UNIVERSITÀ POLITECNICA DELLE MARCHE

DIPARTIMENTO DI SCIENZE DELLA VITA E DELL'AMBIENTE

Corso di Laurea Magistrale in

Biologia Molecolare e Applicata

ANALISI DI TRASCRITTOMI A SINGOLA CELLULA DA BIOPSIE DI TUMORE AL SENO

SINGLE-CELL TRANSCRIPTOME ANALYSIS FORM BREAST CANCER BIOPSIES

Tesi di Laurea Magistrale di
Francesca D'Angelo

Relatore Chiar.mo Prof.
Francesco Piva

Correlatore
Prof. Matteo Giulietti

Sessione di Febbraio
Anno Accademico
2022/2023

1. INTRODUZIONE

1.1 Dati e struttura del tessuto mammario

Stando all'ultimo report dell'OMS del febbraio 2022, sono più di 2,3 milioni i casi di tumore al seno che si verificano ogni anno nel mondo.

Il cancro alla mammella è la forma più comune di cancro tra le donne a livello globale in tutte le fasce di età: circa 1 donna su 8 nella sua vita viene colpita da questo tumore, sebbene la fascia maggiormente a rischio vada dai 50 ai 69 anni.

Il seno racchiude una varietà di strutture anatomicamente distinte, situate sulla parte anteriore del torace. In particolare, si ritrovano le ghiandole mammarie coinvolte nella secrezione e nell'eiezione del latte, circondate da tessuto connettivo, tessuto adiposo, nervi, vasi sanguigni e linfatici [1].

Il tessuto ghiandolare è costituito da circa 15-20 lobi, ognuno composto da strutture sacciformi chiamate lobuli che, durante l'allattamento, producono il latte. Dai lobuli il latte procede attraverso dei canali chiamati dotti che convergono nel capezzolo. Questa struttura prende il nome unità lobulare del condotto terminale (TDLU). (Figura1)

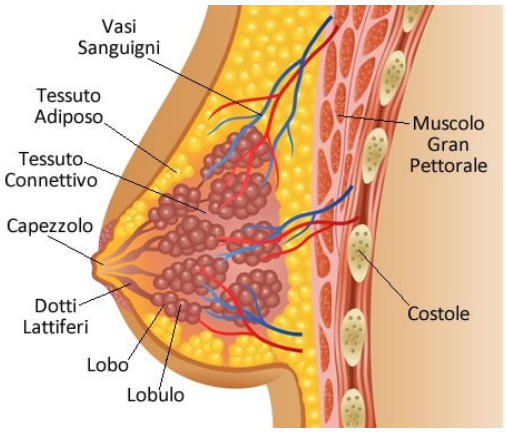


Figura 1 – Anatomia del seno

La ghiandola mammaria, sebbene sia presente in entrambi i sessi, è maggiormente sviluppata nelle donne [2]. Le ghiandole mammarie sono finemente regolate dal sistema endocrino attraverso la secrezione di estrogeni e progesterone.

Durante la pubertà le ovaie iniziano a produrre ormoni femminili che stimolano lo sviluppo del seno: i dotti si allungano e diventano più ramificati fino a formare un sistema vero e proprio.

Nonostante il seno sia anatomicamente maturo dopo la pubertà, il tessuto mammario rimane inattivo fino alla gravidanza. Solo in questa fase, infatti, si completa la formazione dei lobuli che poi, durante l'allattamento produrranno il latte.

Dal punto di vista istologico, il tessuto ghiandolare è costituito da uno strato interno di cellule epiteliali che riveste il lume e uno strato esterno di cellule mioepiteliali che ricopre la membrana basale (Figura 2).

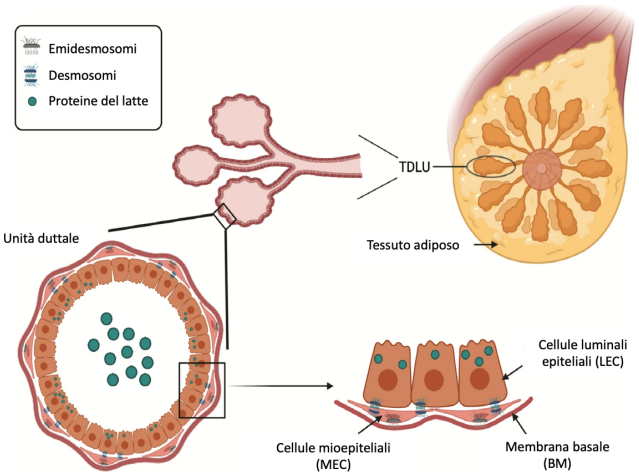


Figura 2 – Istologia del seno

La membrana basale separa le cellule mioepiteliali dallo stroma e permette l'importante differenziazione tra i tumori in situ e i tumori invasivi: i primi sono confinati all'interno della membrana basale, mentre i secondi la oltrepassano.

La popolazione delle cellule mioepiteliali è utile ai fini diagnostici: nel carcinoma mammario invasivo tali cellule sono assenti. La loro identificazione in una lesione mammaria significa che la lesione è sorta in una struttura ghiandolare preesistente. Le cellule mioepiteliali sono facilmente identificabili mediante immunocolorazione con anticorpi contro actina muscolare liscia, miosina ad alto peso molecolare, p63, citocheratina 5-6, ecc. [3]

1.2 Fattori di rischio e prevenzione

Sebbene il tumore alla mammella abbia un'incidenza sempre maggiore negli ultimi decenni, il tasso di sopravvivenza, dopo 5 anni dalla diagnosi, è in crescita ed è superiore all'80% grazie alla tempestiva rilevazione della malattia: una diagnosi precoce, insieme ad accurate misure preventive possono fare la differenza nella storia del tumore.

La mammografia, raccomandata dopo i 40 anni, rappresenta il gold standard tra gli esami di screening ad oggi utilizzati. Insieme a questa, anche l'auto-palpazione del seno, la risonanza magnetica e altri esami clinici sono strumenti a disposizione di ogni donna per monitorare la salute del proprio seno.

Parallelamente, uno stile di vita sano, caratterizzato da una dieta equilibrata, un esercizio fisico regolare e dal mantenimento di un peso corporeo ottimale, possono ridurre il rischio di sviluppare il cancro al seno.

Anche l'allattamento al seno, oltre a fornire innumerevoli benefici per la madre e il bambino, è associato a un rischio leggermente inferiore di sviluppare il tumore, diventando un passo concreto verso la prevenzione.

Tra i principali fattori di rischio imputabili al cancro alla mammella si possono citare: lo scorretto stile di vita, il sesso (le donne sono fortemente a rischio rispetto agli uomini), la storia genetica e familiare, l'invecchiamento, l'esposizione prolungata agli estrogeni (terapie

ormonali sostitutive) o a radiazioni al torace in giovane età (radioterapie) e la storia riproduttiva (donne nullipare o con gravidanze in età avanzata sono maggiormente esposte).

1.2.1 La genetica

La genetica può influenzare il percorso della malattia e, in alcuni casi, richiedere una sorveglianza più attenta o interventi preventivi mirati.

Nei documenti dell'*American Cancer Society* si legge che circa il 5-10% dei casi di cancro al seno è ereditario, derivano cioè, direttamente da mutazioni trasmesse da un genitore.

La causa più comune di cancro al seno ereditario è una mutazione nei geni *BRCA1* e *BRCA2* che nelle cellule normali sono implicati nella produzione di proteine coinvolte nel riparo del DNA danneggiato. Mutazioni di questi geni possono portare ad una crescita cellulare anormale e quindi al cancro.

Le donne con una mutazione del gene *BRCA1* o *BRCA2* hanno una maggiore suscettibilità al cancro al seno in età giovanile e in entrambi i seni, al cancro ovarico e ad alcuni altri tumori.

Altre mutazioni genetiche, meno comuni, che possono portare a tumori al seno ereditari coinvolgono i geni ATM, PALB2, TP53, CHEK2, PTEN, CDH1 e STK11.

1.3 Classificazione del tumore

Il tumore alla mammella presenta un'elevata eterogeneità sotto diversi punti di vista: anatomico, istologico, morfologico e molecolare, a cui corrispondono una prognosi ed una responsività ai trattamenti altrettanto disparati.

La classificazione del cancro al seno è un aspetto cruciale per prevedere il comportamento del tumore nel tempo, per determinare la strategia di trattamento più appropriata e per stabilire le prospettive a lungo termine per il paziente [4].

L’approccio tradizionale prende come riferimento il sistema TNM (Tumor, Node, Metastasis) dell'*American Joint Committee on Cancer (AJCC)* in cui T indica le dimensioni del tumore, N indica l'eventuale coinvolgimento dei linfonodi regionali e M indica la presenza o l'assenza di metastasi a distanza.

Questo sistema è utilizzato come linguaggio comune tra i centri di trattamento in tutto il mondo e per tutti i tipi di tumore, per guidare la pianificazione della terapia.

Tuttavia, con i progressi e il miglioramento delle tecnologie, l'osservazione di tassi di sopravvivenza molto diversi all'interno di tumori con lo stesso gruppo TNM, ha portato alla ricerca di soluzioni alternative e alla necessità di classificare il tumore alla mammella in sottotipi istologici e molecolari con il fine di facilitare la prognosi e la formulazione di trattamenti personalizzati.

1.3.1 Classificazione istologica

La classificazione istologica prende in considerazione il modello di crescita patologico e permette una distinzione tra:

- *carcinoma duttale in situ (DCIS)*, una forma precoce in cui il cancro si trova all'interno dei dotti e non ha ancora invaso il tessuto circostante,
- *carcinoma lobulare in situ (LCIS)* in cui le cellule cancerose sono circoscritte ai lobuli delle ghiandole mammarie,
- *carcinoma duttale infiltrante (IDC)* dove il cancro ha invaso il tessuto circostante attraverso i dotti,
- *carcinoma lobulare infiltrante (ILC)* nel caso in cui il cancro ha invaso il tessuto circostante attraverso i lobuli,

- altre forme meno comuni, con caratteristiche istologiche specifiche (carcinomi mucinosi, cribriforme, micropapillare, papillare, tubulare, midollare, metaplastico e apocrino).

L'IDC rappresenta circa il 70-80% di tutti i tumori invasivi, seguito dai carcinomi lobulari invasivi (ILC) (circa il 10% di tutti i tumori invasivi). IDC non presenta caratteristiche morfologiche specifiche ed esclusive, per questo viene anche detto IDC-NST "di nessun tipo speciale" e abbraccia la maggior parte dei tumori. Per tale ragione, questa classificazione maschera parte dell'ampia eterogeneità biologica e clinica dei tumori al seno; di conseguenza ha implicazioni prognostiche e predittive minime e la sua utilità clinica è piuttosto modesta [5].

1.3.2 Classificazione molecolare

La classificazione molecolare permette, di contro, una più attenta suddivisione dei tumori al seno in sottotipi tumorali sulla base dell'espressione di recettori ormonali e proteici. I recettori ormonali come l'estrogeno (ER) e il progesterone (PR), così come il recettore per il fattore di crescita umano epidermico 2 (HER2), possono influenzare le scelte terapeutiche, per tanto questa classificazione ha dimostrato avere un valore prognostico ed essere predittiva nella risposta alla chemioterapia.

Si distinguono, secondo questa classificazione, quattro sottotipi di tumore alla mammella: luminale A, luminale B, HER2 e triplo negativo.

Il sottotipo *luminale A* è il più comune e presenta un ritmo di crescita più lento rispetto agli altri subtipi molecolari di cancro al seno. Viene definito positivo per l'espressione dei recettori ormonali estrogeno (ER^+) e/o progesterone (PR^+). Questo tipo di tumore è un ottimo candidato per la terapia ormonale (o endocrina) che mira a ridurre l'effetto stimolante degli estrogeni/progesterone sulle cellule tumorali, al fine di bloccarne la proliferazione tumorale.

Il luminale A è anche descritto come HER2-negativo, difatti in questo caso l'espressione del gene HER2 non è aumentata.

Il sottotipo luminale B presenta cellule con una rapida velocità di crescita, è considerato più aggressivo del luminale A, è positivo per i recettori ormonali e presenta una sovraespressione di HER2.

Il sottotipo HER2 è ER e PR negativo, non mostra cioè una sovraespressione di questi recettori; al contrario è caratterizzato da una sovraespressione del gene HER2 che porta ad un'iperproliferazione delle cellule tumorali. HER2 è una proteina normalmente presente sulla superficie delle cellule, la cui funzione principale è quella di regolare la crescita e la divisione cellulare. Tuttavia, la sovraespressione o l'amplificazione del gene HER2 può portare ad un'eccessiva attività di segnalazione, che a sua volta può contribuire alla tumorogenesi.

Questo tipo di tumore può essere trattato con anticorpi monoclonali, come il Trastuzumab, che agiscono bloccando l'attività di HER2 o interferendo con la sua segnalazione, in modo da rallentare la crescita delle cellule tumorali HER2-positive.

È importante sottolineare che la presenza di HER2 può influenzare le opzioni di trattamento e la prognosi in determinati tipi di tumori. Pertanto, l'analisi dello status di HER2 è spesso parte integrante della valutazione del cancro al seno e di altri tumori nei quali è coinvolto.

In fine, nel sottotipo triplo negativo (TNBC) le cellule non presentano recettori né per estrogeni, né per progesterone, né per HER2.

