Analisi dell’espressione differenziale in linee cellulari affette da cancro al seno dopo la sovraespressione di diversi microRNA

Introduzione

I microRNA, secondo la Treccani, sono “una famiglia di piccoli RNA non codificanti che regolano l’espressione genica in maniera sequenza-specifica. Sono lunghi da 21 a 25 nucleotidi e sono a singolo filamento. La regolazione dell’espressione da parte dei microRNA (miRNA) avviene generalmente dopo la trascrizione genica ma sembra che alcuni di loro siano anche in grado di influire sulla trascrizione. Nel genoma umano, i geni per i miRNA possono trovarsi nelle regioni intergeniche, introniche intrageniche o esoniche intrageniche.”.

Differenze nella loro normale espressione sono correlate a varie patologie, tra cui il cancro. Questo progetto consiste nella realizzazione dell’analisi dell’espressione differenziale in due linee cellulari diverse: MCF-7 in seguito alla sovraespressione del miR-455-3p, miR-874, miR-592 e miR-9; MDA-MB-231 in seguito alla sovraespressione del miR-217 e miR-339. Le condizioni sperimentali sono correlate al cancro al seno. In seguito ad una descrizione della tecnica utilizzata, saranno presentati i risultati.

Analisi dell’espressione differenziale

L’analisi dell’espressione differenziale consiste nell’identificazione dei geni che presentano un livello di espressione diverso in campioni sottoposti a condizioni sperimentali diverse. Per definire un gene differenzialmente espresso si tiene conto se la sua espressione genica si discosta dalla condizione di uguale espressione nei due stati in modo significativo, confrontando, ad esempio, con un valore soglia per definire se sono sovraespressi o sottoespressi rispetto a tale valore di espressione.

I trascrittomi analizzati (cioè la totalità degli RNA trascritti a partire da un genoma) sono stati ottenuti a partire dall’RNA-seq, una tecnica per l’analisi del trascrittoma basata sulle tecnologie di Next-Generation Sequencing, come DNBSEQ-T7 e ILLUMINA. Le tecniche di Next-Generation Sequencing si caratterizzano per la produzione di reads, cioè piccole sequenze di DNA che, in caso di sequenziamento reference-based, dovranno essere allineate al genoma di riferimento.

Pipeline di analisi

La pipeline di analisi eseguita per la realizzazione di questo progetto inizia con la quantificazione dei trascritti. Infatti, con le tecniche di Next-Generation Sequencing, l'espressione genica viene calcolata in base ai count, cioè al numero di reads mappate sui geni del trascrittoma di riferimento. Per la creazione dell'indice del trascrittoma e la quantificazione dei trascritti, è stato utilizzato il software *salmon*.
Per la lettura dei risultati del conteggio effettuato da *salmon* è stato utilizzato il pacchetto *tximport.*
Ottenute le conte, dette "conte grezze", devono essere normalizzate per rimuovere errori tecnici nel sequenziamento, come la diversa profondità del sequenziamento e la diversa lunghezza dei trascritti analizzati. I trascritti individuati sono caratterizzati da due valori in particolare: il fold change, il cui segno determina la sovra o sotto-espressione del trascritto e un p-value, che indica quanto il fold change è significativo in base ad un test statistico che indica quanto le medie dei valori misurati nei diversi campioni sono significativamente diverse. Per limitare il numero di falsi positivi, questo valore viene corretto. Ci sono vari metodi disponibili, ma quello di default utilizzato da *limma,* il software utilizzato in questo progetto per l'analisi dell'espressione differenziale, è il metodo di Benjamini-Horchberg.
Limma utilizza un modello lineare per eseguire l'analisi dei geni differenzialmente espressi e nello specifico restituisce una tabella con i seguenti valori per ogni trascritto:

