi differenziale fra kidneyR1L1 vs liverR1L2): boxplot, scatterplot dei campioni di *kidney* e *liver* testati per analisi differenziale e MA-plot.

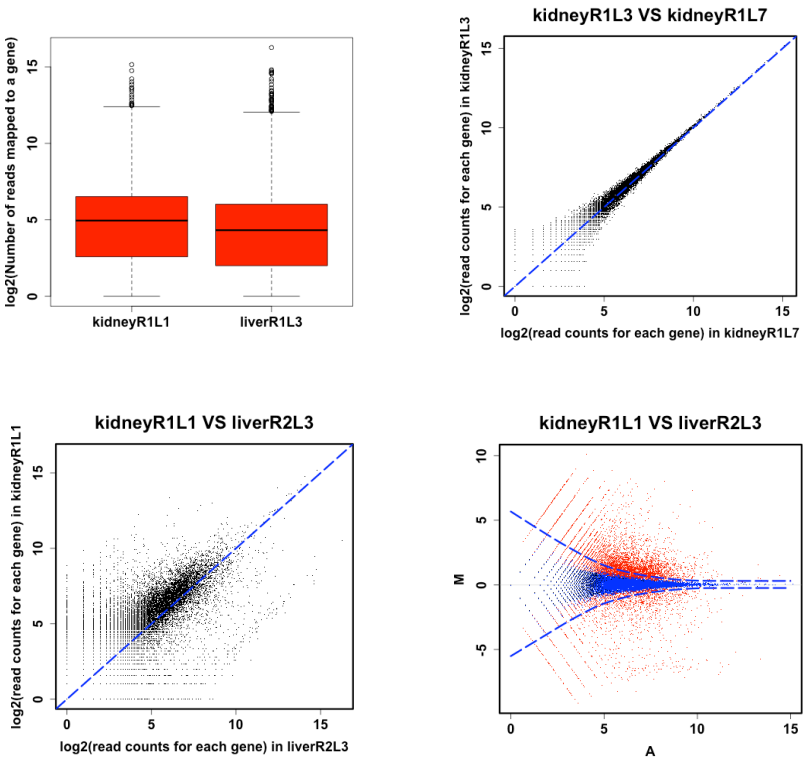


Figura 16. Output grafico prodotto con il metodo MATR di DEGseq per la combinazione 2 (varianza da kidneyR1L3 vs kidneyR1L7, analisi differenziale fra kidneyR1L1 vs liverR2L3): boxplot, scatterplot dei campioni di *kidney* e *liver* testati per analisi differenziale e MA-plot.

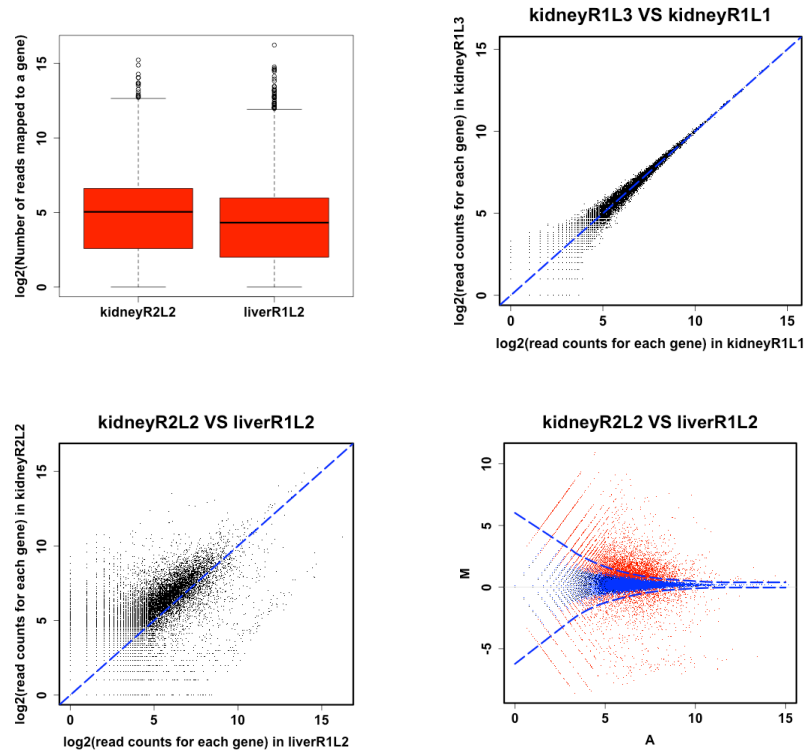


Figura 17. Output grafico prodotto con il metodo MATR di DEGseq per la combinazione 3 (varianza da kidneyR1L3 vs kidneyR1L1, analisi differenziale fra kidneyR2L2 vs liverR1L2): boxplot, scatterplot dei campioni di *kidney* e *liver* testati per analisi differenziale e MA-plot.

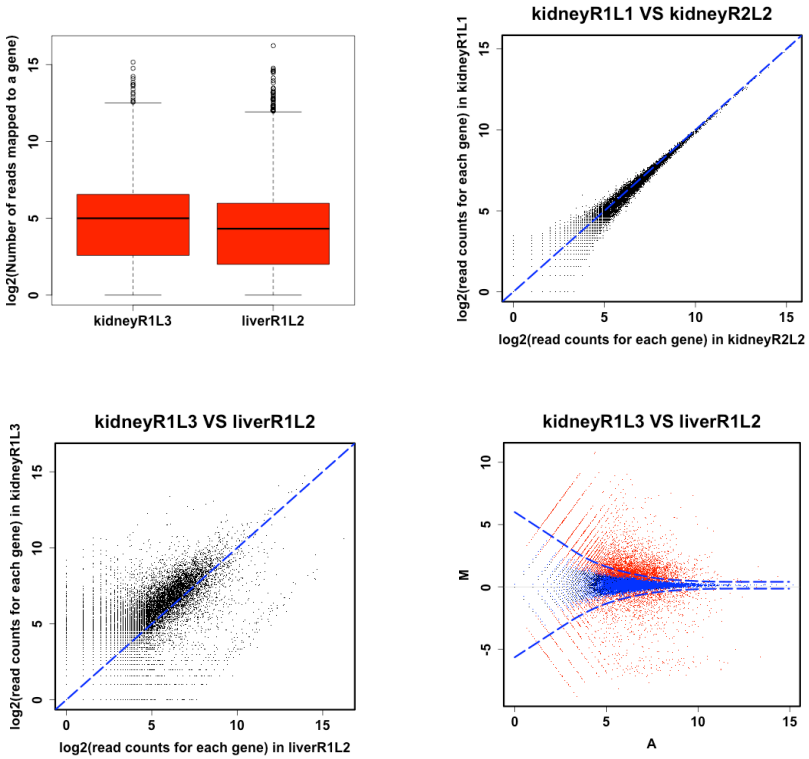


Figura 18. Output grafico prodotto con il metodo MATR di DEGseq per la combinazione 4 (varianza da kidneyR1L1 vs kidneyR2L2, analisi differenziale fra kidneyR1L3 vs liverR1L2): boxplot, scatterplot dei campioni di *kidney* e *liver* testati per analisi differenziale e MA-plot

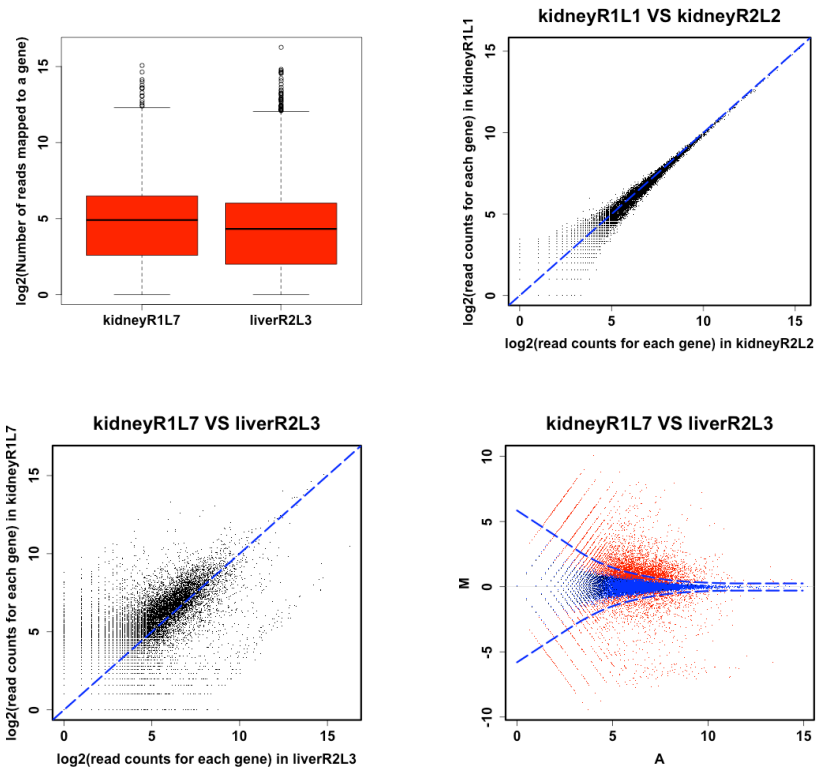


Figura 19. Output grafico prodotto con il metodo MATR di DEGseq per la combinazione 5 (varianza da kidneyR1L1 vs kidneyR2L2, analisi differenziale fra kidneyR1L7 vs liverR2L3): boxplot, scatterplot dei campioni di *kidney* e *liver* testati per analisi differenziale e MA-plot

6.1.3 Metodo CTR

Le repliche tecniche dei due campioni sono state valutate con il metodo CTR (capitolo 4.1.4), che permette di verificare se la variabilità tra queste sia descrivibile da un modello di campionamento uniforme. L'output che si ottiene è un MA-plot in cui sono mostrate le 4sd-curves teoriche $m_{1,2}^{MARS}$ (in rosso) e le 4sd-curves $m_{1,2}^{MATR}$ stimate dalle repliche (in blu). Se le due curve non hanno un andamento coerente fra loro, significa che le variazioni fra le due repliche tecniche non sono spiegabili con il modello adottato. In Figura 20 e Figura 21 sono riportati gli scatterplot (diagonale alta) e gli MA-plot (diagonale bassa) di tutte le possibili combinazioni di repliche di *kidney* e *liver*. Gli scatterplot presentano valori distribuiti lungo la diagonale, coerentemente con le attese di repliche tecniche di uno stesso campione. Dall'analisi degli MA-plot si nota come la stima della varianza dalle repliche tecniche (in blu) ben approssima la varianza "teorica" (in rosso) prevista dal modello di Jiang [22].

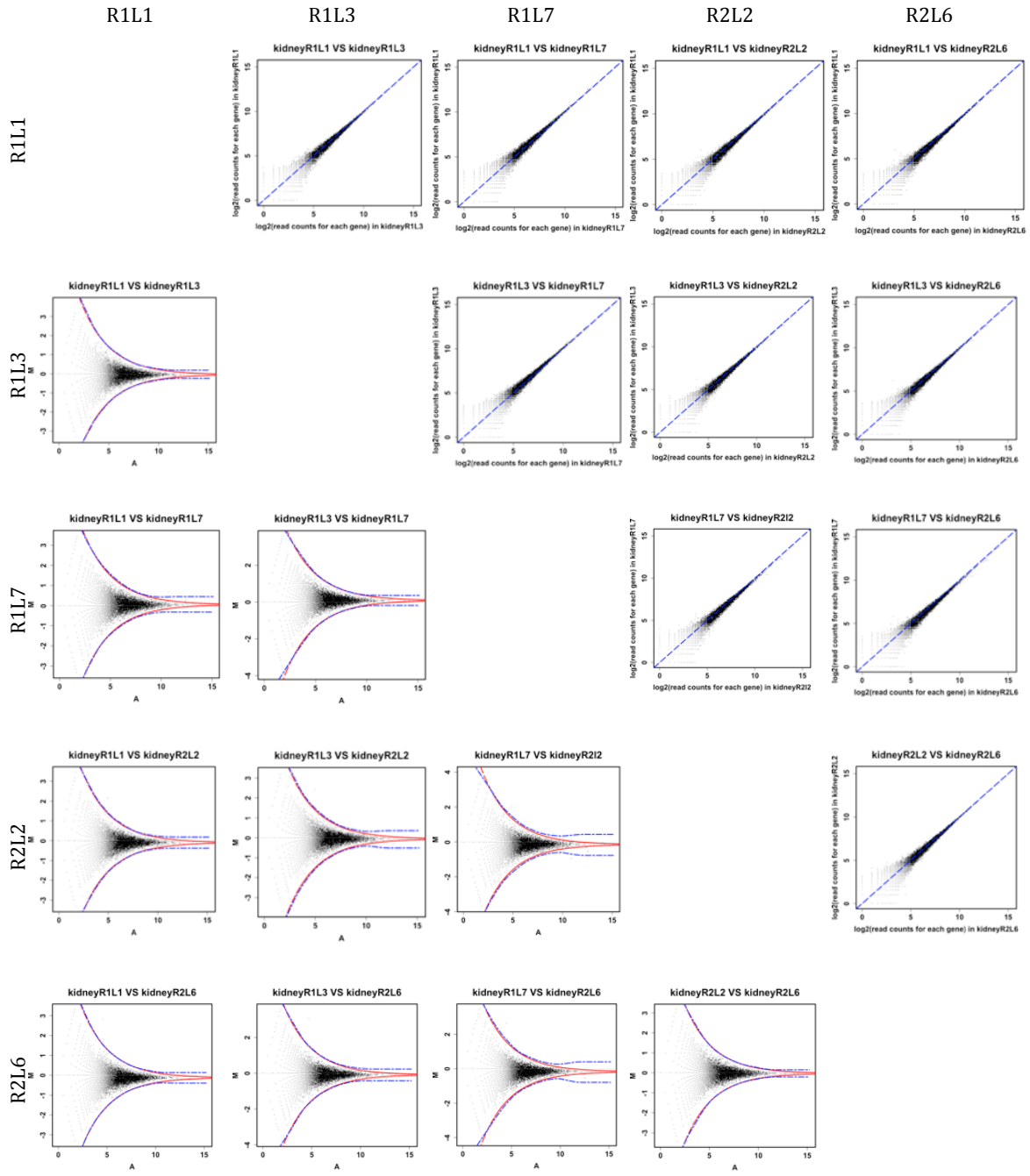


Figura 20. Grafici ottenuti da DEGseq con metodo CTR tra le diverse repliche tecniche di *kidney*