Questo tipo di cancro è solitamente invasivo e le sue caratteristiche molecolari lo rendono resistente alla terapia ormonale e ai farmaci contro HER2 come il Trastuzumab; pertanto, è necessario ricorrere ad altre strategia terapeutiche come la chemioterapia, la radioterapia e la terapia mirata non HER2 [6].

1.4. Single cell RNA-seq

Le classificazioni appena trattate hanno tracciato le grandi linee dei sottotipi tumorali ma la strada rivoluzionaria potrebbe essere segnata da un potente strumento di indagine che permette l'analisi dettagliata dei trascritti di RNA a livello di singola cellula: la tecnica del Single Cell RNA-sequencing (scRNA-seq). Questa tecnica di Next Generation Sequencing (NGS) consente di analizzare il profilo genico di singole cellule, fornendo una visione approfondita delle differenze nell'espressione genica tra le cellule all'interno di un tessuto, offrendo una nuova prospettiva per svelare importanti dettagli genetici che sfuggirebbero alle analisi tradizionali.

È bene ricordare che quasi tutte le cellule del corpo sono costituite dallo stesso materiale genetico; ciò che le differenzia e le rende uniche sono le informazioni sul trascrittoma presenti in ogni cellula, che riflettono l'attività esclusiva di un solo sottoinsieme di geni [7]. Seppure le tecnologie di sequenziamento di nuova generazione abbiano rappresentato un grande progresso nel campo della biologia molecolare e nella pratica clinica, permettendo l'analisi del trascrittoma in modo molto approfondito, queste presentavano dei limiti circa l'unicità cellulare dei trascritti [8].

Nel 2009, Tang et. al hanno segnato una svolta fondamentale nel sequenziamento dell'RNA a singola cellula, eseguendo il primo protocollo per l'esecuzione di scRNA-seq [9].

Lo sviluppo di questa tecnologia ha aperto la strada ad una migliore comprensione della diversità cellulare e delle dinamiche dell'espressione genica a livello di singola cellula in vari contesti biologici e patologici, anche e soprattutto nel campo dell'oncologia. La tecnica ha permesso, infatti, di scrutare il cuore molecolare di ciascuna cellula tumorale, discernendo variazioni sottili e individualità genetica che potrebbero determinare la risposta al trattamento e una terapia personalizzata.

1.5 Fasi del scRNA-seq

Sono innumerevoli i protocolli messi a punto per il sequenziamento dell'RNA a singola cellula ma tutti seguono essenzialmente quattro fasi fondamentali che prevedono: (a) la dissociazione del tessuto di partenza, (b) l'isolamento e la lisi delle singole cellule, (c) la retrotrascrizione delle molecole di mRNA isolate in DNA complementare (cDNA), l'amplificazione del cDNA e la preparazione delle librerie e per ultimo, ma non per importanza, (d) l'analisi dei dati.

1.5.1. Dissociazione

La dissociazione del tessuto può essere ottenuta attraverso trattamenti meccanici ed enzimatici con l'utilizzo di tripsina, collagenasi e papaina. Questi processi possono indurre stress e conseguenti modificazioni trascrizionali nella cellula che devono essere attenzionati e quanto più possibile minimizzati.

È stato visto che sia il processo di dissociazione di per sé che la temperatura di 37°C a cui viene condotto, potrebbero indurre l'espressione di geni dello stress che si traduce in cambiamenti artificiali a livello dei trascritti. È stato quindi suggerito di operare ad una temperatura di 4°C in modo da ridurre ipotetici risultati imprecisi [7].

1.5.2 Isolamento e lisi

Le cellule dissociate vengono trasformate in una sospensione cellulare in modo che possano essere trattate e manipolate individualmente. Possono essere impiegati approcci diversi per l'isolamento delle cellule, distinguibili in approcci automatizzati e approcci manuali.

I primi sono ad alto rendimento e i più utilizzati includono la *FACS* o le *tecniche microfluidiche*. I secondi, applicabili solo in contesti con un numero di cellule limitato, sono la *micromanipolazione meccanica* (micropipettaggio) o la *microdissezione a cattura laser* (LCM).

FACS è l'acronimo di "Fluorescence-Activated Cell Sorting" (o Fluorocitometria a Cellule Attivate da Fluorescenza). È una tecnica che combina la citometria a flusso con la capacità di isolare specifiche popolazioni cellulari in base alle caratteristiche fluorescenti.

Le cellule vengono marcate con anticorpi specifici o coloranti fluorescenti che si legano a precisi marcatori cellulari. Nel caso in cui si sta cercando una popolazione cellulare che esprime un determinato antigene, le cellule verranno marcate con un anticorpo fluorescente mirato per l'antigene di interesse.

Le cellule marcate, in una soluzione liquida e con un flusso costante, sono fatte passare attraverso uno stretto canale in cui vengono esposte ad un raggio laser che eccita i fluorocromi legati, generando segnali fluorescenti. Questi sono individuati da rilevatori che permettono l'identificazione delle cellule con il pattern di colorazione desiderato.

Successivamente, le cellule sono raccolte in piastre con pozzetti contenenti il tampone di lisi, assegnando una cellula ad ogni pozzetto.

Si passa poi alla classificazione e separazione: sulla base delle caratteristiche fluorescenti rilevate, il sistema di FACS classifica le cellule in diverse popolazioni. Inoltre, può separare fisicamente le cellule marcate usando un getto d'aria o un campo elettrico, isolando così le popolazioni di interesse.

Con FACS si riesce ad isolare un numero elevato di cellule, nell'odine delle migliaia, con un alto rendimento e con costi contenuti. L'utilizzo degli anticorpi permette inoltre un arricchimento delle cellule di interesse.

I limiti legati a questa tecnica risiedono nell'elevato numero di cellule di partenza richieste, per cui non è applicabile a campioni con basso numero di cellule, nella contaminazione dell'ambiente circostante e nella potenziale presenza di pozzetti vuoti o con doppietti.

Le *tecniche microfluidiche* permettono la manipolazione e l'analisi di piccoli volumi di liquidi a livello microscopico. Si utilizzano dispositivi microfluidici, come i chip microfluidici, per isolare e catturare singole cellule. Questi chip sono progettati con strutture microscopiche che consentono la cattura e la separazione di singole cellule in microgoccioline.

L'isolamento può essere condotto anche manualmente attraverso una *micromanipolazione meccanica* (o micropipettaggio); dopo aver ottenuto una sospensione con le singole cellule, queste vengono aspirate per mezzo di una pipetta di vetro capillare con l'ausilio di un microscopio.

La micromanipolazione, sebbene sia una tecnica potente per l'isolamento di cellule specifiche, è limitata allo studio di popolazioni cellulari ridotte, richiede un elevato grado di abilità e precisione da parte dell'operatore, è molto dispendiosa sia in termini di tempo che di costi ed infine, espone le cellule a stress meccanico e conseguenti lesioni.

Un'altra tecnica non automatizzata è la *microdissezione a cattura laser* (LCM, Laser Capture Microdissection); questa consente di isolare specifiche cellule da frammenti di tessuto combinando l'uso di un microscopio con la precisione di un raggio laser per "tagliare" e catturare le cellule di interesse fissandole su una pellicola sottile. Quando il film viene sollevato, le cellule selezionate vi rimangono attaccate, mentre il tessuto circostante è sul vetrino. Il film è poi trasferito in una provetta per microcentrifuga.

Sebbene questa metodica abbia il vantaggio di preservare le relazioni spaziali tra le cellule e di non richiedere la dissociazione enzimatica, il tessuto di partenza deve essere sezionato molto sottilmente, comportando una potenziale perdita di materiale qualora il diametro cellulare sia maggiore dello spessore della sezione. Il metodo presenta un throughput

limitato, consentendo l'analisi di piccoli gruppi di cellule; inoltre, è laborioso, ha costi elevati e necessita di attrezzature specializzate.

1.5.3 Retrotrascrizione, amplificazione e costruzione librerie

Dopo aver isolato le singole cellule, queste sono sottoposte a lisi per rilasciare l'mRNA. L'mRNA catturato viene retrotrascritto in cDNA e in questa fase viene marcato con un identificatore unico per ogni cellula, denominato "barcode cellulare" (CB) e con un Identificatore Molecolare Unico (UMI) per ogni trascritto.

Le reazioni procedono in maniera parallela e automatizzata, rendendo il processo altamente produttivo utilizzando piccoli volumi di reagenti.

Il cDNA marcato viene amplificato tramite PCR per generare materiale sufficiente per la preparazione della libreria che verrà poi sottoposta ad un sequenziamento ad alto rendimento dal quale si ottengono delle reads in formato FASTQ.

Le sequenze sono quindi sottoposte ad un'attenta analisi bioinformatica: il processo di demultiplexing, mediante il quale le reads vengono separate in base ai barcode e agli UMI, permette di attribuire ciascuna sequenza alla cellula di origine, mentre il processo di mapping consente di mappare o allineare le reads con un genoma di riferimento; infine, le reads sono utilizzate per quantificare l'espressione genica, generando una matrice grezza di espressione che rappresenterà il profilo di espressione di ogni gene in ogni cellula.

1.5.4 Analisi dei dati

L'analisi dei dati di scRNA-seq è la parte critica del processo, poiché consente di estrarre informazioni biologiche significative dalla grande quantità di dati generati durante il sequenziamento.

Questa è articolata in tre livelli che includono: l'elaborazione dei dati grezzi e il controllo qualità (QC), l'analisi di base dei dati e l'analisi avanzata [10].

Elaborazione dei dati grezzi e il controllo qualità (QC)

Prima di iniziare l'analisi vera e propria, è importante eseguire un controllo di qualità sulla matrice di espressione ottenuta dal pre-processamento per eliminare l'elevata varianza tecnica che costituisce il "rumore" che deve essere necessariamente ridotto al fine di ottenere dati confrontabili e attendibili. Ogni singolo processo di scRNA-seq contribuisce a generare varianza: dalla cattura cellulare alla trascrizione inversa.

Alcuni degli indicatori di qualità possono essere il numero di geni, il numero di UMI, le percentuali di geni mitocondriali e ribosomiali in ciascuna cellula.

La presenza di doppietti o multipletti, ovvero la condizione in cui due o più cellule vengono erroneamente catturate insieme in una stessa gocciolina durante il processo di generazione delle librerie, può portare ad una sovrastima falsata dell'espressione genica, restituendo dati poco precisi.

Un'altra fonte comune di distorsione nei dati di scRNA-seq è il bias di amplificazione: si verifica durante il processo di amplificazione in PCR e può influire sull'accuratezza quantitativa delle misure di espressione genica.

Questo effetto può essere mitigato dall'uso di UMI: indicatori unici di ogni mRNA che permettono di quantificare il numero di trascritti per gene per cellula; ogni UMI rappresenta un singolo trascritto, per cui è facilmente deducibile che trascritti con UMI identici sono frutto dell'amplificazione e non di una reale trascrizione cellulare e per tanto devono essere eliminati.

Anche l'effetto batch, che si verifica quando i campioni sono elaborati in batch separati, con tempi e reagenti differenti o non dallo stesso operatore, può introdurre variabilità tecnica indesiderata [8].

La presenza di RNA mitocondriale può essere un indicatore di stress cellulare o di altri fattori tecnici. L'RNA mitocondriale spesso non è rilevante per l'analisi dell'espressione genica nucleare per cui nel filtraggio viene rimossa l'espressione dei geni mitocondriali.

Anche l'RNA ribosomiale non è di interesse quando si studia l'espressione genica; pertanto, anche questo può essere rimosso o filtrato per migliorare la specificità dell'analisi.

Non esiste uno standard assoluto che permette di giudicare una cellula di buona qualità e una che, al contrario, deve essere rimossa; programmi come Seurat (R), Scarlet (Python), scGEATool (Matlab) e Cellenics permettono l'analisi per controllo qualità dei dati scRNA-seq. Infine, la matrice di espressione viene normalizzata tenendo conto di differenze nei livelli di espressione tra le cellule, considerando la profondità di sequenziamento e il conteggio degli UMI.

Analisi di base

Si passa all'analisi dei dati vera e propria che prevede la feature selection, la riduzione della dimensionalità, il clustering e l'annotazione cellulare.

La feature selection permette di esaminare e prendere in considerazione i geni biologicamente rilevanti per l'analisi a valle, ovvero i geni altamente variabili (HVG) che facilitano la differenziazione tra i vari tipi cellulari.

Tipicamente il numero di HVG considerato è compreso tra 1000 e 5000, il che rende la dimensionalità dei dati molto elevata; per cui viene applicata una *riduzione della dimensionalità* che proietta le cellule da uno spazio ad alta dimensionalità in uno a bassa dimensionalità.

I metodi ampiamente utilizzati per tale scopo sono l'analisi delle componenti principali (PCA), il tSNE e UMAP.

Il clustering cellulare è un'analisi che mira a suddividere le cellule in gruppi omogenei in base a determinate caratteristiche, come il profilo di espressione genica, in modo da ridurre al minimo le differenze intracluster.

I cluster possono rappresentare sottopopolazioni cellulari con caratteristiche simili o con un ruolo biologico comune. Metodi di clustering includono l'algoritmi k-means, Louvian e "nearest-neighbor network for the cells" specifico per i dati di espressione genica [11].