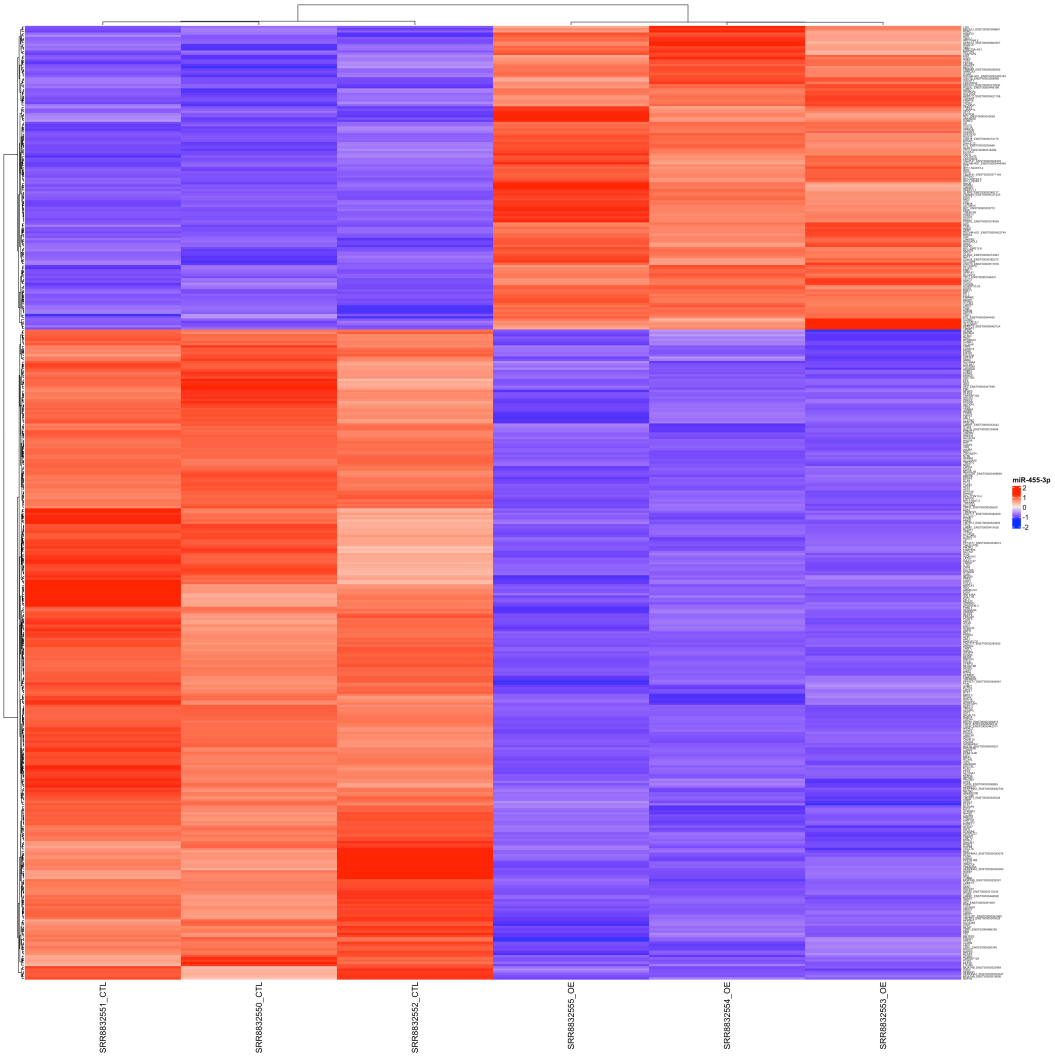
- LogFC: log fold change tra casi e controlli;
- AveExpr: il livello medio di espressione in tutti i campioni;
- t: il risultato del t-test per valutare quanto è grande l'espressione differenziale tra le due condizioni rispetto alla variabilità
- P.Value: il p-value per l'espressione differenziale
- Adj.P.Value: il p-value corretto
- B: il risultato della statistica bayesiana che indica la probabilità che il gene sia realmente differenziato

Dettagli implementativi

Per il download dei campioni dal database Gene di NCBI è stato utilizzato *sra-tools*.
Per la visualizzazione dei dati è stato usato *ComplexHeatmap* per i risultati relativi a miR-455-3p, miR-874 e miR-9; *EnhancedVolcano* per i risultati relativi a miR-592, miR-217 e miR-339. Per quanto riguarda i risultati relativi a miR-455-3p, miR-874 e miR-9, i geni visualizzati nella heatmap sono stati filtrati con quelli per cui il database EnsDb.Hsapiens.v86 ha restituito degli ID e che hanno superato le soglie di Fold Change e P.value in scala logaritmica. Trascritti diversi che si riferiscono allo stesso gene sono stati mantenuti. Le tabelle contenenti i dati sull'espressione genica di ogni trascritto sono disponibili nel relativo repository GitHub.
Di seguito una tabella con i dettagli dei campioni:

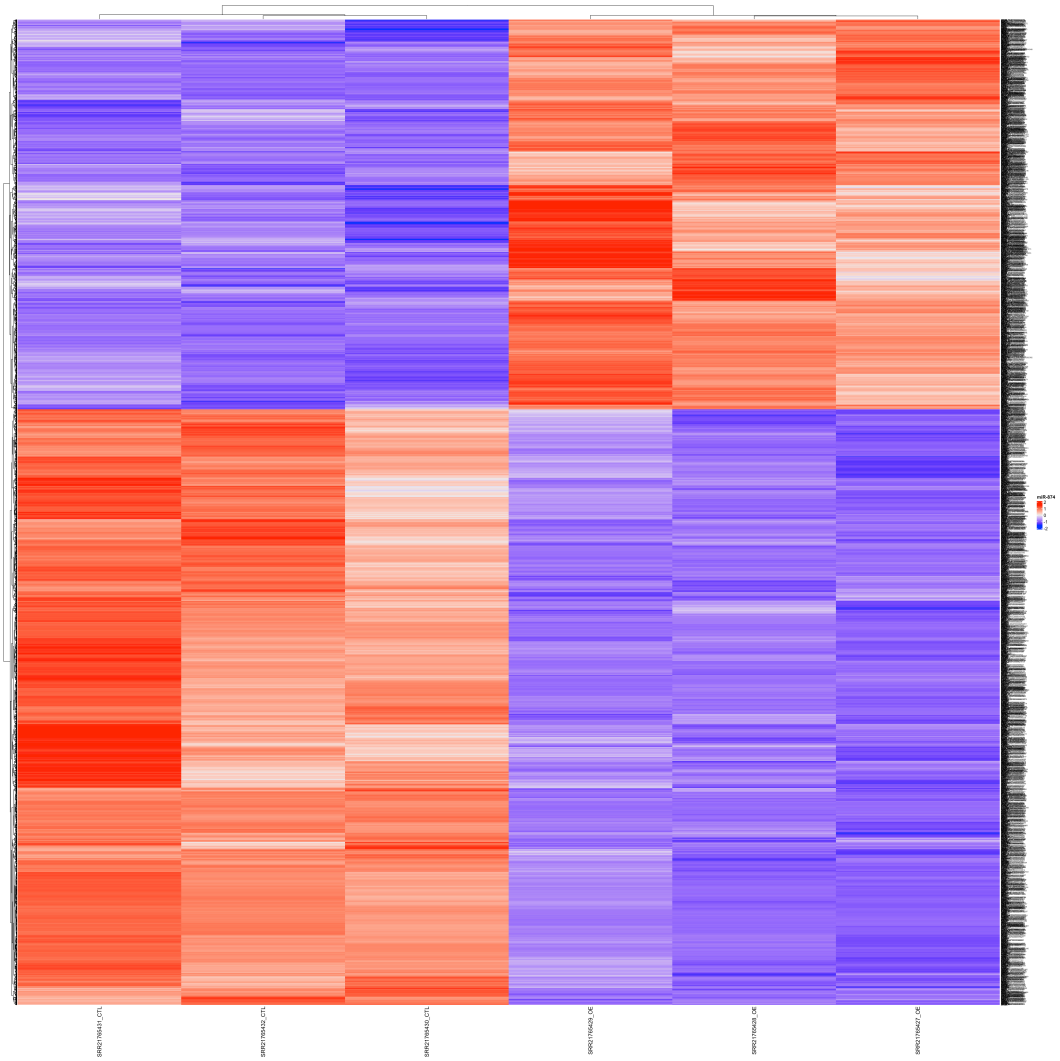
SRA	microRNA	Linea Cellulare	Piattaforma	Tipo
SRR8832553	miR-455-3p	MCF-7	Illumina HiSeq 2000	Overexpressed
SRR8832554	miR-455-3p	MCF-7	Illumina HiSeq 2000	Overexpressed
SRR8832555	miR-455-3p	MCF-7	Illumina HiSeq 2000	Overexpressed
SRR8832550	miR-455-3p	MCF-7	Illumina HiSeq 2000	Control
SRR8832551	miR-455-3p	MCF-7	Illumina HiSeq 2000	Control
SRR8832552	miR-455-3p	MCF-7	Illumina HiSeq 2000	Control
SRR21765427	miR-874	MCF-7	Illumina NextSeq 500	Overexpressed
SRR21765428	miR-874	MCF-7	Illumina NextSeq 500	Overexpressed
SRR21765429	miR-874	MCF-7	Illumina NextSeq 500	Overexpressed
SRR21765430	miR-874	MCF-7	Illumina NextSeq 500	Control
SRR21765431	miR-874	MCF-7	Illumina NextSeq 500	Control
SRR21765432	miR-874	MCF-7	Illumina NextSeq 500	Control
SRR15913265	miR-592	MCF-7	Illumina NovaSeq 6000	Overexpressed
SRR15913266	miR-592	MCF-7	Illumina NovaSeq 6000	Overexpressed