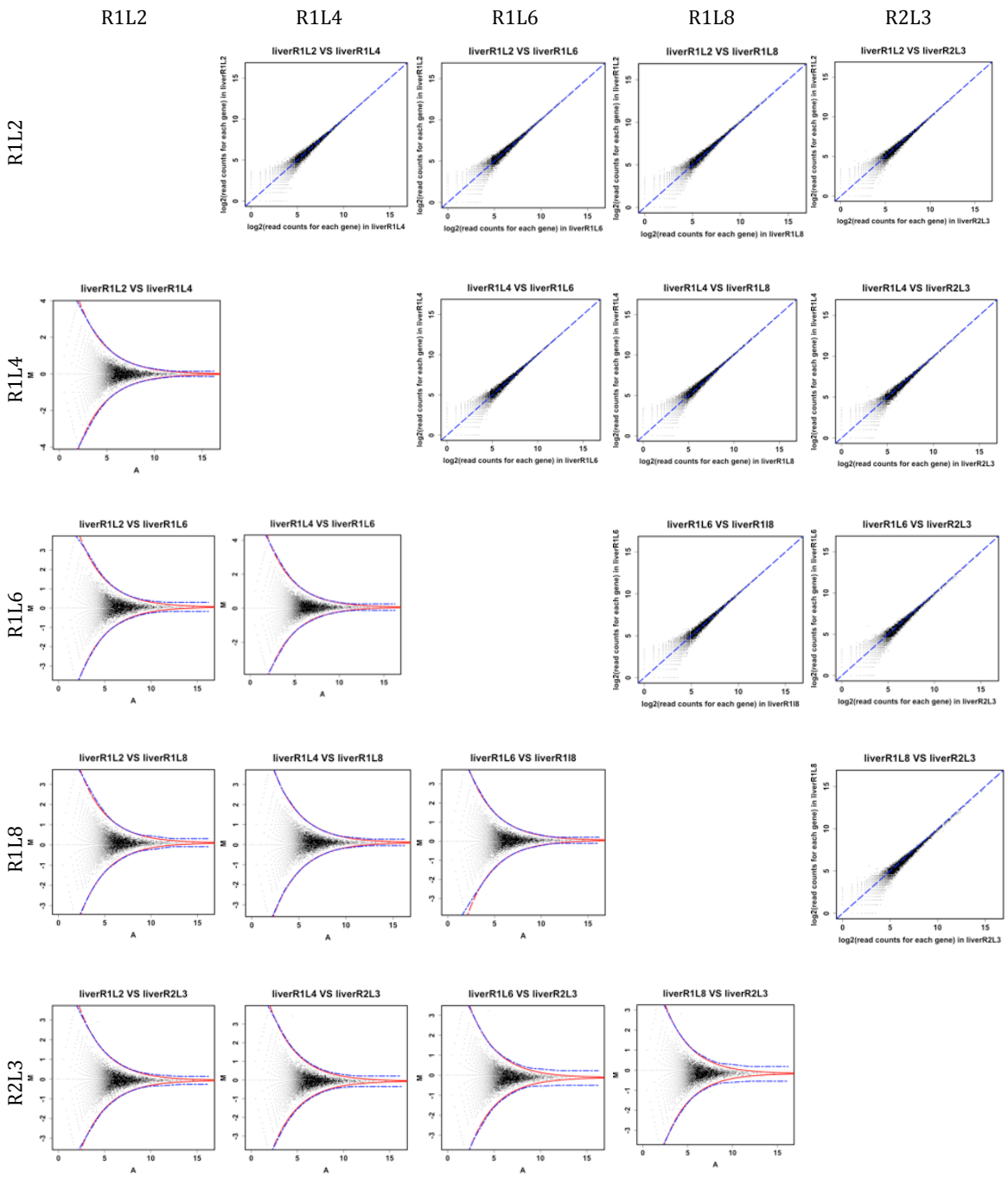


Figura 21. Output ottenuti da DEGseq con metodo CTR tra le diverse repliche tecniche dei campioni di *liver*. I grafici sopra la diagonale sono gli scatterplot fra le due repliche tecniche; sotto la diagonale sono riportati gli MA-plot.

6.2 edgeR

In edgeR, i dati di RNA-Seq vengono convertiti in un oggetto di tipo DGEList. L'oggetto DGEList viene creato dal metodo omonimo a partire dalla matrice dei count e da un vettore di *label* che definisce il design dell'esperimento (definisce cioè a quale gruppo sperimentale appartengono le repliche a disposizione). EdgeR opera direttamente sui dati originali di espressione digitale, senza alcuna pre-elaborazione da parte dell'utente: valori di count normalizzati o trasformati come gli RPKM [1] non sono ammessi.

Ottenuto l'oggetto DGEList, è possibile utilizzare la funzione *calcNormFactors* che calcola, per ciascun campione, un fattore di normalizzazione secondo l'approccio TMM [31]. *calcNormFactors* è in grado di determinare autonomamente il campione da usare come riferimento della normalizzazione. La Tabella 3 elenca, per il dataset di Marioni *et al.*, le diverse library, i gruppi sperimentali a cui appartengono (*liver* o *kidney*), le rispettive library size ed il fattore di normalizzazione determinato da edgeR.

Replica	Gruppo	Library size	Fattore di normalizzazione
R1L1Kidney	Kidney	1804154	1,2208377
R1L2Liver	Liver	1691040	0,8219681
R1L3Kidney	Kidney	1854340	1,2255060
R1L4Liver	Liver	1695637	0,8198550
R1L6Liver	Liver	1630144	0,8151633
R1L7Kidney	Kidney	1741562	1,2187193
R1L8Liver	Liver	1574759	0,8141885
R2L2Kidney	Kidney	1926629	1,2235139
R2L3Liver	Liver	1766529	0,8168980
R2L6Kidney	Kidney	1962514	1,2268305

Tabella 3. Gruppo sperimentale, library size e fattore di normalizzazione determinati da edgeR per ciascuna replica.

Una volta normalizzati i dati, può essere utile visualizzarli in un MDS-plot (*MultiDimensional Scaling plot*), utilizzando la funzione *plotMDS.* Si tratta di un grafico che permette di analizzare visivamente le similitudini o le differenze nei dati, rappresentando ogni osservazione in uno spazio multidimensionale. L'MDS-plot prodotto da edgeR rappresenta le osservazioni su uno spazio a due dimensioni: la dimensione *1* è la direzione che meglio separa i campioni, senza considerare se questi siano trattamenti o repliche; la dimensione *2* è la seconda migliore direzione di separazione dei campioni, scorrelata dalla precedente. Il significato che le due dimensioni assumono dipende dagli specifici dati che si stanno analizzando. Nel caso del dataset considerato, la dimensione *1* separa i due gruppi sperimentali a cui i campioni appartengono, mentre la dimensione *2* separa le diverse repliche tecniche. La distanza tra ogni coppia di campioni corrisponde alla radice quadrata della dispersione comune dei primi 500 geni che mostrano una più marcata distinzione fra i loro campioni. L'MDS-plot è quindi un utile strumento per evidenziare visivamente l'appartenenza ai due gruppi (*kidney* e *liver*) dei dati. La Figura 22 mostra l'MDS-plot ottenuto per i campioni del dataset di Marioni. Visivamente, si nota come vi sia netta separazione, nel senso della dimensione *1*, fra i campioni di *kidney* e quelli di *liver*. Questo indica che ci si può attendere un elevato numero di geni rilevati come differenzialmente espressi.

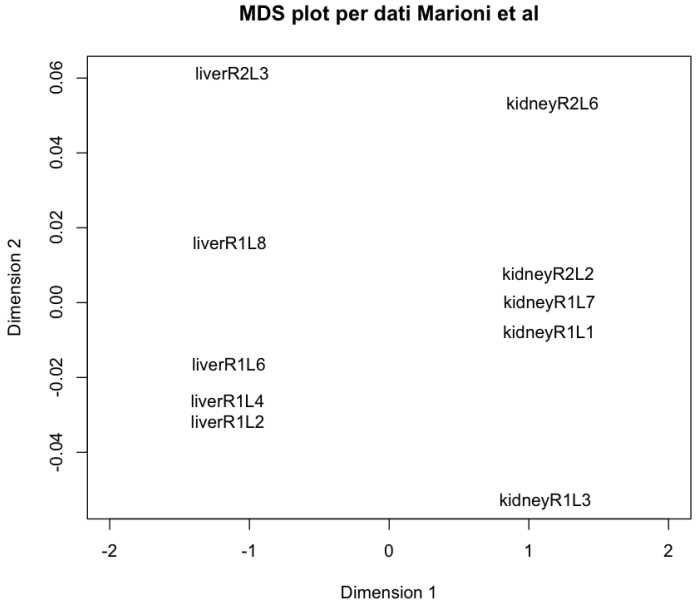


Figura 22. MDS-plot ottenuto con edgeR per le repliche del dataset di Marioni

La selezione dei geni differenzialemente espressi è stata effettuata a valle della fase di stima del parametro dispersione ϕ del modello NB, effettuata con due diversi approcci: stima della dispersione comune e stima della dispersione genewise.

6.2.1 Analisi differenziale con stima della dispersione comune

edgeR fornisce il metodo *estimateCommonDisp* per ottenere la stima della dispersione comune; all'oggetto DGE in input vengono aggiunte le seguenti variabili:

- *common.dispersion*: dispersione comune stimata
- *pseudo.alt*: pseudo-dati generati da edgeR, "correggendo" i count originali rendendoli campioni provenienti da una stessa distribuzione
- *genes*: identificativi dei geni
- *conc*: comprende, per ogni gene, la stima della concentrazione (livello di espressione), determinata come rapporto fra la somma dei count per gene rispetto alla library size comune. Fornisce la stima (*conc$conc.common*) sotto l'ipotesi nulla di assenza di differenze fra i due gruppi e la stima (*conc$conc.group*) di ogni gene all'interno di ogni gruppo
- *common.lib.size*: library size comune s^* a cui sono state riportate le library size originali

Applicato al dataset di Marioni, edgeR stima una dispersione comune pari a 0,0007. Il coefficiente di variazione della variabilità biologica (radice quadrata della dispersione) è CV=0,0264. Una volta stimata la dispersione comune, è possibile valutare la differenza di stima della varianza fra il modello NB adottato dagli autori e il modello di Poisson (adottato da DEGSeq). Indicando con μ la media dei count di un gene, il modello Poissoniano prevede che la varianza sia pari a μ, mentre nel modello NB la varianza vale $\mu(1+\mu\phi)$. Nel caso specifico del dataset di Marioni, considerando che un ipotetico gene abbia media dei count $\mu = 500$, la varianza del modello NB risulterà $1+\mu\phi = 1+500*0{,}0007 = 1{,}35$ volte superiore a quella prevista dal modello di Poisson. La Figura 23 illustra la relazione fra media e varianza nei dati e permette un confronto immediato fra l'andamento della varianza con modello di Poisson, in nero, e della varianza stimata con edgeR (dispersione comune, curva blu, e genewise, pallini azzurri). Entrambe le varianze stimate da edgeR, consistenti tra di loro, descrivono i dati più fedelmente rispetto al modello di Poisson (varianza uguale alla media)

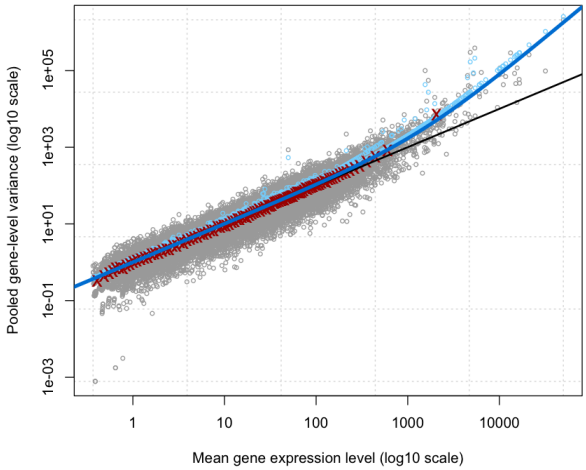


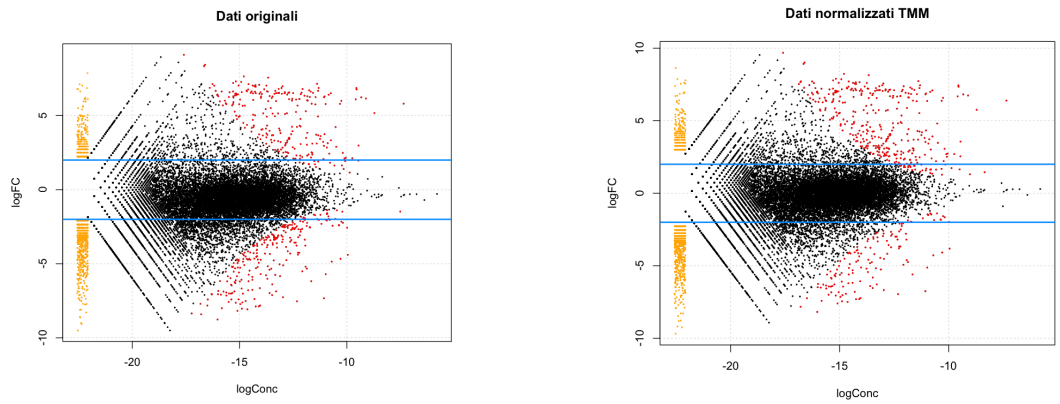
Figura 23. Rapporto media-varianza dei count per ogni gene: i punti grigi indicano la varianza grezza di ciascun gene rispetto alla sua abbondanza (asse x). Le croci rosse indicano la media della varianza grezza dei geni raggruppati in 100 bin sulla base della loro abbondanza. I punti in azzurro mostrano la varianza stimata per ciascun gene, calcolata dalla dispersione genewise. La linea blu rappresenta la varianza stimata utilizzando la dispersione comune. La linea nera rappresenta la varianza prevista dal modello di Poisson, uguale alla media.