Dopo aver ottenuto i vari cluster, è fondamentale assegnare il tipo cellulare ad ogni barcode attraverso il processo di *annotazione cellulare.* L'annotazione cellulare può avvenire con due metodi:

1) metodi manuali basati su geni marcatori specifici (sovraespressi o non espressi) in una data cellula; questi marcatori possono essere reperiti in specifici database (CellMarker, PanglaoDB, ecc.) oppure presenti in letteratura;

2) metodi computazionali di apprendimento automatico supervisionato o non, come SingleR, SciBet, SCINA, ecc. Anche questi strumenti fanno riferimento a database presenti in rete.

Analisi avanzata

L'analisi avanzata dei dati prende in considerazione: l'arricchimento funzionale dei geni, l'inferenza della traiettoria, le comunicazioni cellula-cellula, l'inferenza del regulone e la previsione dell'attività dei fattori di trascrizione (TF), la stima del flusso metabolico e il ciclo cellulare.

L'*analisi di arricchimento funzionale* ha come obiettivo quello di fornire una comprensione più approfondita delle funzioni biologiche che caratterizzano ciascun cluster cellulare.

Questa analisi aiuta a capire quali processi biologici, percorsi molecolari o categorie di geni sono significativamente "arricchiti" all'interno di ciascun cluster attraverso il confronto del profilo genico delle cellule con un set di geni annotati o categorie funzionali, permettendo

così di ottenere preziose informazioni sulla diversità cellulare e su eventuali perturbazioni o cambiamenti nei processi biologici in condizioni diverse o patologiche.

L'*analisi della traiettoria* permette di ordinare le cellule lungo una linea virtuale pseudo-temporale che riflette il loro presunto sviluppo temporale, al fine di esaminarne le traiettorie di sviluppo.

Sono utilizzati algoritmi specifici, come Monocle, che stimano una sequenza temporale basata sull'espressione genica.

Un metodo alternativo per catturare la dinamica dell'espressione genica è utilizzare la velocità dell'RNA, spostando l'attenzione sulla relazione tra trascritti maturi e non maturi.

La *comunicazione cellula-cellula* è un evento importante nello sviluppo e nell'omeostasi dell'organismo, nonché nella generazione e nella progressione della malattia.

Questa dipende essenzialmente dalle interazioni ligando-recettore (LR), che sono quantificate dalla co-espressione LR.

Grazie allo sviluppo di vari strumenti computazionali è possibile dedurre queste interazioni partendo da dati di scRNA-seq. Sono state allestite delle vere e proprie banche dati contenenti le interazioni ligando-recettore note che sono poi interrogate dai diversi programmi che elaborano l'analisi.

Il fine dell'analisi è quello di rivelare come le cellule comunicano tra loro attraverso segnali molecolari e di identificare sottopopolazioni cellulari che potrebbero essere coinvolte in specifiche vie di segnalazione.

La *previsione dell'attività di un fattore di trascrizione* (TF) è un altro strumento utilizzato nell'analisi dei dati di scRNA-seq e consiste nel valutare quanto il TF sia attivo in un particolare campione, sulla base dei livelli di espressione genica dei geni regolati da quel particolare TF.

Vengono identificati dei gruppi di geni regolati da un determinato fattore di trascrizione, che prendono il nome di reguloni.

Per ogni tipo cellulare è quindi possibile individuare reguloni specifici. Una risorsa importante per riconoscere i reguloni sono i database TF-target quali GIASPAR, TRRUST, KnockTF, ecc, che coprono la maggior parte dei fattori di trascrizione.

Sulla base di questi database vengono costruite reti di regolazione trascrizionale specifiche di ogni cluster, attraverso l'individuazione di TF sovraregolati e/o geni TF-bersaglio espressi in modo differenziale.

Infine, *l'analisi metabolica* e quella del *ciclo cellulare* permettono di avere un quadro ancora più completo sul comportamento dei vari tipi cellulari.

La prima consente di monitorare i cambiamenti nell'espressione genica di geni metabolici durante processi fisiologici/patologici.

La seconda ha come obiettivo quello di predire il ciclo cellulare di ogni singola cellula con il fine ultimo di evidenziare l'eventuale stato di divisione delle putative cellule tumorali.

Entrambi si avvalgono dell'utilizzo di strumenti computazionali all'avanguardia e algoritmi sofisticati specifici per il tipo di analisi.

1.6 CNV – Copy Number Variation

Le variazioni del numero di copie (CNV) sono alterazioni nel numero di copie di specifici segmenti di DNA in un genoma. Questi cambiamenti possono includere duplicazioni, delezioni o inversioni di tratti genomici. I CNV possono avere un ruolo significativo nello sviluppo di alcune malattie, compresi i tumori. In particolare, è noto che in alcuni tipi di tumore la sovraespressione di oncogeni può essere dovuta all'aumento del numero di copie

del tratto genico, così come la delezione può coinvolgere regioni codificanti oncosoppressori.

Esistono diversi programmi per inferire i CNV a partire da dati di scRNA-seq. Questi metodi computazionali possono predire le cellule tumorali a partire dalla sequenza fastaq e/o dalla matrice di espressione (InferCNV, SCEVAN, CopyKat, sciCNV, HoneyBADGER, ecc.).

2. SCOPO DELLA TESI

L’eterogeneità tumorale è un fatto certo ma quanto a fondo è stata investigata finora? Con l’obiettivo di scovare nuove differenze, è stata utilizzata la più alta risoluzione possibile all’interno di un frammento di biopsia tumorale, quello della singola cellula, sfruttando il metodo della scRNA-seq. Dopo aver reperito un set di dati di scRNA-seq provenienti da 26 biopsie di tumore al seno, questi sono stati analizzati dal punto di vista bioinformatico.

Mentre recuperare e scaricare i dati di scRNA-seq è un'operazione relativamente semplice, la vera complessità sorge nella successiva fase di elaborazione e interpretazione.

Per affrontare questa sfida, sono stati adottati diversi strumenti specializzati e metodologie avanzate.

L'obiettivo, per ciascun campione, ha visto come primo passo, la ricerca di uno o più strumenti adeguati a predire i tipi cellulari sulla base dell'espressione genica al fine di valutare le differenze tra i sottotipi identificati ed analizzarne le implicazioni biologiche associate.

Questo permette di evidenziare quanto un frammento di biopsia sia eterogeneo dal punto di vista cellulare, sottolineando l’importanza dell’analisi a livello di singola cellula per il sequenziamento, a discapito delle analisi classiche, definite “bulk” o di massa, che rappresentano una “media” di espressioni geniche delle varie cellule del tessuto di partenza.

Per permettere questo tipo di analisi è necessario standardizzare i dati di scRNA-seq per renderli confrontabili tra loro.

3. MATERIALI E METODI

3.1 Acquisizione dati

Prendendo come riferimento lo studio di Sunny Z. Wu et al *"A single-cell and spatially resolved atlas of human breast cancers"* [12] del 2021 sull'analisi trascrittomica a singola cellula del tumore alla mammella, è stato scaricato il relativo set di dati (GSE176079) da GEO (Gene Expression Omnibus) (https://www.ncbi.nlm.nih.gov/geo/), un database pubblico gestito dall'NCBI che archivia dati di espressione genica, dati di microarray e sequenze di RNA-Seq.

Nello studio è riportato il profilo di espressione genica a singola cellula (Chromium, 10X Genomics) di biopsie di 26 diversi pazienti con tumore al seno di cui 11 ER+, 5 HER2+ e 10 TNBC (Tabella 1) tra i quali ne sono stati selezionati 11 preservando quelli con dati di più alta qualità.

Il set di dati per ogni paziente si presenta sottoforma di 4 file: *barcodes.tvs* con all'interno l'elenco completo dei barcodes di ogni cellula, *genes.tsv* contenente i nomi di tutti i geni presi in considerazione nell'analisi, *matrix.mtx* e *metadata.csv* contenente l'annotazione cellulare per singolo barcode.

N° pz	Sesso	Età	Grado	Tipo	Sottotipo IHC	Trattamento	N° geni x cellule
1	Femmina	43	3	IDC	HER2+/ER+	Non trattato	29733 x 6178
2	Femmina	49	3	IDC	HER2+	Non trattato	29733 x 2353
3	Femmina	60	3	IDC	HER2+	Non trattato	29733 x 3024
4	Femmina	50	2	IDC	ER+	Non trattato	29733 x 631
5	Femmina	52	3	IDC	TNBC	Non trattato	29733 x 774
6	Femmina	82	3	IDC	ER+	Non trattato	29733 x 2327
7	Femmina	61	3	IDC	ER+	Trattato	29733 x 3527
8	Femmina	57	3	IDC	ER+	Non trattato	29733 x 2531
9	Femmina	41	2	IDC	HER2+/ER+	Trattato	29733 x 5309
10	Femmina	85	2	IDC	ER+	Non trattato	29733 x 3764
11	Femmina	88	2	IDC	ER+	Non trattato	29733 x 5789
12	Femmina	52	3	IDC	ER+	Trattato	29733 x 4451
13	Femmina	35	3	IDC	TNBC	Non trattato	29733 x 2131
14	Femmina	54	2	IDC	ER+	Non trattato	29733 x 631
15	Femmina	58	2	IDC	ER+	Non trattato	29733 x 1138
16	Femmina	54	3	IDC	TNBC	Non trattato	29733 x 1564
17	Femmina	55	2	ILC	ER+	Non trattato	29733 x 8609
18	Femmina	63	3	IDC	TNBC	Non trattato	29733 x 7985
19	Femmina	49	3	IDC	TNBC	Non trattato	29733 x 7986
20	Femmina	47	3	IDC	TNBC	Non trattato	29733 x 7023
21	Femmina	73	3	MBC	TNBC	Trattato	29733 x 5619
22	Femmina	67	3	IDC	TNBC	Non trattato	29733 x 4149
23	Femmina	58	3	IDC	HER2+	Non trattato	29733 x 2447
24	Femmina	52	3	MBC	TNBC	Trattato	29733 x 1754
25	Femmina	42	2	IDC	ER+	Non trattato	29733 x 4409
26	Femmina	47	2	ILC	ER+	Non trattato	29733 x 3961

Tabella 1 – Dettagli dei pazienti dell'articolo di riferimento

Inizialmente, è stata condotta una ricerca esaustiva per conoscere i vari metodi esistenti per il processing dei dati di scRNA-seq e per l'annotazione cellulare.

Sono stati studiati i metodi di classificazione cellulare su cui ogni software affonda le basi e successivamente, sono stati esaminati uno per uno, valutandone caratteristiche positive e

criticità individuali e, sulla base di queste, sono stati selezionati gli approcci più adatti alle esigenze specifiche.

La difficoltà iniziale è stata che non tutti i software disponibili sono tagliati per tutti i tessuti; la ricerca è stata quindi ristretta agli strumenti in grado di elaborare le analisi di dati di scRNA-seq provenienti da tessuto mammario e che quindi incorporavano nel loro database riferimenti per quel tessuto. Di seguito sono riportati i vari programmi utilizzati per l'analisi.

3.2 Cellenics

Il candidato iniziale per l'analisi dei dati di scRNA-seq è stato Cellenics (https://www.biomage.net).

Cellenics è un pratico strumento disponibile online che permette l'analisi dei dati di scRNA-seq. Dopo la creazione di un account è possibile caricare i file *barcodes.tvs*, *genes.tsv* e *matrix.mtx* e procedere con l'elaborazione dei dati.

In prima istanza Cellenics opera un controllo qualità sui dati grezzi lavorando su sette livelli sequenziali in cui l'output di uno rappresenta l'input per il successivo.

I primi 5 passaggi consistono in filtri per rimuovere dati indesiderati e di scarsa qualità da ogni singolo campione. Nel passaggio 6, vengono integrati più set di dati di esempio per rimuovere gli effetti batch e viene eseguita la riduzione della dimensionalità. Infine, nel passaggio 7, l'incorporamento è configurato (ad es. UMAP o t-SNE) e viene applicato il clustering.

Passaggio 1: filtro classificatore

Mira ad escludere le goccioline vuote utilizzando il metodo "emptydrops" che calcola il False Discovery Rate (FDR), un valore statistico che rappresenta la probabilità che una gocciolina sia vuota. Il valore FDR predefinito è 0,01 per tutti i campioni. Si conservano

solo le goccioline con FDR < 0,01. Pertanto, in questa fase, le goccioline con basso FDR vengono mantenute per l'analisi a valle, mentre le goccioline con un FDR elevato vengono rimosse.

I dati vengono visualizzati in un grafico a ginocchio che classifica le cellule in base al numero di UMI per ogni barcode su scala logaritmica.

Passaggio 2: filtro di distribuzione delle dimensioni delle cellule

A differenza del filtro classificatore che funziona sulla probabilità, questo filtro imposta una soglia rigida sul numero minimo di UMI contenuti in una goccia in modo che quella goccia sia considerata una cellula reale. Quindi, le cellule con UMI inferiori a questa soglia vengono filtrate. I valori di cut-off vengono calcolati automaticamente come punto di flesso del grafico a ginocchio e l'intervallo tipico della soglia minima è di circa 500-2000 UMI per cellula.

Passaggio 3: filtro del contenuto mitocondriale

Assumendo che le cellule morte riversano l'RNA mitocondriale nella cellula, Cellenics considera non vive, ed elimina, le cellule che presentano una percentuale di trascrizioni mitocondriali superiore al valore soglia. Per l'analisi sono stati utilizzati i cut-off di default che ammettono una percentuale massima dell'11,89%.