SRR15913263	miR-592	MCF-7	Illumina NovaSeq 6000	Control
SRR15913264	miR-592	MCF-7	Illumina NovaSeq 6000	Control
SRR18218736	miR-9	MCF-7	Illumina HiSeq 3000	Overexpressed
SRR18218737	miR-9	MCF-7	Illumina HiSeq 3000	Overexpressed
SRR18218734	miR-9	MCF-7	Illumina HiSeq 3000	Control
SRR18218735	miR-9	MCF-7	Illumina HiSeq 3000	Control
SRR24910526	miR-217	MDA-MB-231	DNBSEQ-T7	Overexpressed
SRR24910527	miR-217	MDA-MB-231	DNBSEQ-T7	Overexpressed
SRR24910528	miR-217	MDA-MB-231	DNBSEQ-T7	Control
SRR24910529	miR-217	MDA-MB-231	DNBSEQ-T7	Control
SRR14459543	miR-339	MDA-MB-231	Illumina NovaSeq 6000	Overexpressed
SRR14459544	miR-339	MDA-MB-231	Illumina NovaSeq 6000	Overexpressed
SRR14459545	miR-339	MDA-MB-231	Illumina NovaSeq 6000	Overexpressed
SRR14459540	miR-339	MDA-MB-231	Illumina NovaSeq 6000	Control
SRR14459541	miR-339	MDA-MB-231	Illumina NovaSeq 6000	Control
SRR14459542	miR-339	MDA-MB-231	Illumina NovaSeq 6000	Control