Stimata la dispersione comune è possibile applicare il test esatto fornito da edgeR per l'analisi differenziale (funzione *exactTest*). La funzione fornisce in output una tabella contenente, per ogni gene, il p-value determinato dal test, la correzione false discovery rate (FDR) per test multipli, la log-concentrazione (la misura complessiva, in scala logaritmica, dell'abbondanza del gene) e la log-fold change (cambiamento del livello di espressione fra i due gruppi, calcolata dalla concentrazione del gene in ciascun gruppo). La Tabella 4 mostra i cinque geni per i quali edgeR ha determinato il minor p-value; il p-value risulta addirittura nullo, confermando l'espressione differenziale di tali geni tra i gruppi *liver* e *kidney*.

Gene	log-conc	logFC	p-value	FDR
ENSG00000084734	-17.59696	9.682855	0	0
ENSG00000140107	-16.61076	9.010170	0	0
ENSG00000138075	-16.61076	8.917861	0	0
ENSG00000105697	-14.80108	8.226321	0	0
ENSG00000197901	-16.02883	-8.181032	0	0

Tabella 4. Elenco dei primi cinque geni con minor p-value selezionati da edgeR (con stima della dispersione comune) e rispettive log-concentrazioni, log-fold change e false discovery rate.

edgeR consente di ottenere uno smear plot che rappresenta i risultati dell'analisi differenziale. Si tratta di un grafico in cui sono riportate log-concentrazione (asse x) e log-fold change (asse y), per ogni gene del dataset considerato. Lo smear plot prodotto da edgeR, tramite la funzione *plotSmear,* applica una "correzione grafica" che tiene conto delle condizioni in cui i count totali di un gruppo siano nulli, condizione caratterizzata da log-fold change infinita e log-concentrazione infinita negativa. Per ovviare a questo problema, la funzione *plotSmear* produce una "macchia" (*smear*) di punti all'estrema sinistra del grafico per i geni con count nulli in uno dei gruppi. In Figura 24 sono riportati tutti i geni considerati; in arancione sono indicati i punti *smear*, in rosso sono evidenziati i 500 geni a più alta espressione differenziale. Da Figura 24 è possibile notare come lo shift verso il basso rispetto allo 0 della log-fold change, presente sui dati originali, venga ridotto a seguito della normalizzazione TMM eseguita da edgeR.



(a) (b)

Figura 24. Smear plot di edgeR che confronta log-concentrazione e log-fold change per ciascun gene considerato. Valori di log-fold change positivi indicano geni sovra-espressi in *liver*. In (a) sono riportati i risultati ottenuti sui dati originali e in (b) quelli ottenuti sui dati normalizzati con il metodo TMM. In rosso sono evidenziati i 500 geni a più alto livello di espressione differenziale, mentre le linee blu rappresentano il limite della regione logFC ± 2.

Sia per i dati originali sia per i dati normalizzati, dalle liste dei p-value ottenuti sono stati selezionati i geni differenzialmente espressi tra le due condizioni, impostando quattro diverse soglie sui p-value non corretti (Tabella 5). La normalizzazione riduce il numero di geni selezionati come sovra-espressi in *liver* (*DE+*) e porta all'identificazione di un maggior numero di geni sovra-espressi in *kidney* (*DE-*), rispetto ai dati originali.

p-value	Dati originali			Dati normalizzati TMM		
	N° geni DE	N° geni DE+	N° geni DE-	N° geni DE	N° geni DE+	N° geni DE-
5%	13429	10340	3089	12271	6598	5673
1%	12017	9331	2686	10733	5802	4931
0,05%	10349	8098	2251	8935	4769	4166
0,01%	9718	7617	2101	8219	4362	3857

Tabella 5. Numero di geni selezionati come differenzialmente espressi (*DE*), sovra-espressi (*DE+*) e sotto-espressi (*DE-*) in *liver* da edgeR con stima della dispersione comune, per i quattro valori di soglia di p-value (non corretto per test multipli) considerati. Le tre classificazioni sono state valutate sia sui dati originali sia sui dati normalizzati con il metodo TMM.

6.2.2 Analisi con stima della dispersione genewise

La funzione *estimateTagwiseDisp* di edgeR permette di stimare la dispersione genewise e "moderare" i valori ottenuti per singolo gene verso un valore comune. Seppure sia disponibile un algoritmo sviluppato da Robinson e Smyth [35] per stimare il parametro di smoothing con approccio Bayesiano, gli autori suggeriscono di scegliere questo valore a priori (settabile tramite il parametro *prior.n*) seguendo una regola empirica, basata sul numero di gradi di libertà (il numero di gruppi sottratto al numero di campioni). In particolare, l'indicazione prevede di scegliere il *prior.n* in modo che il prodotto fra questo e i gradi di libertà sia un numero compreso tra 20 e 30. La funzione *getPriorN()* può determinare autonomamente tale

prior, ricevendo in input l'oggetto DGEList ottenuto dalla stima della dispersione comune precedente. Come impostazione di default, *estimateTagwiseDisp* usa il *prior.n* calcolato da *getPriorN*. Per il data set di Marioni, edgeR propone un prior pari a 2,5. Il data set comprende 10 campioni e 2 gruppi, quindi i gradi di libertà sono 8. Si nota come la scelta del prior rispetti la regola empirica suggerita, dato che $2{,}5 * 8 = 20$. La funzione *estimateTagwiseDisp* riceve in input l'oggetto DGEList e restituisce un nuovo oggetto DGEList contenente in aggiunta due nuove variabili riferite alla stima genewise: il prior utilizzato (*prior.n*) e la dispersione calcolata per ogni gene (*tagwise.dispersion*). La Figura 25 mostra la dispersione genewise stimata in relazione al livello di espressione del gene corrispondente (logaritmo della media dei count). La linea blu rappresenta la stima della dispersione comune. Si nota come la stima della dispersione genewise decresce all'aumentare del livello di espressione del gene (espressa della log-concentrazione).

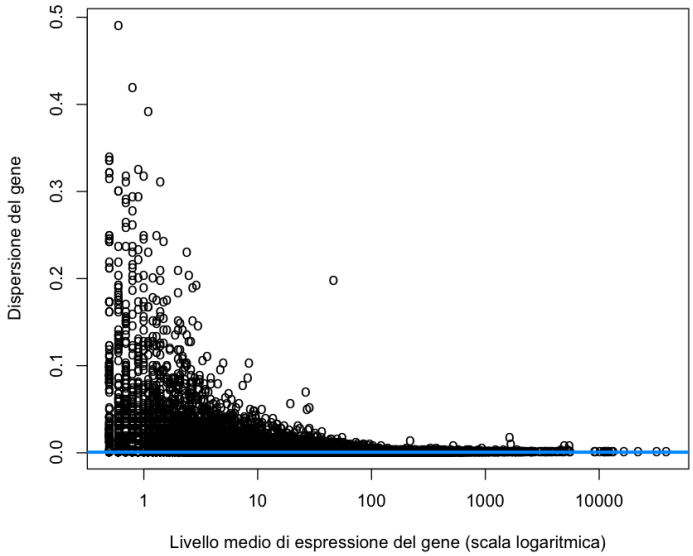


Figura 25. Stima della dispersione genewise con edgeR in relazione al livello di espressione (logaritmo della media dei count). La linea blu rappresenta la dispersione comune stimata.

In modo analogo a quanto visto in precedenza, ottenuta la stima della dispersione genewise è possibile utilizzare il test esatto implementato in *exactTest* per la selezione dei geni differenzialmente espressi. La Tabella 6 mostra i cinque geni a minor p-value, che coincidono con i cinque geni a minor p-value rilevati utilizzando la dispersione comune. Anche in questo caso, i cinque geni riportati presentano p-value nullo.

Gene	log-conc	logFC	p-value	FDR
ENSG00000084734	-17.59644	9.682711	0	0
ENSG00000140107	-16.61081	9.010066	0	0
ENSG00000138075	-16.65683	8.918043	0	0
ENSG00000105697	-14.80112	8.226242	0	0
ENSG00000197901	-16.02888	-8.180912	0	0

Tabella 6. Elenco dei primi cinque geni con minor p-value e rispettive log-concentrazioni, log-fold change e FDR.

La Tabella 7 riporta, per i quattro p-value prefissati, il numero di geni rilevati come differenzialmente espressi (*DE*), sovra-espressi in *liver* (*DE+*) e sovra-espressi in *kidney* (*DE-*). Le selezioni sono simili a quelle ottenute tramite la stima dispersione comune.

p-value	Dati normalizzati TMM		
	N° geni DE	N° geni DE+	N° geni DE-
5%	12175	6546	5629
1%	10635	5744	4891
0,05%	8797	4695	4102
0,01%	8100	4307	3793

Tabella 7. Numero di geni selezionati come differenzialmente espressi (DE), sovra-espressi (DE+) e sotto-espressi (DE-) in *liver* da edgeR con stima della dispersione genewise, per i quattro valori di soglia di p-value (non corretto per test multipli) considerati.

La Figura 26 propone il diagramma di Venn della classificazione dei geni operata da edgeR con stima della dispersione comune e dispersione genewise con un p-value di soglia pari al 5%. La rappresentazione in insiemi permette di evidenziare l'elevato numero di geni classificati come differenzialmente espressi da entrambi gli approcci. Fra i geni rilevati come differenzialmente espressi, oltre il 99% è stato identificato sia con approccio a dispersione comune sia genewise. Questo dimostra che, per il dataset di Marioni, i due metodi offerti da edgeR forniscono risultati robusti.

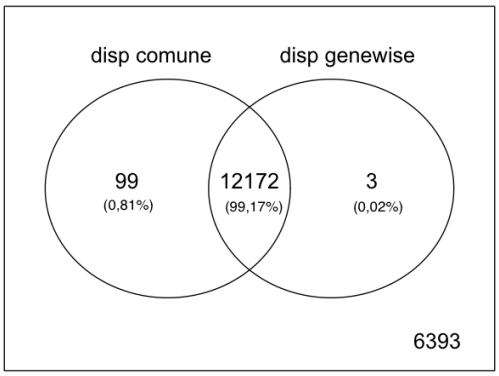


Figura 26. Diagramma di Venn che rappresenta il numero di geni selezionati come differenzialmente espressi da edgeR, con stima della dispersione comune o genewise (p-value≤5%) .

La Figura 27 riporta lo smear plot per l'approccio a dispersione genewise, con i 500 geni a più alta espressione differenziale. I risultati sono altamente confrontabili con quelli in Figura 24.

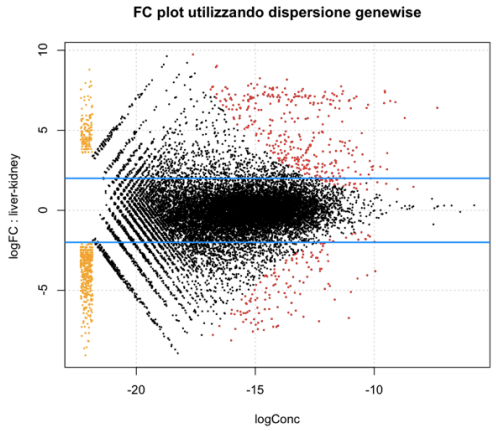


Figura 27. Smear plot di edgeR che confronta log-concentrazione e log-fold change per ciascun gene considerato, utilizzando la stima della dispersione genewise. Valori di log-fold change positivi indicano geni sovra-espressi in *liver*. In rosso sono evidenziati i 500 geni a più alto livello di espressione differenziale, mentre le linee blu rappresentano il limite della regione logFC ± 2

6.3 DESeq

DESeq è stato sviluppato per operare su dataset provvisti di repliche biologiche, condizione non soddisfatta dal data set di Marioni, che comprende repliche tecniche ma non repliche biologiche (tutti i campioni di *liver* e *kidney* sono stati ottenuti da un unico soggetto). Gli autori hanno comunque previsto la possibilità di applicare i metodi di analisi di DESeq in assenza di repliche biologiche, pur precisando che i risultati ottenuti saranno meno robusti e attendibili. In questi casi, gli autori suggeriscono di sommare, per ogni gene, i count misurati su tutti i campioni di una stessa condizione sperimentale [53]. Questo permette di ricondursi alla situazione in cui, per ciascun gene, siano presenti un solo campione (costituito dalla somma dei count delle reali repliche tecniche) per ogni gruppo sperimentale confrontato.