Passaggio 4: Numero di geni rispetto al filtro UMI

Segue il principio secondo il quale il numero di trascrizioni uniche (UMI) aumenta linearmente con il numero di geni. Le cellule che si discostano da questa relazione lineare vengono eliminate perché di scarsa qualità per l'analisi.

Passaggio 5: filtro Doublet

Esclude dall'analisi le cellule che presentano un'alta probabilità di essere un doppietto o multipletto utilizzando l'algoritmo scDblFinder.

Passaggio 6: integrazione dei dati

Rimuove gli effetti batch e riduce la dimensionalità dei dati attraverso 3 metodi alternativi: Harmony, Fast MNN e Seurat v4. Per l'analisi di tutti i campioni è stato utilizzato il metodo di default Harmony.

Prima dell'integrazione, a ciascun campione viene inoltre applicata una normalizzazione attraverso il metodo LogNormalize.

Passaggio 7: configurare l'incorporamento

Attraverso la Principal Component Analysis (PCA) si ottengono le componenti principali utili per eseguire una riduzione della dimensionalità attraverso il metodo UMAP grazie al quale è possibile ridurre la complessità del set di dati preservando la variazione.

I dati filtrati e integrati con il clustering sono quindi disponibili per l'esplorazione e la visualizzazione a valle nei moduli *Data Exploration* e *Plots and Tables* in cui Cellenics restituisce una lunga serie di output utili per l'interpretazione dei dati trattati di seguito.

Come primo risultato si ottiene un grafico con l'incorporamento UMAP in cui le cellule, rappresentate da punti, vengono raggruppate e colorate in base al clustering: un processo di raggruppamento di cellule di alta somiglianza; il metodo utilizzato in Cellenics è Louvian cluster (Figura 3a).

Di seguito si trova l'elenco dei cluster ottenuti secondo il metodo di clusterizzazione di Louvian (Figura 3b). Il numero dei cluster può essere aumentato o diminuito modificando la risoluzione: con una bassa risoluzione si osserva un minor numero di cluster, aumentando la risoluzione si ottengono più cluster. Il paramento di default di 0,8 è risultato in tutti i casi molto alto; pertanto, è stato ritenuto necessario ridurlo, adeguandolo caso per caso, per ottenere un minor numero di cluster e una maggiore eterogeneità tra di essi.

Si ottiene inoltre l'elenco completo dei geni presenti nel set di dati ordinati per dispersione (Figura 3c).

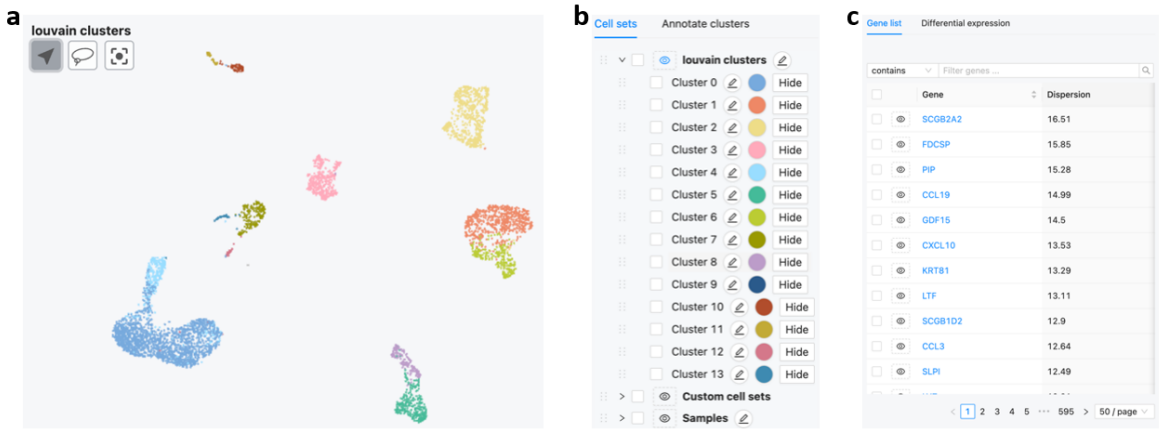


Figura 3 – Esempi di output ottenuti con Cellenics

Nella sezione *Plot and Table* è possibile trovare il Dot Plot (Figura 4) che mostra graficamente, sottoforma di cerchi pieni, la percentuale di cellule che esprime i geni inseriti nell'apposita finestra. Più piccolo è il cerchio, minore è la percentuale di cellule che esprime il gene corrispondente. L'intensità del colore invece rispecchia il livello di espressione del gene. È possibile inserire un elenco di geni, spaziati dalla virgola, e visualizzarne l'espressione nelle cellule dei vari cluster. Nel caso specifico, sono stati inseriti i geni che secondo la letteratura e i database presenti in rete, precedentemente esplorati, mostrano un'elevata espressione nelle cellule tumorali del seno.

In questo modo è stato possibile visualizzare i cluster con un'elevata espressione dei geni inseriti e inferirli tumorali.

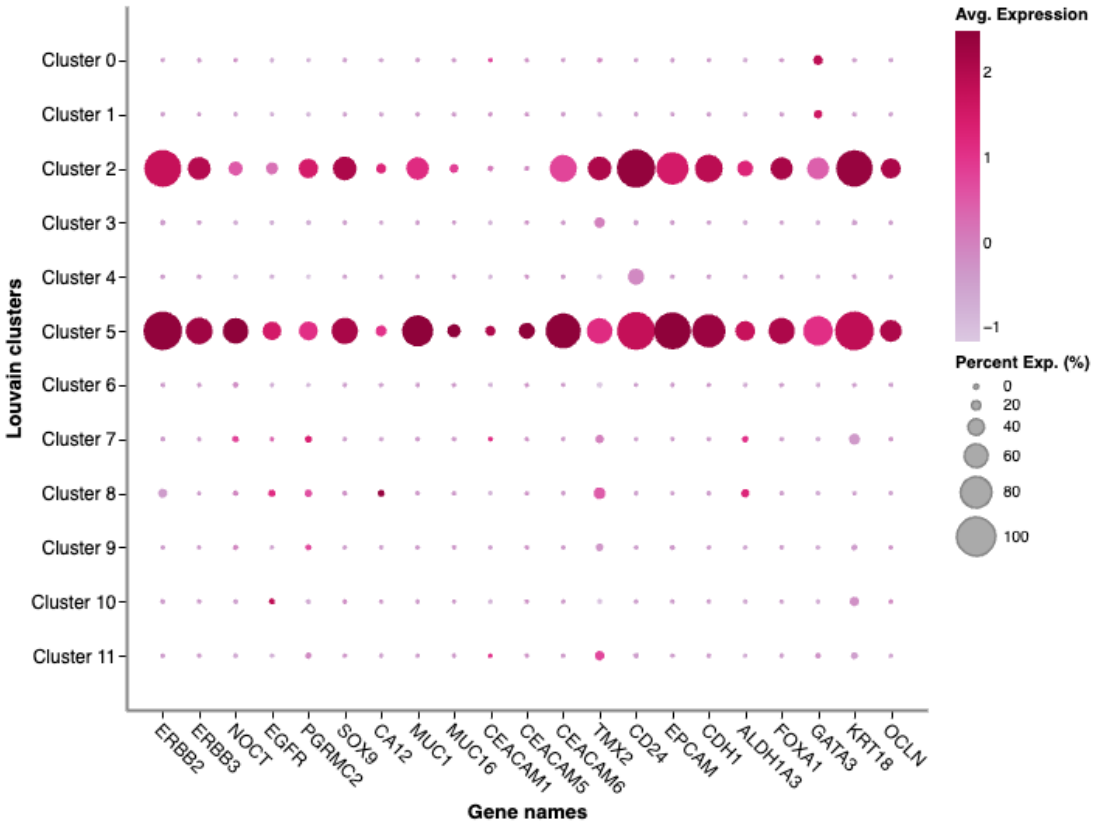


Figura 4 – Dot Plot del Paziente 3. Sono evidenti i cluster presunti tumorali rilevabili dalla sovraespressione dei geni tumorali inseriti

Nel caso del Paziente 3 riportato in esempio, si tratta di una donna di 60 anni con tumore di tipo HER2^{+} in cui il totale delle cellule isolate è di 3024 e di queste 340 sono state inferite tumorali attraverso la sovraespressione dei geni tumorali ERBB2, ERBB3, PGRMC2, NOCT, EGFR, SOX9, CA12, MUC1, MUC16, CEACAM1, CEACAM5, CEACAM6, TMX2, CD24, EPCAM, CDH1, ALDH1A3, FOXA1, GATA3, KRT18, OCLN rilevata nei cluster 2 e 5.

I markers genici tumorali sovraespressi negli altri pazienti, con i quali sono stati predetti i relativi cluster tumorali sono riportati nella Tabella 2.

Paziente	Geni utilizzati
Paziente 1	ADIRF, ALDH1A3, APLP2, C17orf89, CALML5, CD24, CDH1, CEACAM1, EPCAM, ERBB3, ERRFI1, GATA3, KRT15, KRT16, KRT18, KRT19, KRT23, KRT5, KRT7, KRT8, MAGED2, MIEN1, MYC, OCLN, PDZK1IP1, PPP1R1B, SLC39A6, SOX9, TACSTD2
Paziente 3	ALDH1A3, CA12, CD24, CDH1, CEACAM1, CEACAM5, CEACAM6, EGFR, ERBB2, ERBB3, EPCAM, FOXA1, GATA3, KRT18, MUC1, MUC16, NOCT, OCLN, PGRMC2, SOX9, TMX2
Paziente 7	ALDH1A3, CA12, CD24, CD55, CDH1, CEACAM1, EGFR, EPCAM, GATA3, KRT18, MUC1, MUC16, NOCT, OCLN, PGRMC2, SOX9, TACSTD2
Paziente 8	ADIRF, ALDH1A3, APLP2, C17orf89, EGFR, ERBB2, ERRGI1, KRAS, MAGED2, MIEN1, MUC1, MYC, P4HB, PGRMC2, SLC39A6, SOD3
Paziente 9	ALDH1A3, CA12, CD24, CDH1, CEACAM1, CEACAM5, CEACAM6, EPCAM, ERBB2, ERBB3, FOXA1, GATA3, KRT18, MUC1, OCLN, PGRMC2, SOX9, TMX2
Paziente 10	ADIRF, APLP2, C17orf89, CA12, CD24, CDH1, EPCAM, ERBB2, ERBB3, FOXA1, GATA3, KRT15, KRT18, KRT19, KRT23, KRT7, KRT8, MAGED2, MIEN1, MUC1, MYC, PDZK1IP1, PGRMC2, PPP1R1B, SLC39A6, SOX9, TACSTD2, TMX2
Paziente 13	CA12, CALML5, CD24, CD55, CDH1, EPCAM, ERBB2, ERBB3, ERRFI1, FOXA1, GATA3, KRT15, KRT18, KRT19, KRT23, KRT7, KRT8, MIEN1, MUCL1, MYC, OCLN, P4HB, PDZK1IP1, PPP1R1B, SLC39A6, SOX9, TACSTD2, TMX2
Paziente 15	ADIRF, APLP2, C17orf89, CA12, CALML5, CD24, CDH1, EPCAM, ERBB2, ERBB3, ERRFI1, FOXA1, GATA3, INSIG1, KRT18, KRT19, KRT7, KRT8, MAGED2, MIEN1, MUC1, MUCL1, MYC, NOCT, OCLN, PPP1R1B, SLC39A6, TACSTD2, TMX2
Paziente 17	ALDH1A3, C17orf89, CA12, CALML5, CD24, CDH1, CLDN4, EGFR, EPCAM, ERBB2, ERBB3, ERRFI1, FOXA1, GATA3, KRT14, KRT15, KRT16, KRT17, KRT18, KRT19, KRT23, KRT5, KRT7, KRT8, LTF, MUC1, OCLN, PDZK1IP1, PGRMC2, PPP1R1B, SLC39A6, SLPI, SOX9, TACSTD2
Paziente 21	ALDH1A3, BCAM, CA12, CCND1, CHMP2A, CRYAB, CYB5A, EGFR, ERBB2, ERRFI1, FBXO32, KRT15, KRT17, KRT18, KRT19, KRT23, KRT8, MAGED2, MGP, MGST1, MYC, NOCT, PDLIM3, PGRMC2, PIGT, PTN, PVRL2, SDC4, SFN, SFRP1, SLC39A6, SOX9, TACSTD2, TMX2, TNFRSF12A, TRPS1, TSPAN13

Tabella 2 – Geni tumorali sovraespressi nel tumore al seno utilizzati per predire i cluster tumorali nei diversi pazienti.

In Cellenics è disponibile, inoltre, la matrice di espressione normalizzata scaricabile sottoforma di tabella; è possibile selezionare anche soltanto i cluster d'interesse e in questo caso sono state scaricate le matrici dei presunti cluster tumorali in cui sulla prima riga sono riportati tutti i barcode delle cellule appartenenti ai cluster selezionati e sulle colonne l'espressione dei geni presi in considerazione. In questo modo sono stati ottenuti i barcode

delle cellule predette tumorali tramite Cellenics che in una seconda fase sono stati confrontati con i rispettivi risultati degli autori per valutarne la concordanza.