Risultati



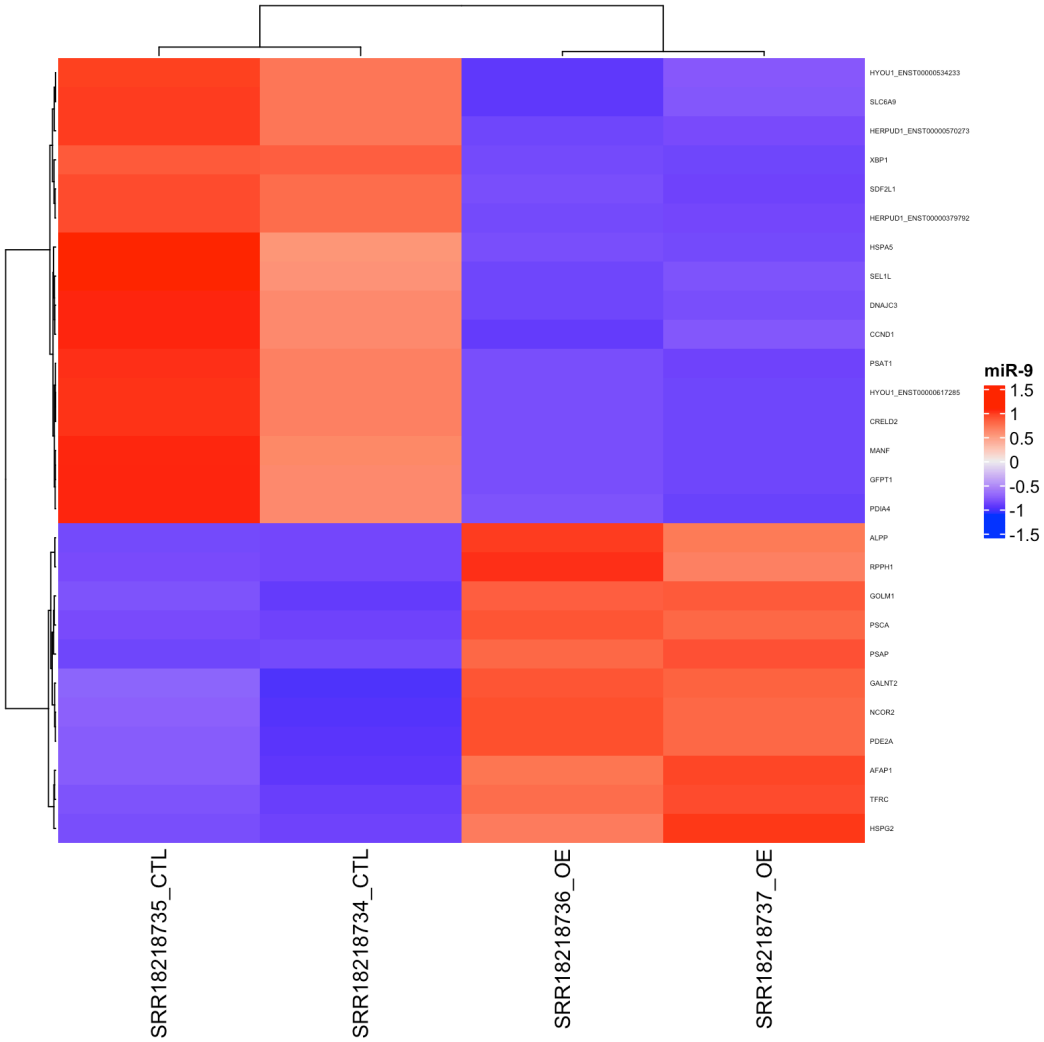
miR-455-3p

Come si può notare, la sovraespressione del miR-455-3p (campioni di destra) in MCF-7 reprime l'espressione della maggior parte dei geni, mentre causa la la sovraespressione di altri.