L'utilizzo di DESeq prevede la creazione di un oggetto di tipo CountDataSet, un formato utilizzato dal pacchetto per la memorizzazione dei count, degli identificativi dei geni e dei campioni e di ulteriori dati che vengono generati nelle analisi successive. L'oggetto viene creato tramite la funzione *newCountDataset* a cui vengono forniti in input la matrice dei count e un array che individua gli identificativi delle due condizioni sperimentali (*liver* e *kidney*). Ottenuto l'oggetto CountDataSet è possibile utilizzare la funzione *estimateSizeFactors* che stima, per ogni campione, la library size. Le due stime sono riportate in Tabella 8.

Campione	**Library size**
Kidney	0,7567361
Liver	1,3214646

Tabella 8. Library size determinate per ciascun campione da DESeq.

Si procede con la stima della dispersione, tramite la funzione *estimateDispersions*. La funzione determina la stima in tre passaggi: inizialmente determina un valore empirico di dispersione per ciascun gene, quindi esegue, per ciascuna condizione, il fit di una funzione sui valori stimati della dispersione contro la media dei count. Infine, assegna ad ogni gene un valore di dispersione, scegliendo fra il valore stimato inizialmente o quello ottenuto tramite il fit della funzione, in base al criterio stabilito dal parametro *sharingMode*.

estimateDispersions restituisce in output l'oggetto CountDataSet a cui aggiunge le variabili:

- *df*: gradi di libertà della stima di ciascun gene
- *perGeneDispEsts*: la dispersione empirica stimata per ogni gene con approccio Bayesiano
- *dispFunc*: funzione che approssima i dati (fit)
- *fittedDispEsts*: riporta, per ogni gene, i valori in accordo con il fit della funzione *dispFunc*

estimateDispersions accetta in input i tre parametri opzionali *method*, *sharingMode* e *fitType* che definiscono l'approccio con cui ottenere le stime. Nel dettaglio, il parametro *method* può assumere tre valori:

- *pooled*: il valore predefinito, da utilizzare in presenza di repliche biologiche. Vengono considerati gli identificativi dei campioni riconoscendo le repliche biologiche di una stessa condizione
- *per-condition*: approccio di stima adottato nelle precedenti versioni di DESeq, mantenuto per compatibilità ma non di interesse per questa analisi
- *blind*: approccio di stima con cui DESeq ignora gli identificativi dei campioni, senza aspettarsi che il dataset in input contenga repliche biologiche, da utilizzare in assenza di tali repliche

Il parametro *sharingMode* può assumerei valori:

- *maximum*: il valore di default, da adottare con dataset provvisti di repliche biologiche. Considera come parametro di dispersione in massimo fra il valore empirico stimato inizialmente e quello ottenuto dal fit della funzione
- *fit-only*: viene utilizzato il solo valore di dispersione ottenuta tramite il fit dei dati. È il valore consigliato per data set privi di repliche biologiche
- *gene-est-only*: viene utilizzato il solo valore empirico Bayesiano, senza applicare nessun fit ai dati. È il valore consigliato quando si hanno a disposizione molte repliche biologiche e il valore empirico della dispersione è ragionevolmente realistico

Infine, il parametro *fitType* può assumere due valori:

- *parametric*: il valore di default, vengono calcolati due coefficienti (indicati nella variabile *dispFunc*) che parametrizzano il fit dei dati con un GLM
- *locfit*: viene eseguito un fit locale ad opera del pacchetto *locfit* [54] . Il reciproco della radice quadrata della varianza stimata dei count normalizzati viene integrato numericamente e l'integrale così ottenuto viene valutato per ogni count

Per l'analisi dei dati di Marioni il metodo *estimateDispersions* è stato invocato con i seguenti parametri: *method="blind"*, *sharingMode="fit-only"*, *fitType="locfit".* Il parametro *fitType* è stato impostato a *locfit* per ragioni di convergenza dell'algoritmo, che con il dataset utilizzato non viene invece raggiunta se impostato a *parametric*.

Quando il parametro *fitType* è settato a *locfit*, DESEq supporta l'ulteriore parametro opzionale *lp_extra_args* che permette di definire le caratteristiche del fit locale. Sono state provate le seguenti parametrizzazioni, come in [49]:

- *nn=0.5*: definisce il parametro di smoothing "nearest neighbor". Utilizzare il valore 0,5 significa che il numero di "dati vicini" considerati per ottenere una stima smooth viene scelto in modo da coprire il 50% dei dati. Il valore di default è 0,7
- *deg=1*: definisce il numero di gradi del polinomio che viene utilizzato. Il valore di default è 2

Dopo avere stimato la dispersione con i tre approcci descritti, è stato applicato il test di DESeq per l'analisi differenziale. In aggiunta, è stata effettuata l'analisi con DESeq sui dati originali di Marioni, considerando separatamente le diverse repliche tecniche, senza sommare i count nei due gruppi sperimentali. In questa analisi, il metodo estimateDispersions è stato invocato con parametri *method="blind"*, *sharingMode="fit-only"*.

DESeq invoca il test di analisi differenziale attraverso la funzione *nbinomTest,* che riceve in input l'oggetto CountDataSet elaborato in precedenza e l'elenco degli identificativi delle due condizioni sperimentali da confrontare. *nbinomTest* fornisce in output una matrice che riporta, per ogni gene:

- *baseMean*: la media dei count, dopo averli trasformati in *common scale*
- *baseMeanA*, *baseMeanB*: la media in common scale, per le due condizioni sperimentali confrontate
- *foldChange*, *log2foldChange*: il livello di espressione differenziale, in scala naturale e logaritmica
- *pval*: il p-value calcolato per l'analisi differenziale fra le due condizioni
- *padj*: il p-value corretto per test multipli (secondo l'approccio Benjamini-Hochberg)

La Tabella 9 riporta il numero di geni differenzialmente espressi, valutati sulle quattro diverse combinazioni di parametri della funzione *estimateDispersions*.

p-value	N° geni DE			
	standard	***lp_extra_args= list(deg=1)***	***lp_extra_args= list(nn=0.5)***	***Repliche tecniche***
5%	1262	1247	1271	6268
1%	610	586	601	4505
0,05%	214	170	212	2862
0,01%	106	64	107	2357

Tabella 9. Numero di geni rilevati come differenzialmente espressi da DESeq, per i quattro valori di soglia di p-value (non corretto per test multipli) considerati, per le quattro modalità di DESeq testate.

Risulta evidente come il numero di geni differenzialmente espressi rilevati da DESeq con le indicazioni fornite dagli autori sia molto basso, a dimostrazione delle limitazioni che il pacchetto incontra se applicato a dati privi di repliche biologiche.

In Figura 28 è riportato l'MA-plot fra la media dei count in common scale (*baseMean)* e il livello di espressione differenziale (*log2foldChange*), per ciascun gene. È interessante notare come alcuni geni a più alta espressione (valori elevati di *baseMean*) non siano stati classificati come differenzialmente espressi, nonostante presentino valori elevati di log-fold change.

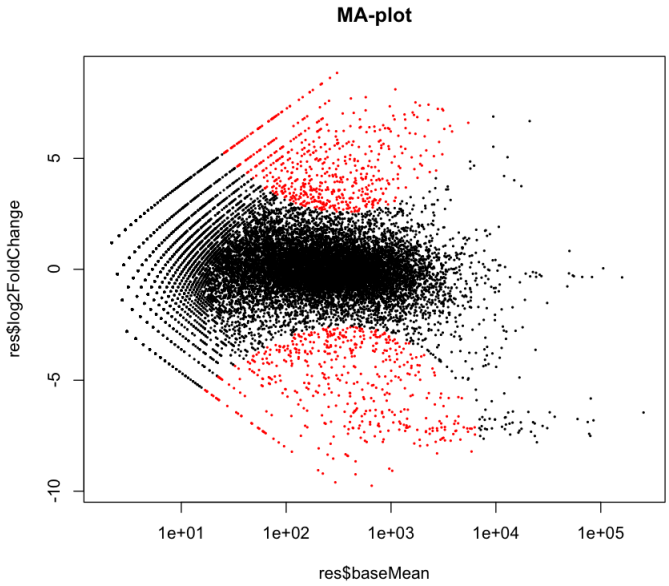


Figura 28. MA-plot di confronto fra livello di espressione (media dei count in common scale, *baseMean*) e log-fold change determinati per ciascun gene da DESeq (configurazione standard di *estimateDispersions*), con p-value≤5%. In rosso sono evidenziati i geni differenzialmente espressi.

6.4 PoissonSeq

L'analisi con PoissonSeq richiede di invocare la sola funzione *PS.Main*: riceve in input la lista *dat* che descrive i dati e restituisce una tabella contenente i risultati dell'analisi differenziale. *dat* deve contenere i seguenti campi:

- *n*: matrice dei count
- *y*: vettore che descrive la condizione sperimentale a cui appartiene ogni campione
- *type*: stabilisce il tipo di confronto da effettuare. I valori previsti sono "*twoclass*" (per esperimenti a due classi), "*multiclass*" (per esperimenti con più di due condizioni sperimentali) e "*quant*" (per quantificazione dati messi in relazione con variabili

quantitative). Il data set di Marioni prevede un confronto a due classi (*liver* contro *kidney*) quindi è stato impostato il valore "*twoclass*"

- *pair*: valore booleano che stabilisce se il confronto debba essere eseguito a coppie o meno. Se impostato a *true* i count di un singolo campione della condizione sperimentale A vengono confrontati con quelli di un singolo campione della condizione B. É stato utilizzato il valore di default, cioè *false* (confronto non a coppie)
- *gname*: vettore che contiene gli identificativi dei geni

L'output che *PS.Main* restituisce contiene le seguenti informazioni:

- *gname:* gli identificativi dei geni, ordinati per livello di espressione differenziale
- *tt:* i valori delle statistiche determinate per ogni gene
- *pval:* i p-value determinati per ogni gene
- *fdr: i* p-value corretti per test multipli]
- *log.fc*: le log-fold change per ogni gene

PoissonSeq permette inoltre di visualizzare la library size stimata per ciascuna replica del data set tramite la funzione *PS.Est.Depth*. La Tabella 10 ne riporta il riassunto.

Replica	**Gruppo**	**Fattore di normalizzazione**
R1L1Kidney	Kidney	1,2755554
R1L2Liver	Liver	0,7702374
R1L3Kidney	Kidney	1,3169896
R1L4Liver	Liver	0,7728523
R1L6Liver	Liver	0,7433143
R1L7Kidney	Kidney	1,2374063
R1L8Liver	Liver	0,7159739
R2L2Kidney	Kidney	1,3605152
R2L3Liver	Liver	0,8033867
R2L6Kidney	Kidney	1,3892718

Tabella 10. Fattore di normalizzazione delle library size determinato, per ciascuna replica, da PoissonSeq.

Dato che sia PoissonSeq sia edgeR permettono di ottenere la stima del fattore di normalizzazione, si desidera investigare se tali stime siano correlate fra loro. L'analisi della correlazione permette di ottenere un'indicazione sull'esistenza di una relazione fra i fattori di normalizzazioni stimati da edgeR (Tabella 3) e da PoissonSeq (Tabella 10). Le correlazione è positiva e l'indice di correlazione è pari a 0,989. Si tratta di un valore molto alto, che dimostra come i fattori di normalizzazione ottenuti dai due metodi mostrino variazioni coerenti fra loro. Inoltre, per entrambi i metodi, sono stati valutati media e deviazione standard dei fattori di normalizzazioni, valutati nei gruppi sperimentali a cui sono riferiti (*liver* e *kidney*). I risultati sono riportati in Tabella 11.

	PoissonSeq		edgeR	
	Liver	**Kidney**	**Liver**	**Kidney**
Media	0,7611529	1,315948	0,8176146	1,2230811
SD	0,0330244	0,061571	0,0032493	0,0033235

Tabella 11. Media e deviazione standard dei fattori di normalizzazioni calcolati da PoissonSeq e edgeR, valutate per ciascun gruppo sperimentale

La Tabella 12 riporta il numero di geni classificati come differenzialmente espressi da PoissonSeq, per i quattro p-value di soglia fissati.

p-value	Geni DE
5%	11828
1%	9853
0,05%	7296
0,01%	6578

Tabella 12. Numero di geni rilevati come differenzialmente espressi da PoissonSeq, per i quattro valori di soglia di p-value considerati

6.5 Confronto con microarray

Per confrontare le prestazioni dei diversi metodi applicati al data set di Marioni sono state determinate le curve ROC a diversi livelli di p-value. Le curve ROC (*Relative Operating Characteristic*) sono dei grafici che rappresentano le prestazioni in termini di sensitività (asse x) e precisione (asse y). Dopo aver determinato il numero di veri positivi (*TP*), veri negativi (*TN*), falsi positivi (*FP*) e falsi negativi (*FN*), sensitività e precisione sono state definite come

$$sensitività = TP/tot_DE_goldstandard$$

$$precisione = TP/(TP + FP)$$

La determinazione di tali parametri richiede di fissare un *gold standard*. Per il data set di Marioni, il gold standard è dato dai risultati che gli autori hanno ottenuto dall'analisi differenziale di dati da microarray sugli stessi campioni di *kidney* e *liver*. Sui dati di espressione genica (microarray) della *SupplementaryTable3.txt* (vedi capitolo 5.1, "Data set di Marioni *et al.*") si è imposta una soglia sulla FDR a 0,1%, selezionando come differenzialmente espressi (*DE*) 8113 geni. Tra questi, è stato calcolato il numero di geni rilevati come sovra-espressi in *liver* rispetto a *kidney* (*DE+*) e quelli sotto-espressi (*DE-*, ovvero sovra-espressi in *kidney*). Nel gold standard sono presenti 3488 geni *DE+* e 4625 geni *DE-*. Sono stati quindi identificati i geni classificati come *nonDE*, *DE+* e *DE-* dai metodi illustrati in precedenza, imponendo una crescente sui p-value (non corretti). I geni selezionati da ciascun metodo sono stati confrontati con quelli del gold standard per identificare veri positivi (*TP*), falsi positivi (*FP*), veri nagativi (*TN*) e falsi negativi (*FN*), secondo il criterio descritto in Tabella 13. Sono stati considerati i soli geni presenti sia nel data set da cui è stato ottenuto il gold standard (i 17708 geni presenti nella *SupplementaryTable3.txt*) sia nella tabella di output di ciascun metodo (i geni per i quali ciascun metodo ha determinato un p-value).