Sempre nei risultati è possibile consultare la Heatmap (Figura 5) in cui è riportata l'espressione dei 5 geni rappresentativi di ogni cluster, per tutte le cellule; presenta i geni sulle colonne e le cellule sulle righe. I colori più caldi come il giallo, l'arancione e il fucsia indicano un'espressione molto elevata di quel gene, mentre i colori più freddi come il blu indicano un'espressione minore. Attraverso questa mappa è possibile identificare visivamente i livelli diversi di espressione genica che distinguono i cluster tra loro. Inoltre, tenendo conto dei cluster tumorali predetti con il Dot Plot, è stata posta l'attenzione su quelli e sono stati osservati i geni maggiormente rappresentativi di quei cluster.

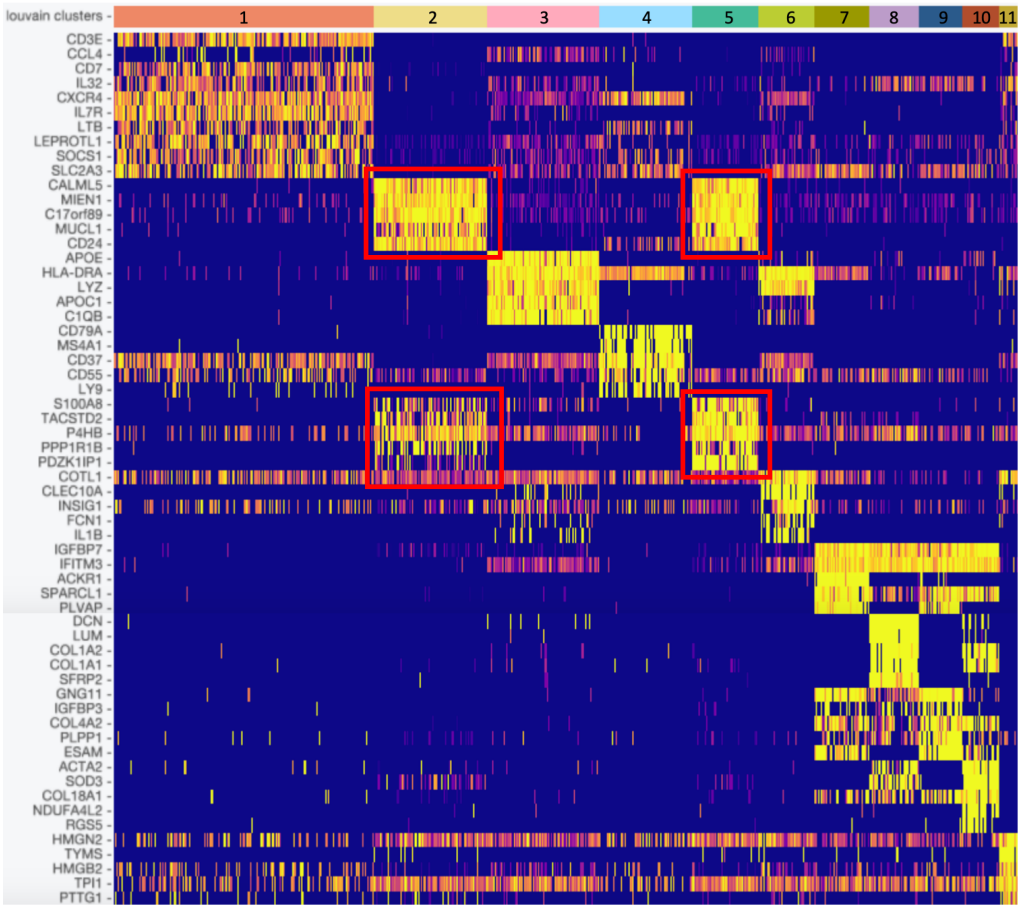


Figura 5 – Heatmap del Paziente 3. Sono evidenziati i geni sovraespressi nei presunti cluster tumorali (Cluster 2 e 5).

Dalla stessa interfaccia è possibile eseguire un'analisi dell'espressione differenziale che consente di determinare quali geni sono espressi in maniera differenziale tra due gruppi di cellule. In questo caso sono state osservate le differenze di espressione tra le cellule predette tumorali e quelle sane.

Il calcolo dell'espressione differenziale utilizza l'implementazione *presto* del test della somma dei ranghi di Wilcoxon e l'analisi auROC. Nel caso specifico sono stati prima creati due cluster: il Cluster A, costituito dalle cellule dei presunti cluster tumorali e il Cluster B con le cellule dei restanti cluster. È stata calcolata l'espressione differenziale tra i due cluster e si è ottenuto come risultato una tabella (Figura 6) contenente un elenco di geni in ordine decrescente di log fold change (logFC) insieme ai seguenti parametri:

- *LogFC*: è il rapporto tra l'espressione di un gene tra i due gruppi confrontati, trasformato in log;
- *Adj p-value*: la probabilità di osservare la differenza nell'espressione di un dato gene assumendo che questo non sia espresso in modo differenziale. Inoltre, il valore viene corretto utilizzando la correzione Benjamini-Hochberg per il test di ipotesi multiple, per tenere conto che quando si testano migliaia di geni, alcuni potrebbero avere un valore p basso per via del caso. Più è piccolo, maggiore è la probabilità che il gene sia effettivamente espresso in modo differenziale.
- *Pct1*: la percentuale di cellule del primo gruppo in cui è espresso il gene
- *Pct2*: la percentuale di cellule del secondo gruppo in cui è espresso il gene
- *AUC*: area sotto la curva caratteristica operativa del ricevitore (ROC). È proporzionale alla statistica Wilcoxon U calcolata mediante il test della somma dei ranghi. Più è grande, più è probabile che il gene corrispondente sia espresso in modo differenziale.

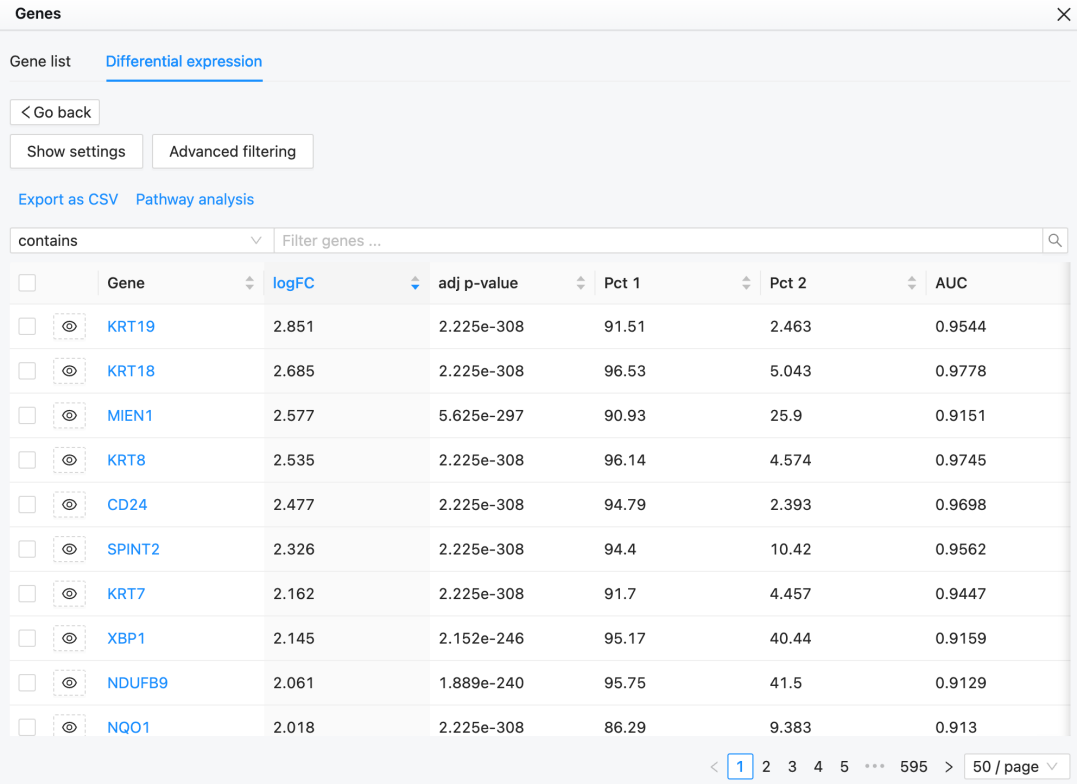


Figura 6 – Tabella ottenuta dall'espressione differenziale tra i cluster predetti tumorali e quelli sani nel Paziente 9. I geni che presentano un logFC maggiore sono quelli che hanno un'elevata espressione nel cluster A e una bassa espressione nel cluster B.

3.3 SciBet

SciBet (Single Cell Identifier Based on Entropy Test) è un altro uno strumento computazionale disponibile in rete oppure eseguibile in R che è stato investigato. È basato sull'inferenza bayesiana, per mezzo del quale è possibile prevedere l'identità cellulare di una cellula sequenziata mediante la tecnica del scRNA-seq sfruttando l'entropia statistica. https://doi.org/10.1101/645358

SciBet fa affidamento ad un set di dati di riferimento e il suo algoritmo prevede 4 passaggi: pre-elaborazione, selezione delle caratteristiche, addestramento del modello e assegnazione del tipo di cellula [13].

Per mitigare il rischio di annotazioni errate, a causa della limitata completezza nella raccolta dei dati di riferimento, SciBet utilizza un set di dati nullo come "sfondo".

In senso pratico, dopo aver fornito come input la matrice di espressione in formato .csv (Comma Separated Values), SciBet permette di caricare un file di riferimento da prendere in considerazione per l'annotazione cellulare o di sceglierne uno tra i 93 modelli presenti all'interno del proprio sito per poi proseguire con l'elaborazione alla fine della quale restituisce come risultati: un grafico, sottoforma di istogramma, in cui si evidenziano a colpo d'occhio i vari tipi cellulari e la relativa abbondanza all'interno del campione (Figura); un file in cui sono riportati i barcode con i rispettivi tipi cellulari predetti ed infine una tabella con la probabilità, che va da 0 a 1, di ogni cellula di appartenere ai vari tipi cellulari presi in considerazione nel set di dati di riferimento.

È stato utilizzato il classificatore online e predetti i tipi cellulari per ogni paziente.

Come primo dataset di riferimento è stato scelto *30 major human cell types* che contiene 30 principali tipi di cellule umane tra cui: astrociti, cellule B, chemochine (CC), endoteliali, epiteliali, FGC neuroni, cellule H9, miscela HEK e 3T3, HSC, ILC, macrofagi, mastociti, microglia, monociti, muscolari, NK, isole pancreatiche, Plasma B, mioblasti del muscolo scheletrico, soma, cellule T, cellule staminali embrionali, cellule fibroblastiche del prepuzio, epatociti, interneuroni, cellule progenitrici neurali, cellule progenitrici degli oligodendrociti e blastomero preimpianto.

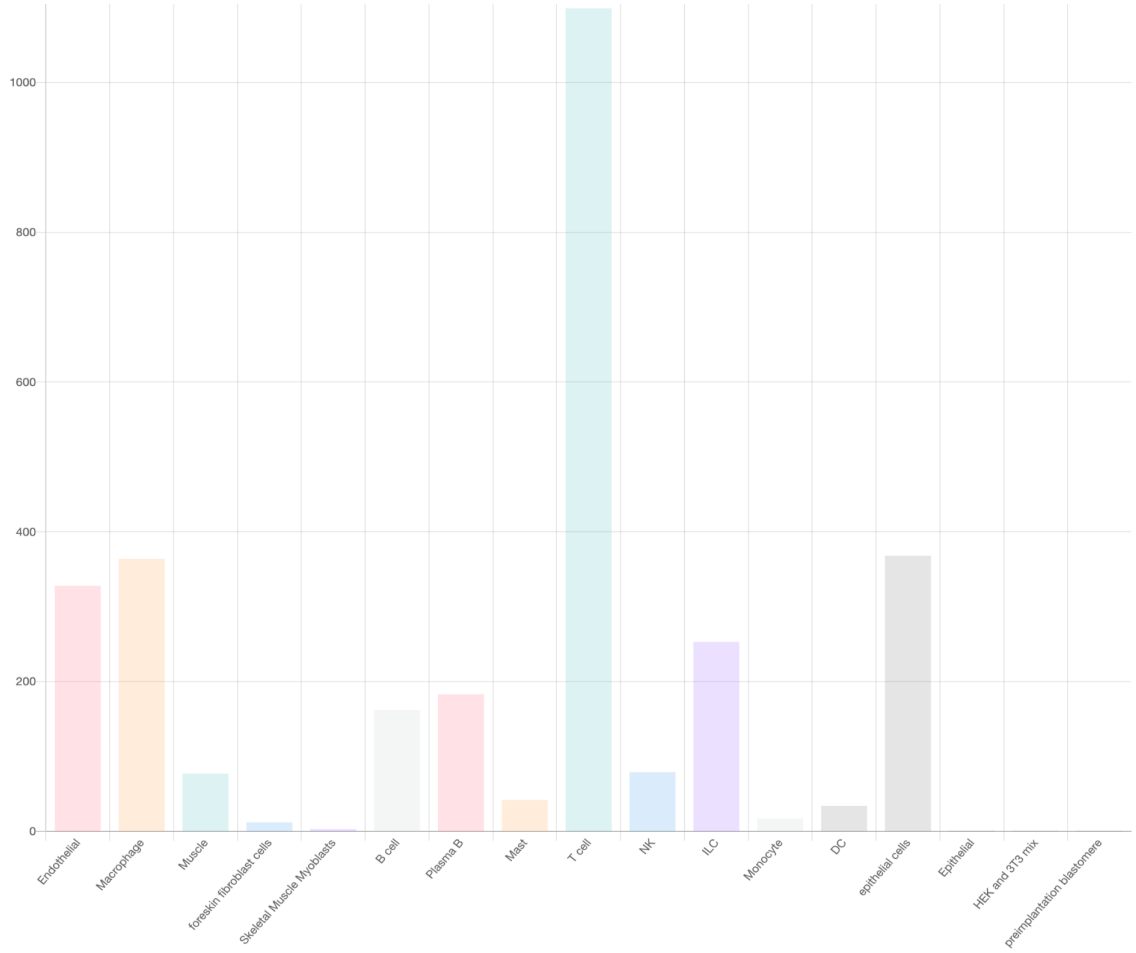


Figura 7 – Tipi cellulari predetti da SciBet nel Paziente 3.