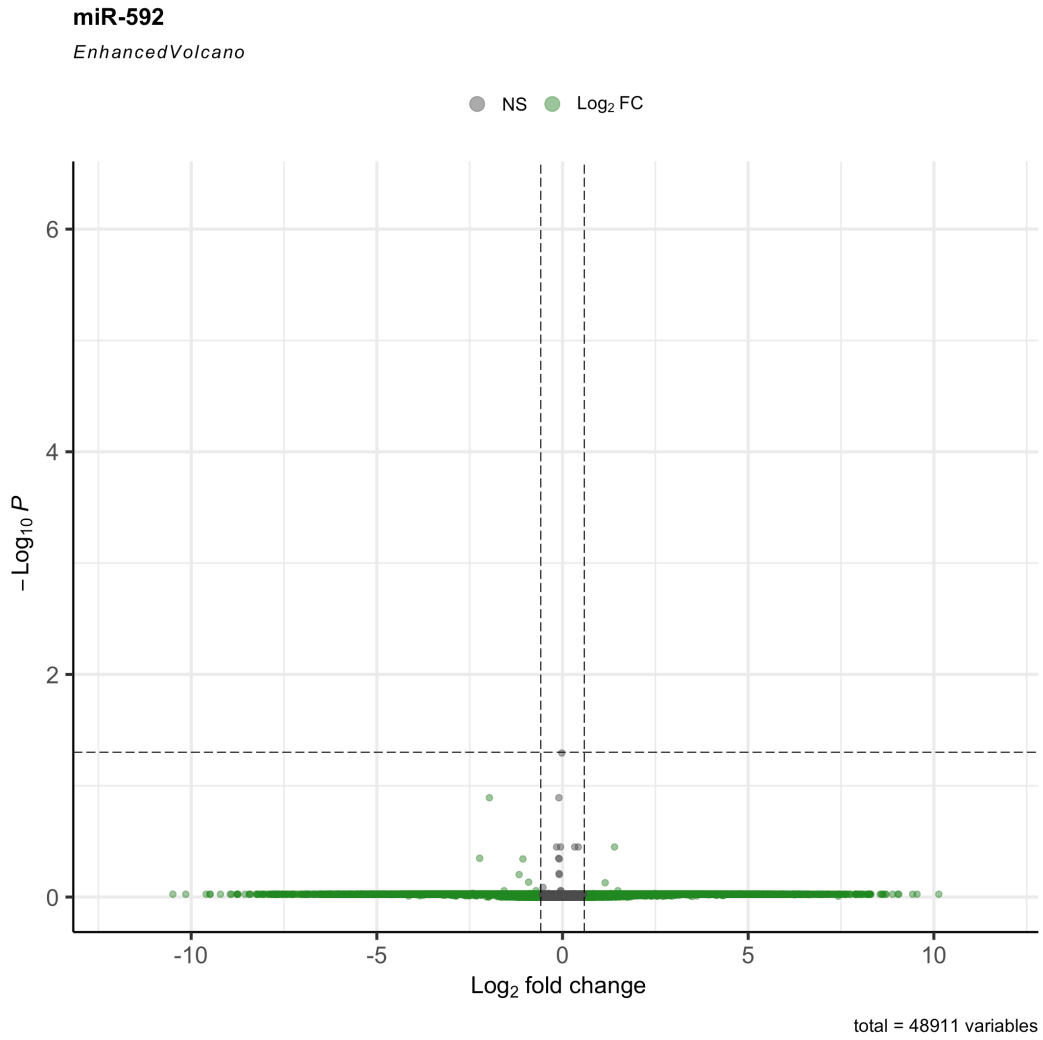
miR-874

Dato il grande numero di geni condizionati dalla sovraespressione del miR-874, le dimensioni di questo grafico sono piuttosto onerose (essendo i geni differenzialmente espressi individuati oltre 5000) e non si riescono bene a distinguere i nomi associati ad ogni riga, anche con dimensione dell'immagine 6000x6000 pixel. Ma nella metà di sinistra troviamo, come anche nelle altre heatmap i campioni di controllo e nella metà di destra i campioni dove il miR-874 è stato sovraespresso.