		Metodo RNA-Seq		
		DE+	DE-	nonDE
Gold standard	DE+	TP	FP	FN
	DE-	FP	TP	FN
	nonDE	FP	FP	TN

Tabella 13. Tabella di contingenza tramite cui si definiscono le prestazioni della classificazione di ciascun metodo (RNA-Seq) rispetto al gold standard (dati da microarray). Le intersezioni permettono di determinare TP, TN, FP e FN.

É quindi possibile ottenere la curva ROC per ciascun metodo considerato, determinando:

$$sensitività = TP/tot_DE_goldstandard$$

$$precisione = TP/(TP + FP)$$

La precisione è ottenuta seconda la formulazione tradizionale mentre la sensitività è definita come numero dei veri positivi rispetto al numero di geni selezionati *DE* nel gold standard, come in [50]. È così possibile considerare l'intero set di geni differenzialmente espressi secondo il gold standard, che comprende oltre ai *TP* e *FP*, anche quei geni *DE* correttamente selezionati come *DE* da un metodo ma con un errata direzione di espressione differenziale (*FP* di Tabella 13). La Figura 29 propone le curve ROC di tutti i metodi applicati al data set di Marioni.

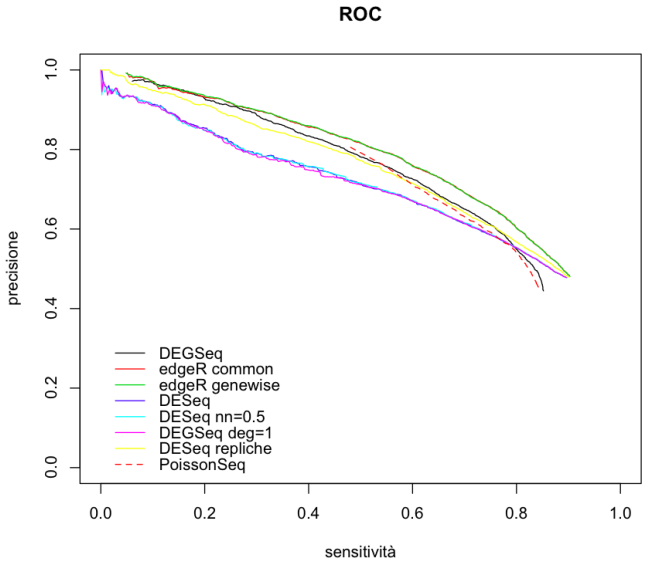


Figura 29. Curve ROC (precisione-sensitività) determinate per i metodi applicati al dataset di Marioni.

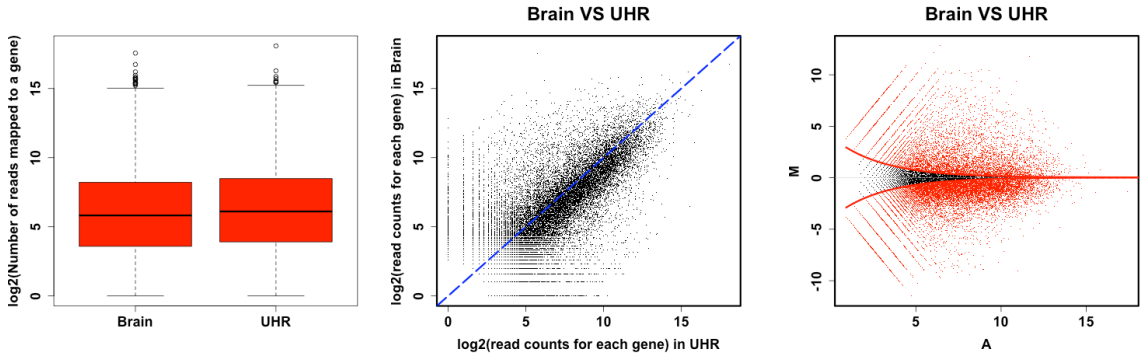
L'analisi visiva delle curve ROC permette di notare come edgeR risulti il metodo che offre le migliori prestazioni, in termini di sensitività e precisione, sul data set di Marioni. Su questo data set, la stima con dispersione comune e dispersione *genewise* ottengono performance praticamente identiche. DESeq applicato al "pooling" delle repliche tecniche ottiene il risultato peggiore, in tutte le quattro configurazioni testate; i risultati migliorano invece applicando DESeq al dataset originale.
È inoltre interessante notare come la curva ROC di PoissonSeq abbia origine in valori di sensitività molto superiori rispetto agli altri metodi (0,48 contro 0,05 circa). Ciò è dovuto all'elevato numero di geni a cui il metodo assegna un basso p-value.

7 Analisi data set MAQC

Prima di procedere con l'elaborazione, ai dati MAQC è stato applicato lo stesso filtraggio utilizzato sul data set di Marioni *et al.* (capitolo 6, "Analisi data set di Marioni *et al.*") in modo da eliminare i geni con media dei count (valutata sulle 14 repliche disponibili) inferiore a 0,5. Applicando tale correzione si passa dai 51760 geni originali a 23817 geni.

7.1 DEGSeq MARS

In Figura 30 sono riportati boxpot, scatterplot e MA-plot ottenuti applicando al data set il metodo MARS di DEGSeq, con p-value ≤ 5%. Visivamente, si nota come i campioni di *UHR* presentino un maggior numero di geni espressi. Tale situazione produce lo shift verso il basso dello scatterplot e dell'MA-plot.



(a) (b) (c)

Figura 30. L'output grafico prodotto da DEGseq con il metodo MARS: (a) boxplot, (b) scatterplot e (c) MA-plot. p-value ≤ 5%

L'analisi differenziale è stata valutata per quattro diversi livelli di significatività. La Tabella 14 riassume il numero di geni identificati come differenzialmente espressi, fissando il p-value di soglia (non corretto) al 5%, 1%, 0,05% e 0,01%.

p-value	N° geni DE
5%	18119
1%	16239
0,05%	13835
0,01%	12985

Tabella 14. Numero di geni selezionati come differenzialmente espressi dal metodo MARS, per i quattro valori di soglia di p-value (non corretto per test multipli) considerati.

7.2 edgeR

In Figura 31 è riportato l'MDS-plot generato da edgeR sul data set MAQC. Si nota come la dimensione 1 separi nettamente i campioni di *Brain* dai campioni di *UHR*, indicando che è lecito attendersi un elevato numero di geni differenzialmente espressi tra i due gruppi.
É inoltre interessante notare come diverse repliche di UHR vengano rappresentate praticamente sovrapposte, ad indicare una bassa variabilità nei loro dati.

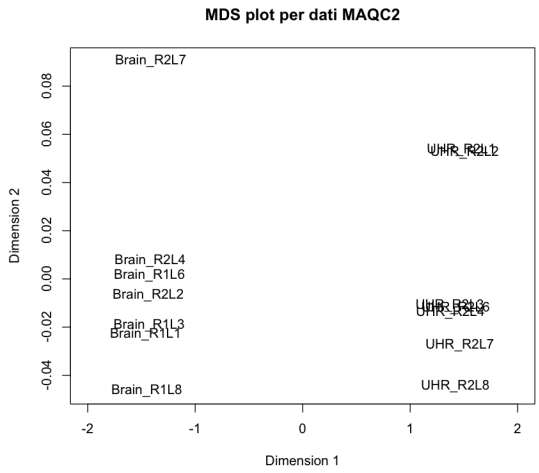


Figura 31. MDS-plot ottenuto con edgeR per le repliche del dataset MAQC

L'analisi differenziale del data set è stata svolta con entrambi gli approcci di stima della dispersione permessi da edgeR: dispersione comune e genewise.

7.2.1 Analisi differenziale con stima della dispersione comune

La dispersione comune stimata è pari a 0,0004, dunque il coefficiente di variazione è di 0,0197.
Ciò significa che considerando un ipotetico gene con media dei count $\mu = 500$, la varianza del modello NB risulterà $1 + \mu\phi = 1 + 500 * 0{,}0004 = 1{,}20$ volte superiore a quella prevista dal modello di Poisson. La Figura 32 rappresenta la relazione fra media e varianza nei dati descritta secondo il modello di Poisson e i due approcci di edgeR. Si nota come, anche in questo caso, entrambe le varianze stimate da edgeR descrivono i dati più fedelmente rispetto al modello di Poisson.

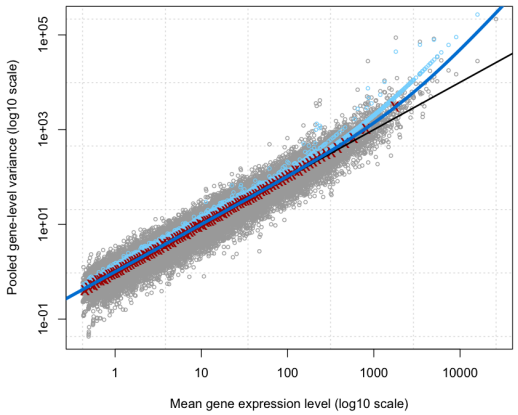
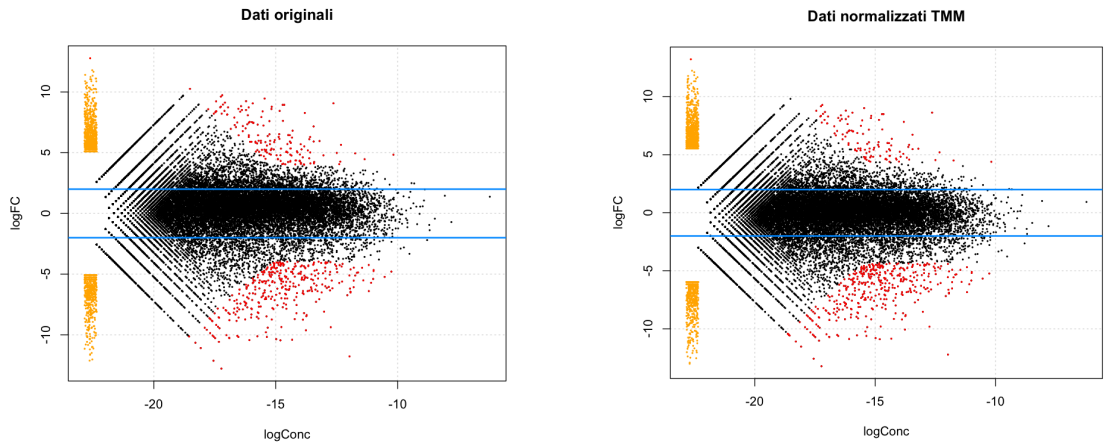


Figura 32. Rapporto tra media-varianza dei count per ogni gene: i punti grigi indicano la varianza grezza di ciascun gene rispetto alla sua abbondanza (asse x). Le croci rosse indicano la media della varianza grezza dei geni raggruppati in 100 bin sulla base della loro abbondanza. I punti in azzurro mostrano la varianza stimata per ciascun gene, calcolata dalla dispersione genewise. La linea blu rappresenta la varianza stimata utilizzando la dispersione comune. La linea nera rappresenta la varianza prevista dal modello di Poisson, uguale alla media.

Stimata la dispersione comune è possibile applicare il test esatto fornito da edgeR per l'analisi differenziale.
Lo smear plot di Figura 33 evidenzia la distribuzione dei primi 500 geni a più alta espressione. È immediato notare l'effetto che la normalizzazione TMM (capitolo 4.2.2, "Normalizzazione") introduce sulla distribuzione dei dati: lo smear plot dei dati non normalizzati (Figura 33a) presenta uno shift verso l'alto dovuto all'elevato numero di geni sovra-espressi in *UHR*.



(a) (b)

Figura 33. Smear plot di edgeR che confronta log-concentrazione e log-fold change per ciascun gene considerato. Valori di log-fold change positivi indicano geni sovra-espressi in *Brain*. In (a) sono riportati i risultati ottenuti sui dati originali e in (b) quelli ottenuti sui dati normalizzati con il metodo TMM. In rosso sono evidenziati i 500 geni a più alto livello di espressione differenziale, mentre le linee blu rappresentano il limite della regione logFC ± 2. La maggior distribuzione di punti rossi nella parte basse evidenza come un maggior numero di geni sia sovra-espresso in *UHR*.