Tra i set di dati di riferimento è disponibile anche uno specifico del tessuto mammario che permette però la distinzione soltanto tra cellule mioepiteliali e luminali epiteliali.

Le cellule mioepiteliali, come già anticipato, rappresentano una sorta di barriera tra l'epitelio del seno e lo stroma circostante, la cui perturbazione provoca il rilascio di fattori di crescita e di fattori angiogenici e delle specie reattive dell'ossigeno che causano un'alterazione del microambiente promuovendo la proliferazione delle cellule circostanti e aumentando l'invasività delle cellule tumorali [14].

Questo set di dati è stato utilizzato per condurre un'analisi dettagliata sulla composizione del tessuto in termini di cellule mioepiteliali e luminali.

3.4 Single R

SingleR è uno strumento disponibile pubblicamente come pacchetto in R, utilizzato per predire il tipo cellulare dei dati scRNA-seq; si basa su un set di geni di riferimento e non richiede ulteriori marcatori genetici predefiniti come input.

Single R si fonda sulla “correlazione con ottimizzazione iterativa", che utilizza, cioè, la correlazione (la relazione o dipendenza tra variabili) come base e incorpora un processo iterativo di ottimizzazione.

Il funzionamento di SingleR prevede:

Generazione del set di geni di riferimento: vengono raccolti e caratterizzati i profili di espressione genica di diverse popolazioni cellulari noti come set di geni di riferimento. Questi profili rappresentano le firme geniche tipiche di specifici tipi cellulari.

Calcolo della similarità: per ciascuna cellula del campione scRNA-seq, SingleR confronta il profilo di espressione genica con il set di geni di riferimento. Viene calcolata la similarità tra il profilo della cellula e quelli delle popolazioni di riferimento.

Assegnazione dell'etichetta di tipo cellulare: la cellula viene quindi assegnata al tipo cellulare che mostra la maggiore similarità con il suo profilo di espressione genica. In altre parole, SingleR cerca di identificare quale tipo cellulare noto assomigli di più alla cellula in esame.

Valutazione della precisione: SingleR fornisce anche misure di affidabilità dell'assegnazione, aiutando a valutare quanto è sicuro l'assegnare un tipo cellulare specifico a una particolare cellula.

L'obiettivo di SingleR è quindi semplificare e automatizzare il processo di assegnazione dei tipi cellulari alle singole cellule in un campione di dati scRNA-seq, facilitando l'analisi e l'interpretazione della complessità cellulare [15].

Anche con questo programma è stata eseguita un'annotazione cellulare dei campioni, prendendo come riferimento il database Human Primary Cell Atlas Data presente all'interno del pacchetto celldex.

Per condurre questa analisi sono stati scaricati i pacchetti SingleR, Scater e Celldex. Di seguito è riportato lo script utilizzato (Figura 8).

```
> library(SingleR)
  library(scater)
  library(celldex)

  # Reference dataset
  hpca.se <- celldex::HumanPrimaryCellAtlasData()
  # ImmGenData BlueprintEncodeData MonacoImmuneData MouseRNAseqData NovershternHematopoieticData

  # New dataset
  X <- read.csv('C:/Users/FRANCESCA/OneDrive - Università Politecnica delle Marche/Desktop/PazientiSingleR
  /Paziente3.tsv', sep = '\t', row.names = 1, na.string=".")
  X <- SummarizedExperiment(list(counts = X))
  X <- logNormCounts(X)

  # Cell types
  CT <- SingleR(test = X, ref = hpca.se, labels = hpca.se$label.main)
  writeLines(CT$labels, 'Paziente3.csv')
```

Figura 8 – Script utilizzato in SingleR per l'annotazione dei tipi cellulari

Al termine dell'analisi il programma restituisce una tabella contenente i barcode con i relativi tipi cellulari individuati i quali sono stati poi confrontati con i tipi cellulari predetti dagli altri software utilizzati.

3.5 Matlab

MATLAB è un linguaggio di programmazione ampiamente utilizzato nell'ambito dell'ingegneria, della matematica, della bioinformatica e in molti altri settori tecnici e scientifici.

Il termine MATLAB deriva dalle parole "Matrix Laboratory" e riflette la sua forza nel trattamento delle operazioni su matrici, elemento centrale delle sue funzionalità.

MATLAB include una varietà di strumenti grafici per la visualizzazione dei dati e può essere esteso tramite l'aggiunta di toolbox specializzati che forniscono funzionalità supplementari

per specifici settori o applicazioni. Nell'ambito del scRNA-seq, MATLAB offre il pacchetto scGEATool specifico per l'analisi di dati di questo tipo.

Dopo l'installazione del tool è possibile procedere con l'analisi dei dati passando prima attraverso il controllo qualità fino ad arrivare all'annotazione dei diversi tipi cellulari.

Questo strumento offre, inoltre, la possibilità di assegnare ad ogni cellula la propria fase del ciclo cellulare.

Dopo aver installato il pacchetto scGEATool sono stati caricati i 3 file matrix, barcode e feautures e dato il via all'analisi. È stato effettuato un primo controllo di qualità con i valori di cut-off predefiniti (Figura 9), seguito dalla rimozione dei geni mitocondriali e quelli ribosomiali. Successivamente è stata applicata una riduzione della dimensionalità attraverso il metodo UMAP per poi procedere con il clustering. Alla fine, sono stati ottenuti sia l'annotazione cellulare che le fasi del ciclo cellulare per ogni cellula.

I risultati sono scaricabili dal workspace che offre delle tabelle in cui vengono riportati i barcode e i relativi tipi cellulari, i barcode e i relativi cluster, i barcode e la relativa fase del ciclo cellulare.

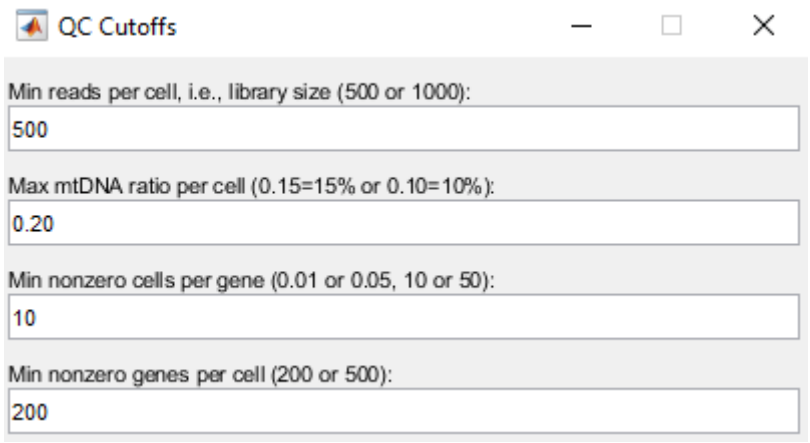


Figura 9 – Valori di cut-off utilizzati per l'analisi dei dati in tutti i pazienti in esame

3.6 Icarus

ICARUS (Interactive single cell RNA-seq analysis with R-shiny using Seurat) si presenta come un server web facilmente accessibile (https://launch.icarus2-scrnaseq.cloud.edu.au/app/ICARUS) attraverso il quale è possibile analizzare e processare dati di scRNA-seq sfruttando funzioni di Seurat [16].

Permette di operare il controllo qualità, la riduzione della dimensionalità, il raggruppamento delle cellule, l'annotazione dei cluster, ecc.

Gli aggiornamenti più recenti hanno inoltre permesso l'inserimento di funzionalità aggiuntive, tra cui, l'identificazione di doppietti, attraverso l'utilizzo di DoubletFinder.

Attualmente sono supportati undici organismi, tra cui uomo, cane, topo, ratto, zebrafish, moscerino della frutta, nematodi, lieviti, bovini, pollo e maiale.

ICARUS è stato utilizzato per sfruttare la funzione DoubletFinder e predire i doppietti dei campioni. Come input ICARUS richiede una matrice di geni per cellule sottoforma di tabella delimitata da tab con le cellule nelle colonne e i geni nelle righe, o come oggetto Seurat R (file RDS) sottoforma di dati di output 10X CellRanger (barcodes.tsv, features.tsv e matrix.mtx). Per motivi di privacy dei dati, questi non vengono conservati sul server dopo la fine della sessione dell'utente [17]

3.7 Scevan

SCEVAN (Single CEll Variational Aneuploidy aNalysis) è un algoritmo variazionale che sfrutta la stima delle comuni alterazioni del numero di copie (CNV) caratteristiche delle cellule neoplastiche per rilevare cluster tumorali da dati di scRNA-seq [18].

Partendo dalla matrice di conteggio grezza, questa subisce una pre-elaborazione attraverso la trasformazione logaritmica, l'eliminazione delle cellule con basso numero di trascrizioni e la selezione dei geni più espressi.

SCEVAN opera una distinzione tra le cellule sane e quelle maligne; queste ultime sono poi analizzate attraverso un algoritmo di segmentazione articolare basato sull'ottimizzazione.

Per tutti i pazienti è stata condotta un'analisi di SCEVAN in R utilizzando come parametro specifico di segmentazione 0,5 (beta_vega).

4. RISULTATI

4.1 Cellenics

Attraverso l'analisi con Cellenics è stato possibile, partendo dai 3 file scaricati direttamente da NCBI GEO, operare un'analisi dettagliata ponendo l'attenzione sui cluster tumorali. Questi sono stati predetti inserendo manualmente i geni altamente espressi nel tumore al seno precedentemente indagati attraverso la letteratura e interrogando database pubblici come Cell Marker 2.0 (http://117.50.127.228/CellMarker/CellMarker_help.html).

I marcatori genici utilizzati per individuare i cluster tumorali dei vari pazienti sono riportati nella Tabella 2.

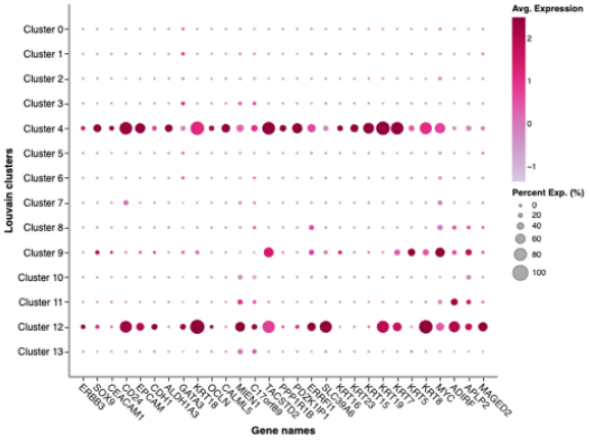
Seppure dall'analisi sia emersa una congruenza relativamente elevata tra i geni maggiormente espressi nei vari cluster predetti tumorali nei diversi pazienti, è stato visto che non tutti i geni sono sovraespressi in tutti i pazienti. Di seguito è riportato il confronto tra 3 pazienti come esempio (Figura 10).