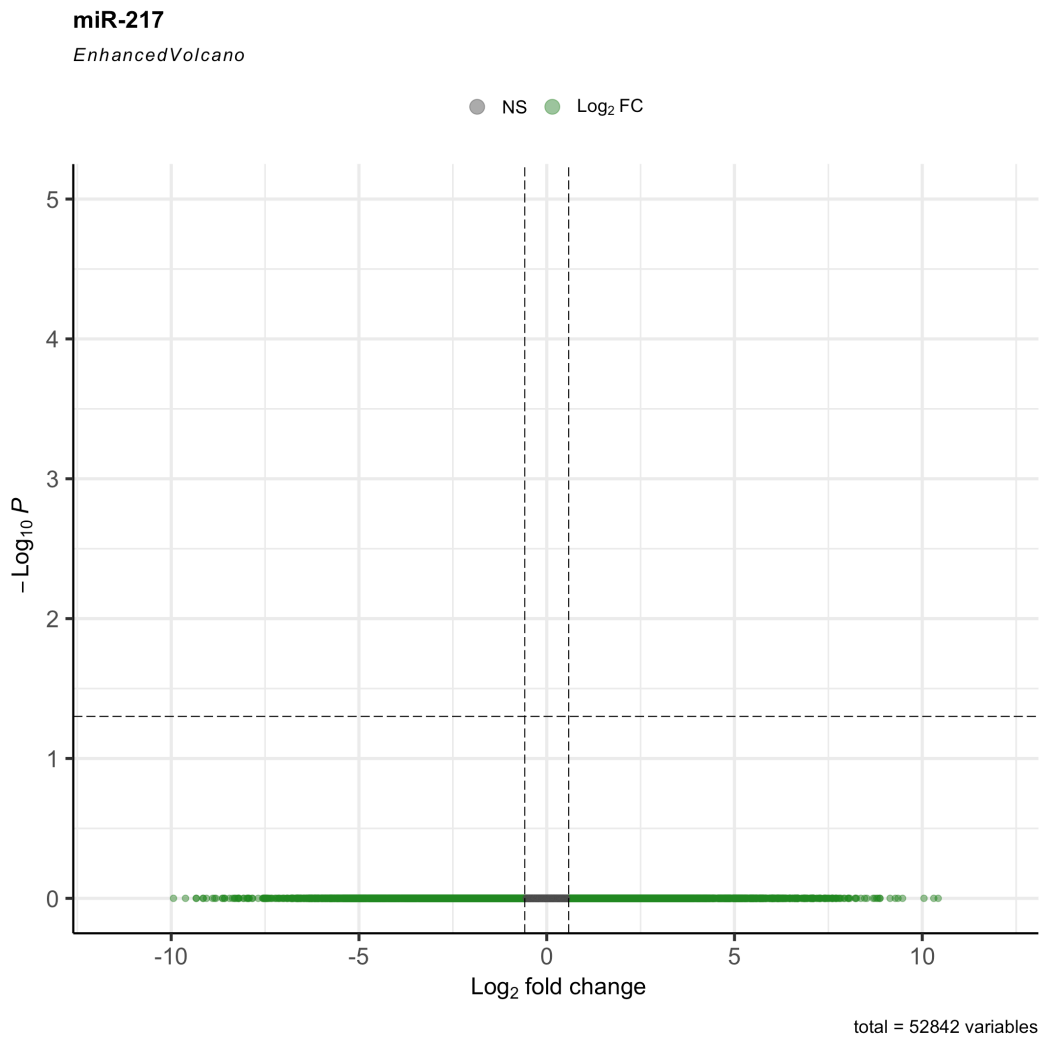
miR-9

In questo caso la sovraespressione del miR-9 ha provocato una un numero di geni differenzialmente espressi più contenuto.