Dalle liste dei geni e relativi p-value ottenuti, sono stati selezionati i geni differenzialmente espressi tra le due condizioni sperimentali, impostando le quattro diverse soglie sui p-value non corretti. Il riassunto della selezione è riporato in Tabella 15.

p-value	Dati originali			Dati normalizzati TMM		
	N° geni DE	N° geni DE+	N° geni DE-	N° geni DE	N° geni DE+	N° geni DE-
5%	17387	6135	11252	16783	7899	8884
1%	15742	5554	10188	15021	7217	7804
0,05%	13534	4817	8717	12814	6265	6549
0,01%	12675	4532	8143	11957	5906	6051

Tabella 15. Numero di geni selezionati come differenzialmente espressi (DE), sovra-espressi (DE+) e sotto-espressi (DE-) in *Brain* da edgeR con stima della dispersione comune, per i quattro valori di soglia di p-value (non corretto per test multipli) considerati. Le tre classificazioni sono state valutate sia sui dati originali sia sui dati normalizzati con il metodo TMM.

7.2.2 Analisi con stima della dispersione genewise

La Figura 34 mostra la distribuzione delle stime della dispersione genewise in relazione al livello di espressione del gene corrispondente (espresso come logaritmo della media dei count), mentre la linea blu indica la stima della dispersione comune. Anche in questo caso, le stime della dispersione genewise decrescono all'aumentare del livello di espressione del gene (espressa nel grafico della log-concentrazione).

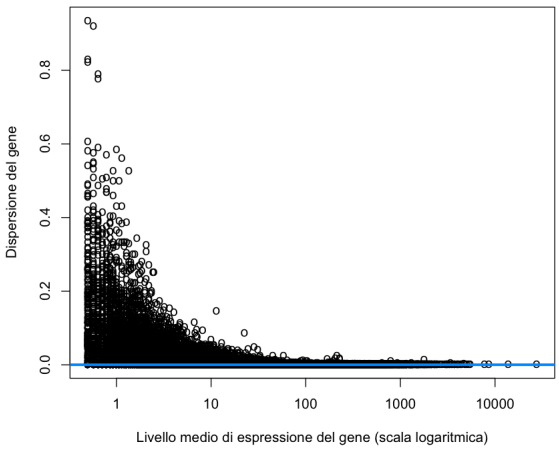


Figura 34. Stima della dispersione genewise con edgeR in relazione al livello di espressione (logaritmo della media dei count). La linea blu rappresenta la dispersione comune stimata.

Come per l'approccio a stima della dispersione comune, una volta determinata la dispersione genewise è possibile applicare il test esatto per la selezione dell'analisi differenziale. La Tabella 16 riporta, per i quattro p-value prefissati, il numero di geni rilevati come differenzialmente espressi (*DE*), sovra-espressi in *Brain* (*DE+*) e sovra-espressi in *UTH* (*DE-*). Si nota come le selezioni sono abbastanza simili a quelle ottenute tramite la stima dispersione comune (Tabella 15).

p-value	N° geni DE	N° geni DE+	N° geni DE-
5%	16660	7838	8822
1%	14864	7144	7720
0,05%	12626	6179	6447
0,01%	11787	5826	5961

Tabella 16. Numero di geni selezionati come differenzialmente espressi (DE), sovra-espressi (DE+) e sotto-espressi (DE-) in *Brain* da edgeR con stima della dispersione genewise, per i quattro valori di soglia di p-value (non corretto per test multipli) considerati.

La Figura 26 propone il diagramma di Venn della classificazione dei geni operata da edgeR con stima della dispersione comune e dispersione genewise con un p-value di soglia pari al 5%. Fra i geni rilevati come differenzialmente espressi, oltre il 99% è stato identificato sia con approccio a dispersione comune sia genewise.

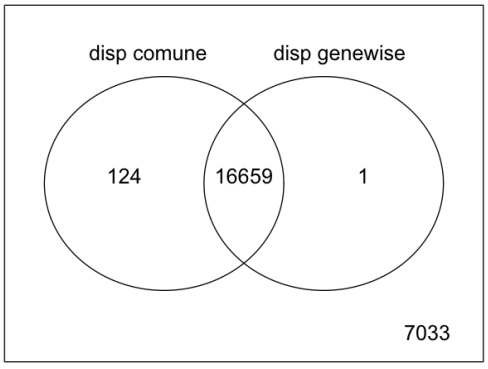


Figura 35. Diagramma di Venn che rappresenta il numero di geni selezionati come differenzialmente espressi da edgeR, con stima della dispersione comune o genewise (p-value≤5%) .

La Figura 36 propone lo smear plot per l'approccio a dispersione genewise, con i 500 geni a più alta espressione differenziale evidenziati in rosso. I risultati sono ben confrontabili con quelli ottenuti con approccio di stima della dispersione comune.

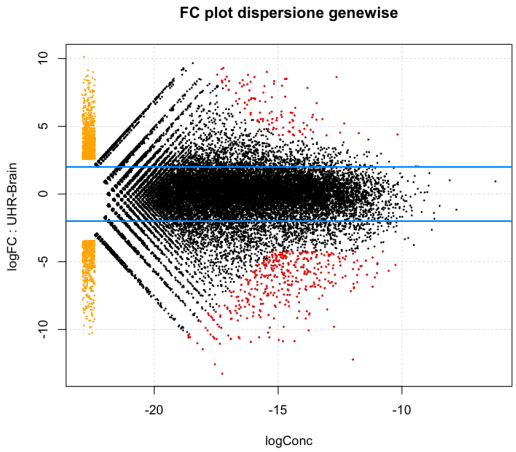


Figura 36. Smear plot di edgeR che confronta log-concentrazione e log-fold change per ciascun gene considerato, utilizzando la stima della dispersione genewise. Valori di log-fold change positivi indicano geni sovra-espressi in *Brain*. In rosso sono evidenziati i 500 geni a più alto livello di espressione differenziale, mentre le linee blu rappresentano il limite della regione logFC ± 2. La maggior distribuzione di punti rossi nella parte basse evidenza come un maggior numero di geni sia sovra-espresso in *UHR*.

7.3 DESeq

Come il data set di Marioni, anche MAQC non prevede repliche biologiche ma solo repliche tecniche. Quindi, come suggeriscono gli autori, sono stati sommati, per ogni gene, i count misurati su tutti i campioni di una stessa condizione sperimentale.

DESeq prevede la stima della dispersione tramite la funzione *estimateDispersions*, a cui possono essere forniti dei parametri che definiscono l'algoritmo di stima da utilizzare. Per l'analisi del data set di MAQC sono state utilizzate le stesse quattro combinazioni di parametri utilizzate per l'analisi del data set di Marioni *et al.* (capitolo 6.3, "DESeq").

La Tabella 17 riporta il numero di geni differenzialmente espressi, valutati con quattro valori di soglia per il p-value sulle quattro diverse combinazioni di parametri della funzione *estimateDispersions*.

p-value	**N° geni DE**			
	standard	***lp_extra_args= list(deg=1)***	***lp_extra_args= list(nn=0.5)***	***Repliche tecniche***
5%	1726	1687	1725	11085
1%	687	647	686	8438
0,05%	51	40	51	5745
0,01%	0	1	0	4867

Tabella 17. Numero di geni rilevati come differenzialmente espressi da DESeq, per i quattro valori di soglia di p-value (non corretto per test multipli) considerati, per le quattro modalità di DESeq testate.

Risulta evidente come il numero di geni differenzialmente espressi rilevati da DESeq con le indicazioni fornite dagli autori sia molto basso. Anche in questo caso, DESeq mostra forti limitazioni se applicato a dati privi di repliche biologiche.

In Figura 37 è riportato l'MA-plot fra la media dei count in common scale (*baseMean)* e il livello di espressione differenziale (*log2foldChange*), per ciascun gene compreso nel data set.

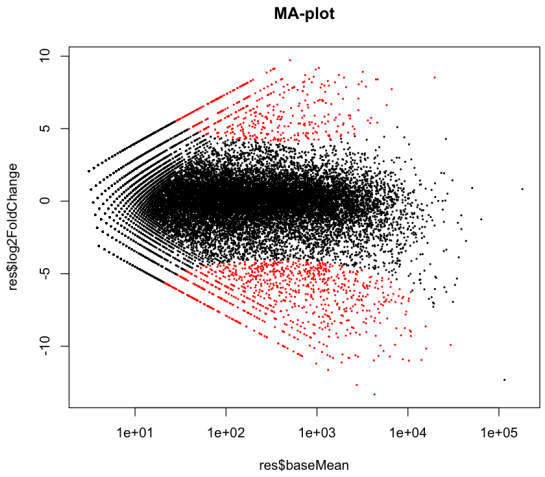


Figura 37. MA-plot di confronto fra livello di espressione (media dei count in common scale, *baseMean*) e log-fold change determinati per ciascun gene da DESeq (configurazione standard di *estimateDispersions*), con p-value≤5%. In rosso sono evidenziati i geni differenzialmente espressi.

7.4 PoissonSeq

La Tabella 18 riporta il numero di geni classificati come differenzialmente espressi da PoissonSeq, per i quattro p-value di soglia fissati.

p-value	Geni DE
5%	15883
1%	13209
0,05%	9906
0,01%	8957

Tabella 18. Numero di geni rilevati come differenzialmente espressi da PoissonSeq, per i quattro valori di soglia di p-value considerati

7.5 Confronto con qRT-PCR

Per confrontare le prestazioni dei diversi metodi applicati al data set MAQC sono state determinate le curve ROC (capitolo 6.5, "Confronto con microarray") per crescenti livelli di p-value.

Il gold standard da intendere come rappresentazione della realtà è stato individuato nelle classificazioni su dati ottenuti da qRT-PCR, elaborati come descritto di seguito. Gli identificativi delle misure ottenute con qRT-PCR sono stati confrontati con i geni annotati nel database Ensembl [46], scartando le misure che ottenevano il match con più geni. Sono stati poi mantenuti i soli geni identificati come espressi in almeno tre delle quattro run dell'esperimento, con errore standard misurato sulle otto run inferiore a 0,25. La log-fold change fra *Brain* e *UHR* è stata determinata come il rapporto, in scala logaritmica, fra la media delle repliche di *UHR* e la media delle repliche di *Brain*. Quindi, i geni con un valore assoluto di log-fold change superiore a 0,5 e inferiore a 0,2 sono stati selezionati rispettivamente come differenzialmente espressi (*DE*) e non differenzialmente espressi (*nonDE*). I geni *DE*, sono infine stati separati, in base al segno della log-fold change, in sovra-espressi in Brain (*DE+*) e

sotto-espressi (*DE*-). La lista finale di geni che compongono il gold standard da qRT-PCR comprende 721 geni.

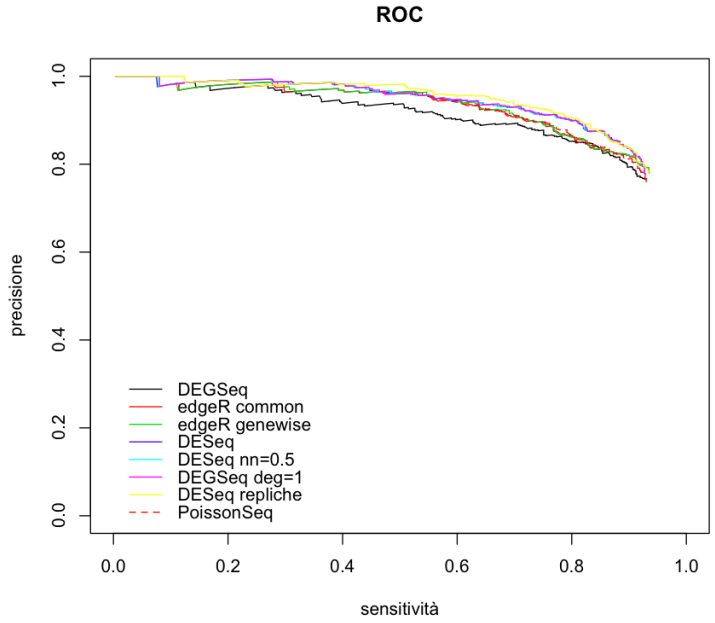


Figura 38. Curve ROC determinate per i metodi applicati al dataset MAQC.

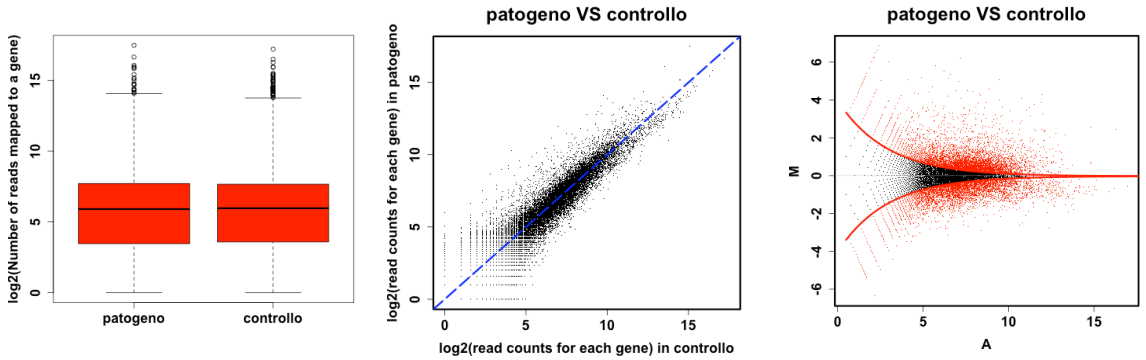
L'analisi visiva delle ROC per il data set MAQC permette di notare come i diversi metodi testati ottengano prestazioni di classificazione molto simili fra loro. Su questi dati, DESeq applicato al data set originale, ottiene risultati lievemente migliori rispetto agli altri metodi. Le prestazioni di DEGSeq risultano le più basse, seppure lo scostamento tra le curve ROC fra metodo peggiore e metodo migliore sia molto limitato su questo dataset.