PAZIENTE 1

6178 cellule

43 anni – IDC – Grado 3

PR	+
ER	+
HER2	+



Geni utilizzati:
ADIRF, ALDH1A3, APLP2, C17orf89, CALML5, CD24, CDH1, CEACAM1, EPCAM, ERBB3, ERRFI1, GATA3, KRT15, KRT16, KRT18, KRT19, KRT23, KRT5, KRT7, KRT8, MAGED2, MIEN1, MYC, OCLN, PDZK1IP1, PPP1R1B, SLC39A6, SOX9, TACSTD2

Cluster predetti tumorali:
4 e 12

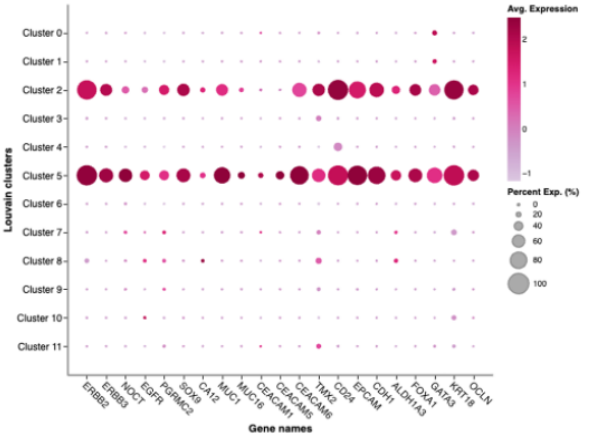
Totale cellule tumorali predette:
615 (9,95%)

PAZIENTE 3

3024 cellule

60 anni – IDC – Grado 3

PR	-
ER	-
HER2	+



Geni utilizzati:
ERBB2, ERBB3, PGRMC2, NOCT, EGFR, SOX9, CA12, MUC1, MUC16, CEACAM1, CEACAM5, CEACAM6, TMX2, CD24, EPCAM, CDH1, ALDH1A3, FOXA1, GATA3, KRT18, OCLN

Cluster predetti tumorali:
2 e 5

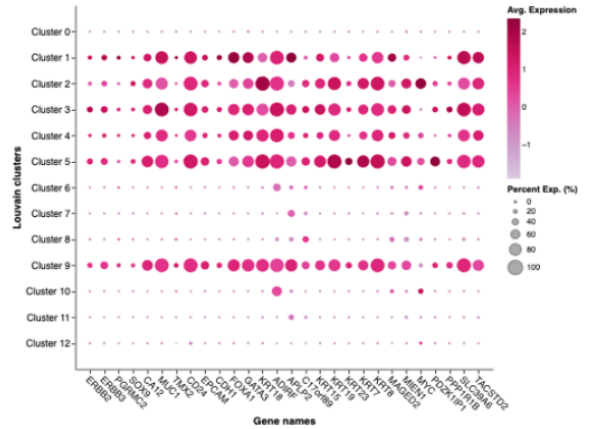
Totale cellule tumorali predette:
340 (11,24%)

PAZIENTE 10

3764 cellule

85 anni – IDC – Grado 2

ER	+
PR	+
HER2	+



Geni utilizzati
ADIRF, APLP2, C17orf89, CA12, CD24, CDH1, EPCAM, ERBB2, ERBB3, FOXA1, GATA3, KRT15, KRT18, KRT19, KRT23, KRT7, KRT8, MAGED2, MIEN1, MUC1, MYC, PDZK1IP1, PGRMC2, PPP1R1B, SLC39A6, SOX9, TACSTD2, TMX2

Cluster predetti tumorali:
1, 2, 3, 4, 5 e 9

Totale cellule tumorali predette:
998 (26,51%)

Figura 10 – DotPlot dei Pazienti 1, 3 e 10 ricavati da Cellenics attraverso l'utilizzo di specifici marker genici tumorali

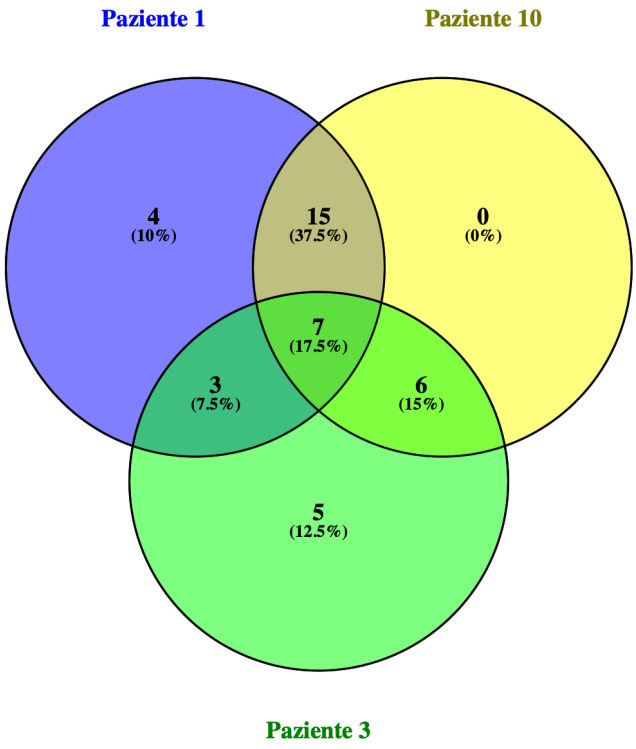


Figura 11 – Geni sovraespressi nei cluster predetti tumorali dei Pazienti 1, 3 e 10

In comune tra i 3 campioni: CD24, CDH1, EPCAM, ERBB3, GATA3, KRT18, SOX9.

In comune tra il Paziente 1 e 3: ALDH1A3, CEACAM, OCLN

In comune tra il Paziente 1 e 10: ADIRF, APLP2, C17orf89, KRT15, KRT19, KRT23, KRT7, KRT8, MAGED2, MIEN, MYC, PDZK1IP1, PPP1R1B, SLC39A6, TACSTD2

In comune tra il Paziente 3 e 10: CA12, ERBB2, FOXA1, MUC1, PGRMC2, TMX2

Si può notare che il Paziente 1 e il Paziente 10 mostrano una più alta corrispondenza per i geni tumorali sovraespressi. Confrontando i 3 fenotipi tumorali emerge una più stretta corrispondenza anche in questo caso tra il Paziente 1 e il Paziente 10, entrambi con tumore di tipo Luminale A, caratterizzato dalla sovraespressione dei recettori ER, PR e HER2.

Al contrario, il Paziente 3 presenta un diverso fenotipo tumorale, di tipo $HER2^{+}$, caratterizzato dalla mancanza di sovraespressione dei recettori ER e PR, che riflette a un'altresì diversa espressione genica dei presunti cluster tumorali.

Questa correlazione rende conferma il fatto che la classificazione tra subtipi tumorali delinea fenotipi differenti in cui la sovraespressione di pattern genici cambia tra un sottotipo tumorale e l'altro.

È stato inoltre osservato che i geni predittivi di un cluster tumorale per un paziente, non sempre hanno mostrato la stessa sovraespressione in cluster presunti tumorali di un altro paziente.

La figura 12 riporta il DotPlot del Paziente 21 in cui i geni marcatori tumorali inseriti (CALML5, CD24, CD55, CDH1, EPCAM, FOXA1, GATA3, KRT7, MIEN1, MUCL1, OCLN, P4HB, PDZK1IP1, PPP1R1B, SLPI, CEACAM1, CEACAM5, CEACAM6, MUC16, AZGP1, SPINT2, WFDC2, XBP1, SCGB2A2, FXYD3, CLDN7, SAA1, DSP, IER3, PIP, CLDN3, FBXO32, VAMP8, AGR2, KRT19, MUC1, PDZK1IP1, NOCT) non hanno predetto nessun cluster tumorale. La predizione è stata resa possibile testando un'altra serie di geni tumorali (ALDH1A3, BCAM, CA12, CCND1, CHMP2A, CRYAB, CYB5A, EGFR, ERBB2, ERRFI1, FBXO32, KRT15, KRT17, KRT18, KRT19, KRT23, KRT8, MAGED2, MGP, MGST1, MYC, NOCT, PDLIM3, PGRMC2, PIGT, PTN, PVRL2, SDC4, SFN, SFRP1, SLC39A6, SOX9, TACSTD2, TMX2, TNFRSF12A, TRPS1, TSPAN13), in questo caso rilevati sovraespressi nei cluster 7, 8, 11, 12 e 13 che hanno delineato i presunti cluster tumorali, mascherati invece nella precedente analisi.

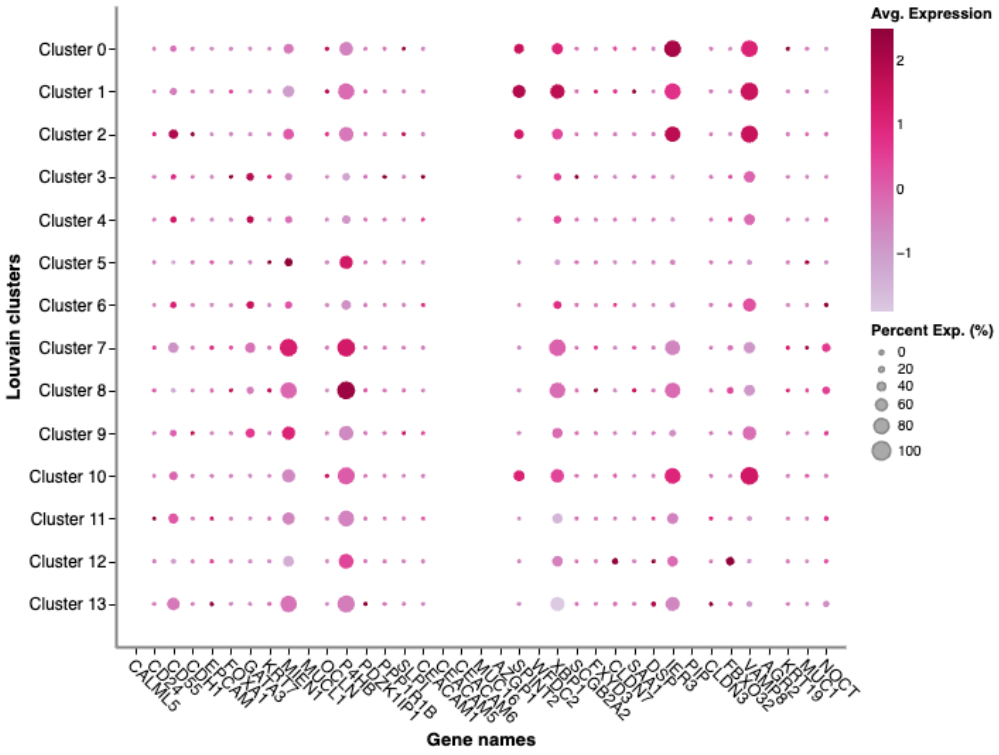


Figura 12 – Dotplot del paziente 21 in cui sono stati inseriti marker genici tumorali che non hanno rivelato nessun cluster tumorale

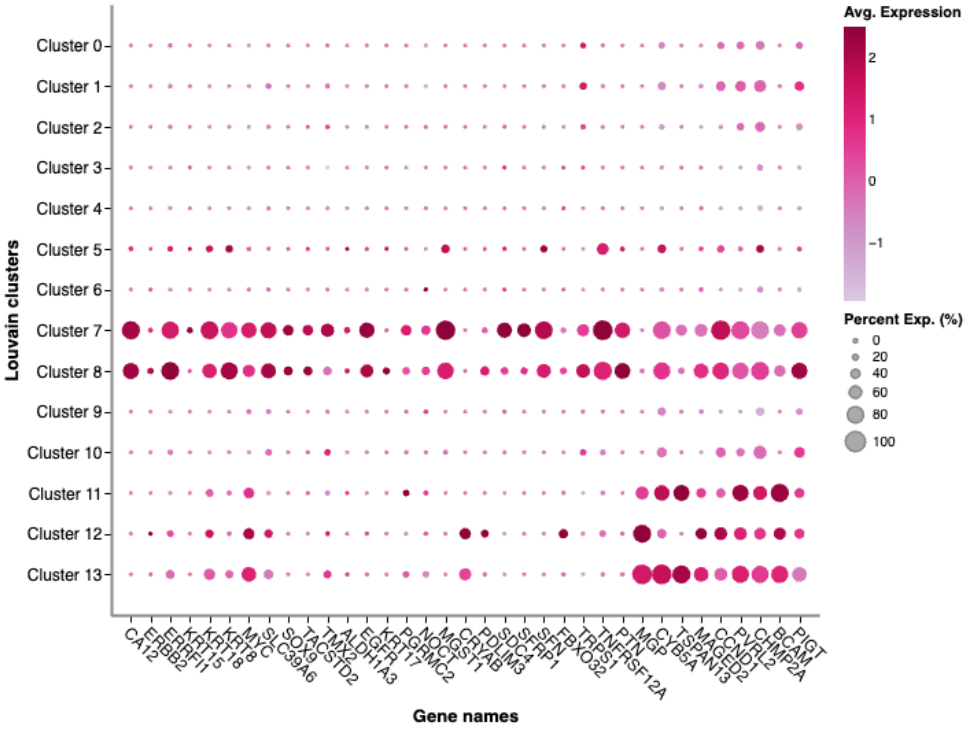
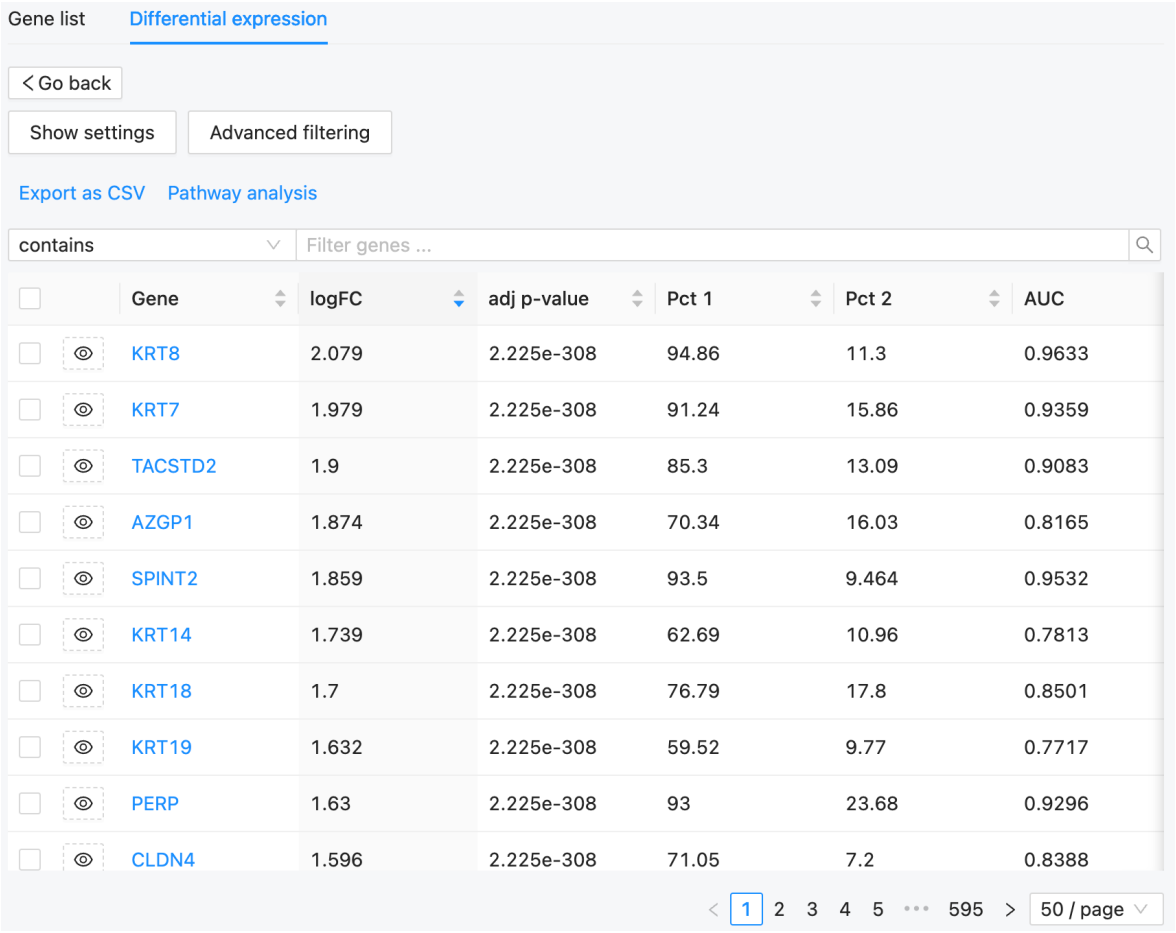