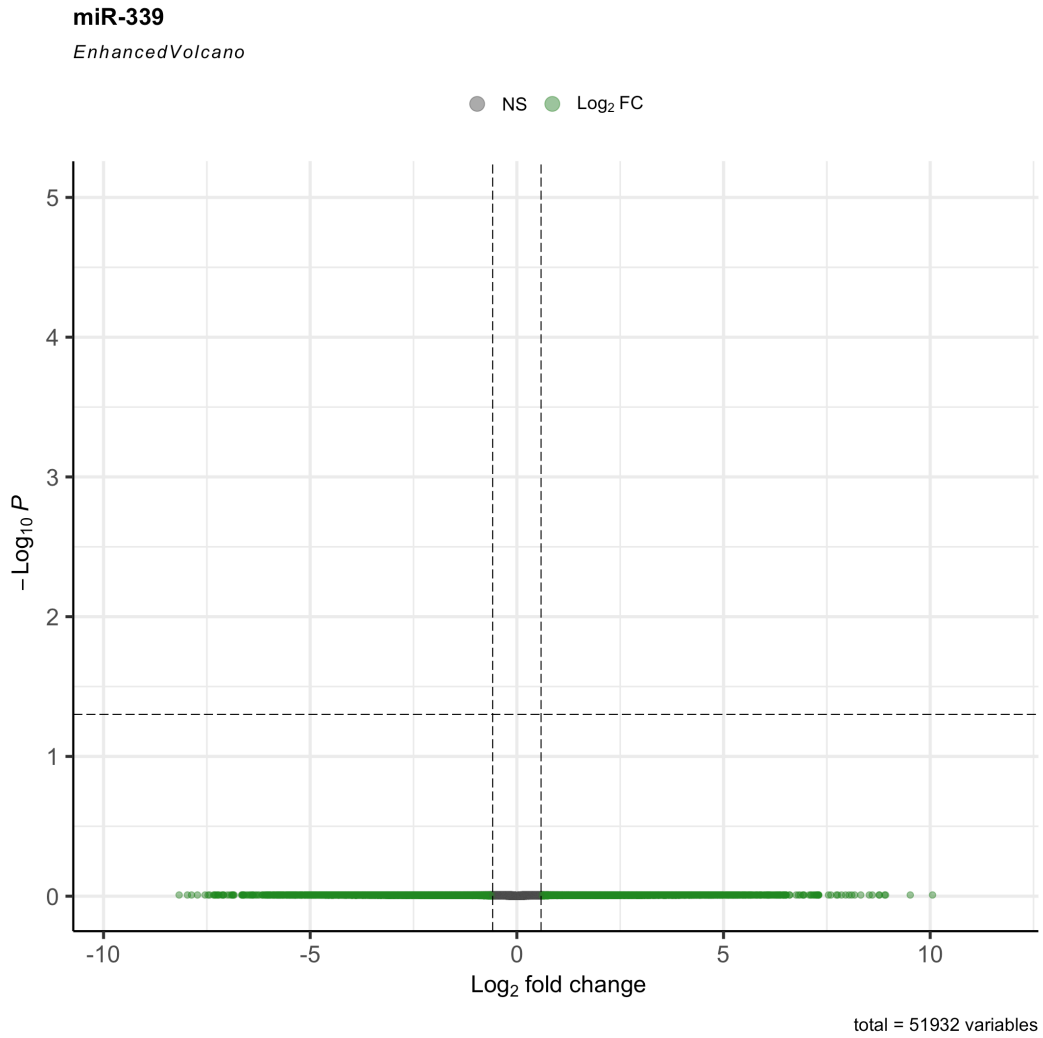
miR-592

Il miR-592 invece non ha prodotto alcun cambiamento significativo dell'espressione genica: nessun trascritto, come si nota dal VulcanoPlot, ha superato le soglie di Fold Change e di P-Value contemporaneamente.



miR-217

Per quanto riguarda il comportamento di entrambi i microRNA sovraespressi analizzati nella linea cellulare MDA-MB-231, miR-217 e miR-339, come si potrà verificare anche dal grafico successivo, i risultati ottenuti sono sono simili: nessun gene ha una differenza di espressione genica statisticamente significativa.



miR-339

Come si può osservare, i risultati sono simili a quelli ottenuti analizzando il miR-217.

Conclusioni

Il progetto consisteva nella nell'analisi dell'espressione differenziale in due linee cellulari diverse: MCF-7 in seguito alla sovraespressione del miR-455-3p, miR-874, miR-592 e miR-9 e MDA-MB-231 in seguito alla sovraespressione del miR-217 e miR-339 in condizioni sperimentali correlate al cancro al seno. I trascritti analizzati sono stati scaricati dal database *gene* di NCBI, ottenuti utilizzando tecniche di sequenziamento NGS, Next-Generation Sequencing.
L'analisi condotta ha portato alla realizzazione di diversi grafici del tipo HeatMap nel caso in cui si fosse registrata la presenza di geni differenzialmente espressi, del tipo VulcanoPlot nel caso in cui non fosse individuato alcun gene significativamente differenzialmente espresso. I risultati ottenuti sono stati molteplici. Per quanto riguarda i miR-455-3p, miR-874, e miR-9, la loro sovraespressione ha causato effettivamente una differenza di espressione significativa in diversi geni, individuandone un grande numero nell'analisi dei campioni relativi a miR-874, oltre 5000; viceversa un piccolo numero nei campioni relativi a miR-9. I risultati ottenuti analizzando gli effetti della sovraespressione dei miR-592, miR-217 e miR-339 registrano invece l'assenza di geni sovraespressi o sottoespressi ad un livello tale da avere rilevanza da un punto di vista statistico.
Sebbene questo studio si sia concentrato sull'analisi dell'espressione differenziale, i risultati ottenuti pongono le basi per future indagini funzionali. Studi futuri potrebbero quindi concentrarsi sulla validazione di questi risultati attraverso esperimenti in vitro e in vivo, al fine di valutare il potenziale terapeutico.

Il download dei dati completi in formato csv è disponibile al seguente repository GitHub: https://github.com/Hacstyle/Progetto-Bioinformatica