8 Analisi data set di Cumbie *et al.*

Nella fase di pre-elaborazione dei dati, al data set di Cumbie *et al.* è stato applicato lo stesso filtraggio utilizzato sui precedenti data set, mantenendo i soli geni aventi media dei count (valutata sulle 6 repliche disponibili) pari ad almeno 0,5. Applicando tale correzione si passa dei 26220 geni originali ne vengono mantenuti 23963.

8.1 DEGSeq MARS

In Figura 30 sono riportati boxpot, scatterplot e MA-plot ottenuti applicando al data set il metodo MARS di DEGSeq, con p-value ≤ 5%. Visivamente, si nota come il boxplot evidenzi una distribuzione molto simile fra *patogeno* e *controllo*. Difatti, i dati sono rappresentati sullo scatterplot con con una buona simmetria rispetto alla diagonale e rispetto a *M=0* sull'MA-plot.



(a) (b) (c)

Figura 39. L'output grafico prodotto da DEGseq con il metodo MARS: (a) boxplot, (b) scatterplot e (c) MA-plot. p-value ≤ 5%

L'analisi differenziale è stata valutata per quattro diversi livelli di significatività. La Tabella 14 riassume il numero di geni identificati come differenzialmente espressi, fissando il p-value di soglia (non corretto) al 5%, 1%, 0,05% e 0,01%.

p-value	N° geni DE
5%	11543
1%	8544
0,05%	5955
0,01%	5159

Tabella 19. Numero di geni selezionati come differenzialmente espressi dal metodo MARS, per i quattro valori di soglia di p-value (non corretto per test multipli) considerati.

8.2 edgeR

L'analisi differenziale del data set è stata svolta con entrambi gli approcci di stima della dispersione permessi da edgeR: dispersione comune e genewise.

8.2.1 Analisi differenziale con stima della dispersione comune

La dispersione comune stimata è molto elevata, pari a 0,343 (coefficiente di variazione 0,585).
Ciò significa che per un ipotetico gene con media dei count $\mu = 500$, la varianza del modello NB risulterà $1 + \mu\phi = 1 + 500 * 0{,}343 = 172$ volte superiore a quella prevista dal modello di Poisson. L'analisi di questo data set mostra quindi in maniera decisamente evidente l'inadeguatezza del modello di Poisson: la Figura 32 rappresenta la relazione fra media e varianza nei dati descritta secondo il modello Poissoniano e i due approcci di edgeR. È subito evidente come le varianze stimate da edgeR descrivano i dati molto più fedelmente rispetto al modello di Poisson.

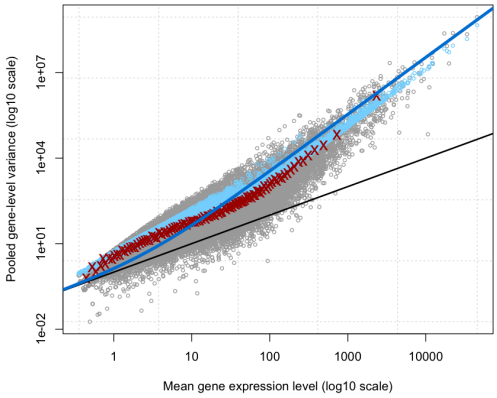


Figura 40. Rapporto tra media-varianza dei count per ogni gene: i punti grigi indicano la varianza grezza di ciascun gene rispetto alla sua abbondanza (asse x). Le croci rosse indicano la media della varianza grezza dei geni raggruppati in 100 bin sulla base della loro abbondanza. I punti in azzurro mostrano la varianza stimata per ciascun gene, calcolata dalla dispersione genewise. La linea blu rappresenta la varianza stimata utilizzando la dispersione comune. La linea nera rappresenta la varianza prevista dal modello di Poisson, uguale alla media.

Ottenuta la stima della dispersione comune si può applicare il test esatto fornito da edgeR per l'analisi differenziale.
Lo smear plot di Figura 33 evidenzia la distribuzione dei primi 500 geni a più alta espressione. Si nota la presenza di un maggior numero di geni fortemente espressi in *patogeno*.

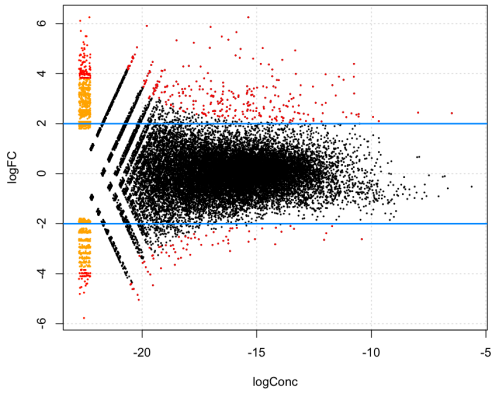


Figura 41. Smear plot di edgeR che confronta log-concentrazione e log-fold change per ciascun gene considerato. Valori di log-fold change positivi indicano geni sovra-espressi in *patogeno*. In rosso sono evidenziati i 500 geni a più alto livello di espressione differenziale, mentre le linee blu rappresentano il limite della regione logFC ± 2. La maggior distribuzione di punti rossi nella parte alta evidenza come un maggior numero di geni sia sovra-espresso in *patogeno*.

Dalle liste dei geni e relativi p-value ottenuti, sono stati selezionati i geni differenzialmente espressi tra le due condizioni sperimentali, impostando le quattro diverse soglie sui p-value non corretti. Il riassunto della selezione è riporato in Tabella 15.

p-value	N° geni DE	N° geni DE+	N° geni DE-
5%	1974	719	1255
1%	722	190	532
0,05%	168	18	150
0,01%	90	2	88

Tabella 20. Numero di geni selezionati come differenzialmente espressi (DE), sovra-espressi (DE+) e sotto-espressi (DE-) in *patogeno* da edgeR con stima della dispersione comune, per i quattro valori di soglia di p-value (non corretto per test multipli) considerati.

8.2.2 Analisi con stima della dispersione genewise

La Figura 34 mostra l'andamento della dispersione genewise stimata, in relazione al livello di espressione del gene corrispondente (logaritmo della media dei count). La linea blu indica la stima della dispersione comune.

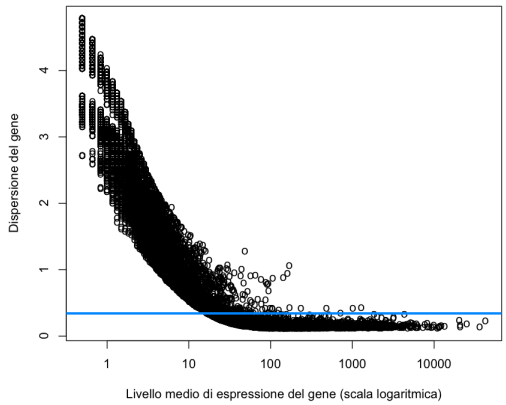


Figura 42. Stima della dispersione genewise con edgeR in relazione al livello di espressione (logaritmo della media dei count). La linea blu rappresenta la dispersione comune stimata.

Come per l'approccio a stima della dispersione comune, una volta determinata la dispersione genewise è possibile applicare il test esatto per la selezione dell'analisi differenziale. La Tabella 16 riporta, per i quattro p-value prefissati, il numero di geni rilevati come differenzialmente espressi (*DE*), sovra-espressi in *patogeno* (*DE+*) e sovra-espressi in *controllo* (*DE-*). Si nota come le selezioni siano più numerose rispetto a quelle ottenute tramite la stima dispersione comune (Tabella 15), ma con un'eccezione: per p-value ≤ 5%, il numero di geni *DE* con dispersione genewise risulta più basso del suo corrispettivo con dispersione comune.

p-value	N° geni DE	N° geni DE+	N° geni DE-
5%	1848	805	1043
1%	853	282	571
0,05%	325	51	274
0,01%	232	29	203

Tabella 21. Numero di geni selezionati come differenzialmente espressi (DE), sovra-espressi (DE+) e sotto-espressi (DE-) in *Brain* da edgeR con stima della dispersione genewise, per i quattro valori di soglia di p-value (non corretto per test multipli) considerati.

La Figura 26 propone il diagramma di Venn della classificazione dei geni operata da edgeR con stima della dispersione comune e dispersione genewise con un p-value di soglia pari al 5%. A differenza dei data set di Marioni e MAQC, i geni rilevati come differenzialmente espressi sia con approccio a dispersione comune sia genewise sono solamente il 28% del totale dei geni identificati come differenzialmente espressi.

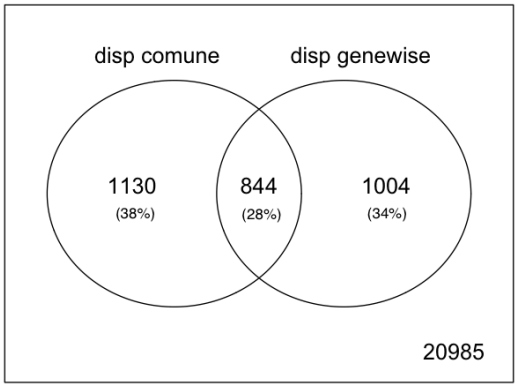


Figura 43. Diagramma di Venn che rappresenta il numero di geni selezionati come differenzialmente espressi da edgeR, con stima della dispersione comune o genewise (p-value≤5%) .

La Figura 36 propone lo smear plot per l'approccio a dispersione genewise, con i 500 geni a più alta espressione differenziale evidenziati in rosso. Si nota come i geni sovra-espressi in *patogeno* siano molto più numerosi di quelli sovra-espressi in *controllo*. Inoltre, molti geni classificati come differenzialmente espressi presentano log-fold change (livello di espressione differenziale) sensibilmente inferiore a ±2.

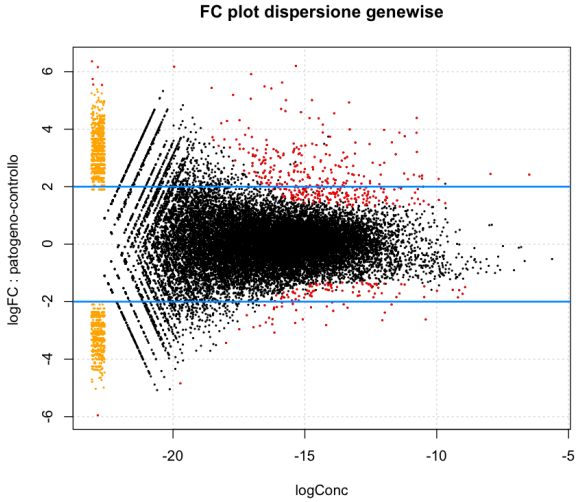


Figura 44. Smear plot di edgeR che confronta log-concentrazione e log-fold change per ciascun gene considerato, utilizzando la stima della dispersione genewise. Valori di log-fold change positivi indicano geni sovra-espressi in *patogeno*. In rosso sono evidenziati i 500 geni a più alto livello di espressione differenziale, mentre le linee blu rappresentano il limite della regione logFC ± 2.

8.3 DESeq

A differenza dei data set precedenti, nel design sperimentale adottato da Cumbie *et al.* i campioni corrispondono a repliche biologiche (capitolo 5.2, "Data set di Cumbie *et al.*"). Ciò permette di utilizzare DESeq nella sua configurazione predefinita, in cui la funzione *estimateDispersions* può essere invocata senza precisare parametri aggiuntivi. Tale configurazione viene di seguito indicata come *standard*.

In aggiunta all'impostazione predefinita, l'analisi differenziale è stata condotta anche specificando i parametri opzionali *sharingMode="fit-only"*, *fitType="locfit"*, mentre *lp_extra_args* è stato impostato sia a *nn=0.5* sia a *deg=1*. La Tabella 17 riporta il numero di

geni differenzialmente espressi, valutati con quattro valori di soglia per il p-value sulle tre diverse configurazioni di input per la funzione *estimateDispersions*.

p-value	**N° geni DE**		
	standard	***lp_extra_args=list(deg=1)***	***lp_extra_args=list(nn=0.5)***
5%	1136	1525	1638
1%	531	695	771
0,05%	200	285	212
0,01%	130	207	311

Tabella 22. Numero di geni rilevati come differenzialmente espressi da DESeq, per i quattro valori di soglia di p-value (non corretto per test multipli) considerati, per le quattro modalità di DESeq testate.

In Figura 37 è riportato l'MA-plot fra la media dei count in common scale (*baseMean)* e il livello di espressione differenziale (*log2foldChange*), per ciascun gene compreso nel dataset.

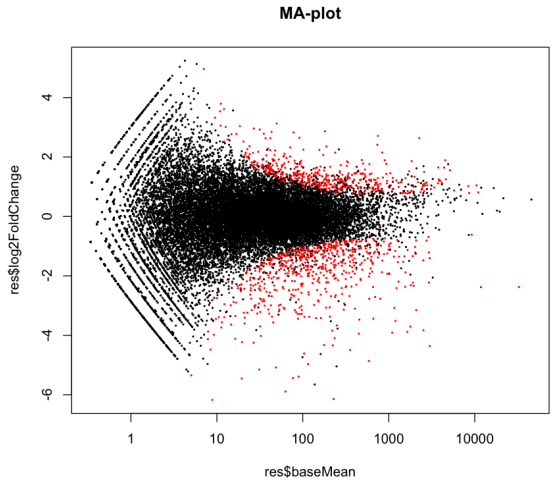


Figura 45. MA-plot di confronto fra livello di espressione (media dei count in common scale, *baseMean*) e log-fold change determinati per ciascun gene da DESeq (configurazione standard di *estimateDispersions*), con p-value≤5%. In rosso sono evidenziati i geni differenzialmente espressi.

8.4 PoissonSeq

La Tabella 18 riporta il numero di geni classificati come differenzialmente espressi da PoissonSeq, per i quattro p-value di soglia fissati.

p-value	**N° geni DE**
5%	1999
1%	549
0,05%	50
0,01%	10

Tabella 23. Numero di geni rilevati come differenzialmente espressi da PoissonSeq, per i quattro valori di soglia di p-value considerati

9 Riassunto analisi data set

I grafici di relazione media-varianza prodotti da edgeR hanno permesso di evidenziare come il modello di Poisson risulti generalmente poco adatto a descrivere i dati di RNA-Seq. La relazione media-varianza secondo il modello binomiale negativa permette un fit più corretto nei due data set provvisti di repliche tecniche (Figura 23 e Figura 32); nel data set di Cumbie *et al.*, caratterizzato da repliche biologiche, il modello di Poisson risulta del tutto inadeguato alla descrizione dei dati che presentano una varianza molto maggiore alla media (sovradispersione) (Figura 40).

L'analisi differenziale ottenuta da edgeR con stima della dispersione comune e genewise ha ottenuto selezioni molto simili sui data set provvisti di repliche tecniche, selezionando oltre il 99% di geni in comune (Figura 26 e Figura 35). Le differenze di selezione diventano più marcate nel data set con repliche biologiche, in cui i geni selezionati in comune scendono al 28% (Figura 43).

Il metodo DESeq ha dimostrato le sue limitazioni quando applicato a data set sprovvisti di repliche biologiche, selezionando un numero bassissimo di geni differenzialmente espressi, rispetto agli altri metodi applicati agli stessi dati. Il numero di geni selezionati aumenta solo se si considerano separatamente le repliche tecniche.

L'analisi delle curve ROC per i due data set provvisti di un gold standard non ha permesso di identificare un unico metodo come migliore in senso assoluto. Nel data set di Marioni *et al.* il metodo migliore è risultato edgeR (Figura 29), mentre sul data set MAQC i metodi testati presentano risultati molto simili tra di loro, con la curva ROC di DESeq lievemente al di sopra degli altri (Figura 38).

10 Conclusioni

L'analisi dell'espressione differenziale è una delle più interessanti applicazioni dell'RNA-Seq. Ad oggi, sono stati sviluppati molti metodi di analisi, diversi per modello statistico utilizzato per la descrizione dei dati e per il test statistico adottato per l'analisi dell'espressione differenziale. Si tratta di studi molto recenti, per i quali non esiste ancora una descrizione comune.
In questa tesi si è quindi organizzata l'informazione in un modello generale della distribuzione dei dati, uniformando le notazioni e la terminologia delle diverse descrizioni. In particolare, sono stati descritti i modelli statistici di Poisson e Binomiale Negativo.

Sono stati scelti quattro metodi di analisi differenziale che implementano i due modelli statistici sopra citati: DEGSeq, edgeR, DESeq e PoissonSeq. I quattro metodi sono stati confrontati, valutandone i risultati di classificazione su tre data set pubblici. Due dei data set (Marioni *et al.* e MAQC) sono provvisti di sole repliche tecniche, il terzo (Cumbie *et al.*) contiene invece repliche biologiche.

Il confronto ha permesso di evidenziare come, in tutti i data set, il modello di Poisson, che ipotizza una stringente eguaglianza tra media e varianza dei dati, sia poco adatto a descrivere i dati di RNA-Seq. L'inadeguatezza diventa evidente in presenza di repliche biologiche, che rendono i dati sovra-dispersi, cioè caratterizzati da una varianza molto maggiore della media. Su tutti i tre data set il modello Binomiale Negativo mostra una rappresentazione più realistica della relazione media-varianza presente nei dati.

Per due data set (Marioni *et al.* e MAQC) è stato identificato un gold standard, cioè una selezione dei geni differenzialmente espressi ottenuta con tecniche alternative a RNA-Seq e ritenuta un'affidabile rappresentazione della realtà. Dal confronto con i gold standard, sono stati calcolati due indici di confronto, precisione e sensitività, che hanno permesso di costruire le curve ROC per i diversi metodi. Dall'analisi delle curve ROC sul data set di Marioni, il metodo edgeR è risultato il migliore in termini di precisione e sensitività. I risultati relativi al data set MAQC sono invece molto simili tra i diversi metodi, con la curva ROC di DESeq lievemente sopra alle altre. Non è stato quindi possibile identificare un metodo che ottenesse prestazioni sensibilmente migliori rispetto agli altri su entrambi i data set.

Le principali limitazioni di questo studio sono rappresentate dal limitato numero di data set a cui i metodi sono stati applicati. Le curve ROC sono state determinate per due soli data set, non permettendo di indicare quale metodo sia il migliore in senso generale.
Per il data set provvisto di repliche biologiche non è stato possibile identificare un gold standard da usare come confronto della classificazione. A causa dei costi non trascurabili che realizzazione di esperimenti con repliche biologiche comporta, sono infatti rari dati di questo tipo disponibili in forma pubblica. Questa configurazione del design dell'esperimento sarebbe invece di grande interesse, permettendo di evidenziare l'effetto della variabilità biologica sull'analisi differenziale.

Oggetto di studi futuri potrà essere l'applicazione dei metodi su un numero più elevato di data set, possibilmente comprendenti repliche biologiche.
Un altro approccio interessante potrebbe essere quello di realizzare un data set di RNA-seq caratterizzato da un gold standard più affidabile, ottenuto sequenziando, assieme all'RNA endogeno in studio, molecole di RNA a sequenza e concentrazione note (*spike-in* RNA).

11 Bibliografia

[1] A Mortazavi, BA Williams, K McCue, L Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nat Methods*, July 2008.

[2] Ugrappa Nagalakshmi et al., "The transcriptional landscape of the yeast genome defined by RNA Sequencing," *Science*, vol. 320, June 2008.

[3] Zhong Wang, Mark Gerstein, and Snyder Michael, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, 2008.

[4] John C. Marioni, Christopher E. Mason, and Shrikant M. Mane, "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, 2008.

[5] F Sanger, Nicklen S, and Coulson AR, "DNA sequencing with chain terminating inhibitors," *Proc Natl Acad Sci U S A.*, Dec. 1977.

[6] Mahendar Thudi, Yupeng Li, Scott A. Jackson, Gregory D. May, and Rajeev K. Varshney, "Current state-of-art of sequencing technologies for plant genomics research," *Briefing in functional genomics*, vol. 2, no. 1, 2012.

[7] 454 Life Sciences - Roche. [Online]. http://www.454.com

[8] Applied Biosystems. [Online]. http://www.appliedbiosystems.com

[9] Illumina Inc. [Online]. http://www.illumina.com/

[10] Likun Wang, Zhixing Feng, Xi Wang, Xiaowo Wang, and Xuegong Zhang, "DEGseq: an R package for identifying differentially expressed genes from RNA-seq data," *Bioinformatics*, 2009.

[11] Jay Shendure and Hanlee Ji, "Next-generation DNA sequencing," *Nature Biotechnology*, vol. 26, no. 10, Oct. 2008.

[12] M Margulies, M Egholm, WE Altman, S Attiya, and JS, et al. Bader, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, Sep. 2005.

[13] OMICS Publishing Group. [Online]. http://omicsonline.org

[14] 454/Roche. GS FLX+ System. [Online]. http://my454.com/products/gs-flx-system/index.asp

[15] G. Turcatti, A. Romieu, M. Fedurco, and A.P. Tairi, "A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis," *Nucleic Acids Research*, 2008.

[16] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, and et al, "Sequence-specific error profile of Illumina sequencers," *Nucleic Acids Research Advance Access*, vol.

1, no. 13, May 2011.

[17] Jay Shendure et al., "Accurate multiplex polony sequencing of an evolved bacterial genome ," *Science*, Sep. 2005.

[18] Applied Biosystems. (2010) Applied Biosystems SOLiD 4 System - Templated Bead Preparation Guide. [Online]. http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_081748.pdf

[19] Applied Biosystems. 5500xl Genetic Analyzer. [Online]. http://products.invitrogen.com/ivgn/product/4460730

[20] Applied Biosystems. 5500 Series Genetic Analysis Systems. [Online]. http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html

[21] R Treffer and V Deckert, "Recent advances in single-molecule sequencing," *Curr Opin Biotechnol*, Feb. 2010.

[22] Hui Jiang and Wing Hung Wong, "Statistical inferences for isoform expression in RNA-Seq," *Bioinformatics*, 2009.

[23] Jun Li, Hiu Jiang, and Wing Hung Wong, "Modeling non-uniformity in short-read rates in RNA-Seq data," *Genome Biology*, 2010.

[24] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, 2009.

[25] Ugrappa Nagalakshmi, "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing," *Science*, vol. 320, no. 1344, 2008.

[26] D. M. Robinson, J. D. McCarty, and K. G. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, 2009.

[27] Likun Wang and Xi Wang. Bioconductor. [Online]. http://bioconductor.org/packages/2.6/bioc/html/DEGseq.html

[28] Mark D. Robinson and Gordon K. Smyth, "Moderated statistical tests for assessing differences in tag abundance," vol. 23, no. 21, pp. 2881-2887, 2007.

[29] Mark D. Robinson and Gordon K. Smyth, "Small-sample estimation of negative binomial dispersion, with applications to SAGE data," *Biostatistics*, pp. 321-332, 2008.

[30] DJ McCarthy, Yunshun Chen, and Gordon K. Smyth, "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation," 2011.

[31] Mark D. Robinson and Oshlack Alicia, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biology*, 2011.

[32] George Casella and Roger L. Berger, *Statistical Inference*.: Duxbury Press, 2001.

[33] D. R. Cox and N. Reid, "Parameter orthogonality and approximate conditional inference," *Journal of the Royal Statistical Society*, 1987.

[34] Gordon K. Smyth and Ari P. Verbyla, "A conditional approach to residual maximum likelihood estimation in generalized linear models," *Journal of the Royal Statistical Society*, 1996.

[35] Gordon K. Smyh, "Linear model and empirical Bayes method for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, 2004.

[36] Simon Anders and Wolfgang Huber, "Differential expression analysis for sequence count data," *Genome Biology*, 2010.

[37] Simon Anders, *Analysing RNA-Seq data with the DESeq package.*: European Molecular Biology Laboratory.

[38] Clive Loader, *Local Regression and Likelihood.*: Springer, 1999.

[39] Jun Li, Daniela M. Witten, Iain M. Johnstone, and Robert Tibshirani, "Normalization, testing, and false discovery rate estimation for RNA-sequencing data ," *Biostatistics*, 2011.

[40] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Natl Acad. Sci. USA*, 2001.

[41] John D. Storey and Robert Tibshirani, "Statistical significance for genomewide studies," *Proc Natl Acad Sci USA*, vol. 100, no. 16, pp. 9440-9445, 2003.

[42] John D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society*, vol. 64, no. 3, pp. 479-498, 2002.

[43] John C. Marioni, Christopher E. Mason, and Shrikant M. Mane. (2009) Genome Research. [Online]. http://genome.cshlp.org/content/suppl/2008/08/01/gr.079558.108.DC1/SupplementaryTable2.txt

[44] Affymetrix GeneChip Human Genome U133 Plus 2.0. [Online]. http://www.affymetrix.com/products_services/arrays/specific/hgu133plus.affx

[45] L Gautier, L, Cope, BM Bolstad, and RA Irizarry, "affy--analysis of Affymetrix GeneChip data at the probe level," *Bioinformatics*, Feb. 2004.

[46] P Flicek et al., "Ensembl 2008," *Nucleic Acids Research*, 2006.

[47] John C. Marioni, Christopher E. Mason, and Shrikant M. Mane. (2008) Genome Research. [Online]. http://genome.cshlp.org/content/suppl/2008/08/01/gr.079558.108.DC1/SupplementaryTable3.txt

[48] Jason S. Cumbie et al., "GENE-Counter: A Computational Pipeline for the Analysis of RNA-Seq Data for Gene Expression Differences ," *PLoS ONE*, vol. 6, no. 10, Oct. 2011.

[49] Yanming Di, Daniel W. Schafer, Jason S. Cumbie, and Jeff H. Chang, "The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq," *Statistical Applications in Genetics and Molecular Biology* , vol. 10, no. 1, 2011.

[50] JH Bullard, E Purdom, KD Hansen, and S Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments," *Bioinformatics*, vol. 11, no. 94, Feb. 2010.

[51] NCBI. SRA. [Online]. http://www.ncbi.nlm.nih.gov/sra

[52] NCBI. Gene Expression Omnibus. [Online]. http://www.ncbi.nlm.nih.gov/geo/

[53] Simon Anders, *Analysing RNA-Seq data with the DESeq package.*: European Molecular Biology Laboratory, 2011.

[54] Catherine Loader, Package 'locfit', 2010.

[55] Hairi Li et al., "Determination of tag density required for digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model," *Proceedings of the National Academy of Sciences of the USA*, 2008.

[56] Hairi Li et al. (2008) pnas_expression.txt. [Online]. https://sites.google.com/site/davismcc/useful-documents/pnas_expression.txt?attredirects=0&d=1