Figura 13 – DotPlot del paziente 21 in cui sono stati inseriti marker geni che hanno rivelato come tumorali i cluster 7, 8, 11, 12 e 13

Dalla figura 13 è interessante notare come all'interno di uno stesso campione i cluster tumorali presentano una sovraespressione differente di geni tumorali. I cluster 7-8 e i cluster 11-12-13 presentano un pattern diverso di espressione di geni tumorali.
Questo potrebbe predire l'esistenza di ulteriori subtipi all'interno di quelli già esistenti.
Il paziente 21, di 73 anni, presenta un tumore di tipo triplo negativo metastatico.
Da un'analisi in letteratura è stato osservato che i geni sovraespressi nei cluster 11, 12 e 13 (MGP, CYB5A, TSPAN13, MAGED2, CCND1, PVRL2, CHMP2A, BCAM e PIGT) potrebbero essere coinvolti nella progressione tumorale e nelle metastasi del tumore al seno, oltre che essere correlati ad una prognosi. La sovraespressione di questi geni riflette il sottotipo tumorale del paziente, confermando ancora una volta la relazione tra sottotipo e espressione genica e quanto l'analisi di quest'ultima sia importante per formulare una diagnosi quanto più precisa possibile.

L'analisi dell'espressione differenziale ha permesso di scovare i geni che presentano una sovraespressione esclusivamente nei cluster predetti tumorali e che invece sono presenti a bassi livelli nelle cellule predette sane. Come risultato si ottiene una tabella (Figura 14) recante i valori di logFC, adj p-value, Pct1, Pct2 e AUC.

I geni sono stati ordinati secondo il valore di logFC decrescente in modo da avere all'inizio quelli sovraespressi nel Cluster A, cioè nel cluster delle putative cellule tumorali e con una bassa espressione nel Cluster B, delle cellule predette sane.



Gene	logFC	adj p-value	Pct 1	Pct 2	AUC
KRT8	2.079	2.225e-308	94.86	11.3	0.9633
KRT7	1.979	2.225e-308	91.24	15.86	0.9359
TACSTD2	1.9	2.225e-308	85.3	13.09	0.9083
AZGP1	1.874	2.225e-308	70.34	16.03	0.8165
SPINT2	1.859	2.225e-308	93.5	9.464	0.9532
KRT14	1.739	2.225e-308	62.69	10.96	0.7813
KRT18	1.7	2.225e-308	76.79	17.8	0.8501
KRT19	1.632	2.225e-308	59.52	9.77	0.7717
PERP	1.63	2.225e-308	93	23.68	0.9296
CLDN4	1.596	2.225e-308	71.05	7.2	0.8388

Figura 14 – Espressione differenziale tra i putativi cluster tumorali e cluster sani nel Paziente 17

Dalla lista sono stati selezionati i primi 50 geni e visualizzati sulla Heatmap (Figura 15).

È stato visto che alcuni geni sovraespressi nei presunti cluster tumorali, presentano una sovraespressione anche negli altri cluster, questi solo stati filtrati eliminando i geni con un Pct2 > 50 ottenendo una Heatmap più pulita (Figura 16).

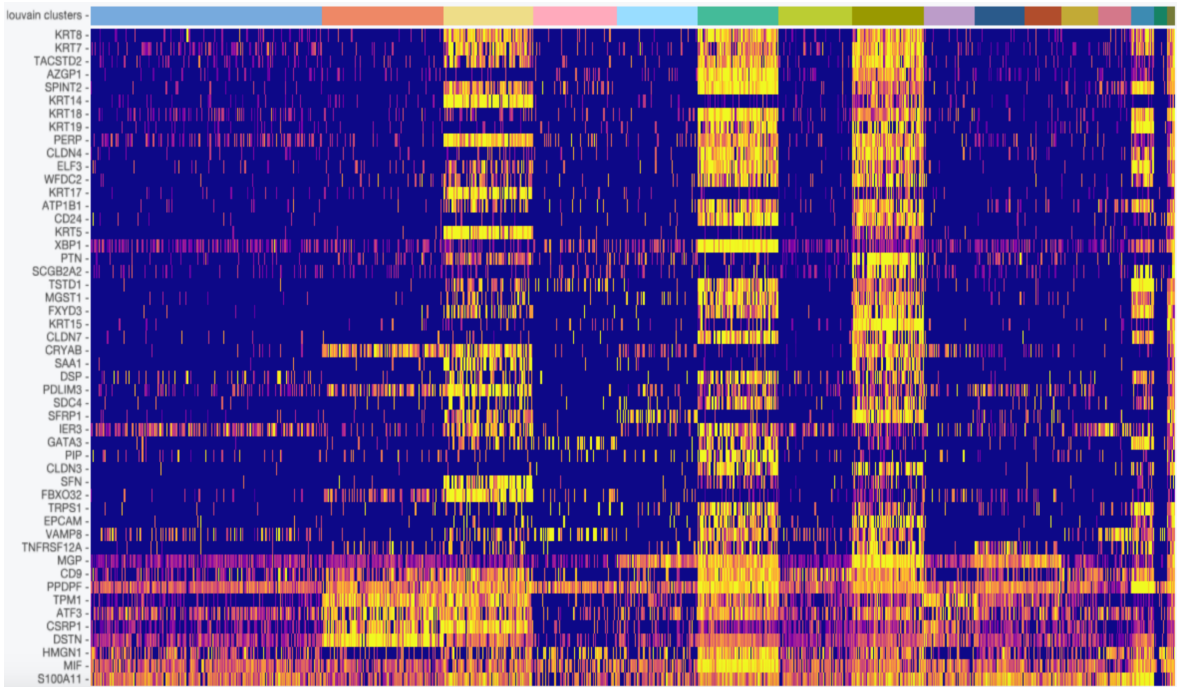


Figura 15 – Heatmap del paziente 17 con i 50 geni sovraespressi nei presunti cluster tumorali

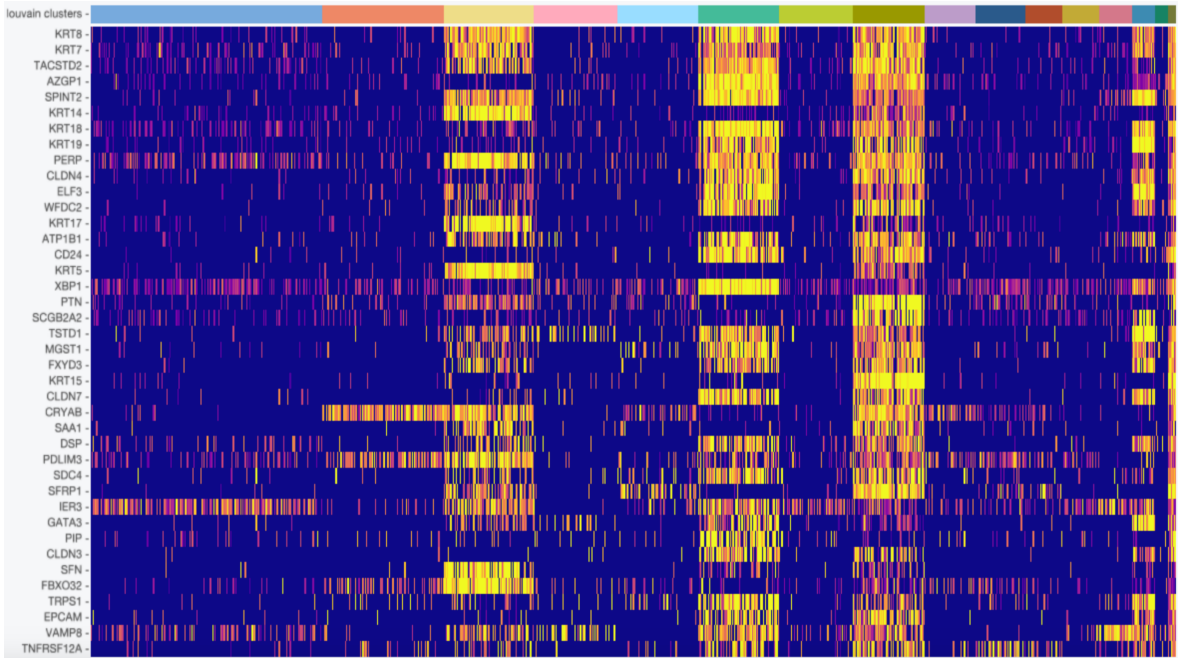


Figura 16 – Heatmap del paziente 17 in cui sono stati rimossi i geni sovraespressi sia nelle cellule predette tumorali che in quelle sane

Ad esempio, nel Paziente 17 i primi 50 geni sovraespressi nel cluster A sono: KRT8, KRT7, TACSTD2, AZGP1, SPINT2, KRT14, KRT18, KRT19, PERP, CLDN4, MGP, ELF3, WFDC2, KRT17, ATP1B1, CD24, KRT5, CD9, XBP1, PTN, SCGB2A2, TSTD1, MGST1, PPDPF, FXYD3, KRT15, CLDN7, TPM1, CRYAB, SAA1, ATF3, DSP, PDLIM3, CSRP1, SDC4, SFRP1, IER3, GATA3, PIP, CLDN3, SFN, FBXO32, DSTN, HMGN1, MIF, TRPS1, S100A11, EPCAM, VAMP8, TNFRSF12A.

Da questi sono stati eliminati i geni sovraespressi anche negli altri cluster ottenendo una lista con i geni sovraespressi esclusivi dei cluster potenzialmente tumorali.

Geni sovraespressi esclusivamente nei cluster presunti tumorali del Paziente 17:

KRT8, KRT7, TACSTD2, AZGP1, SPINT2, KRT14, KRT18, KRT19, PERP, CLDN4, ELF3, WFDC2, KRT17, ATP1B1, CD24, KRT5, XBP1, PTN, SCGB2A2, TSTD1, MGST1, FXYD3, KRT15, CLDN7, CRYAB, SAA1, DSP, PDLIM3, SDC4, SFRP1, IER3, GATA3, PIP, CLDN3, SFN, FBXO32, TRPS1, EPCAM, VAMP8, TNFRSF12A.

Sovraespressi anche negli altri cluster: MGP, CD9, PPDPF, TPM1, ATF3, CSRP1, DSTN, HMGN1, MIF, S100A11.

I geni sovraespressi nei cluster tumorali sono stati indagati ed è stata vista un'alta correlazione tra questi e la presenza di tumore al seno.

Le cheratine (KRT), ad esempio, attivano vie di segnalazione coinvolte nella migrazione cellulare, nell'invasione e nelle metastasi. È stato visto che l'espressione di KRT19 è particolarmente elevata nei tumori al seno ed è correlata alla loro invasività [19].

Inoltre, la sovraespressione di CD24 è un fattore prognostico sfavorevole indipendente nel cancro al seno, specialmente per i sottotipi luminali A e TNBC; è stato evidenziato che CD24 può essere un bersaglio terapeutico promettente per sottotipi specifici di cancro al seno [20].

Infine, è stato condotto un confronto tra le cellule predette tumorali attraverso Cellenics e quelle tumorali predette dagli autori tramite InferCNV.

Nella maggior parte di casi si è riscontrata una correlazione di circa l'80%, confermando un notevole grado di concordanza tra i due metodi, fatta eccezione per alcuni campioni in cui gli autori non hanno inferito cellule tumorali.

Attraverso l'annotazione manuale condotta con Cellenics è stato sempre possibile rivelare presunti cluster tumorali.

4.2 SciBet, Single R e Matlab

Come prima analisi con SciBet è stato utilizzato il data base *30 major human cell types* attraverso cui è stato possibile ricavare l'annotazione cellulare per ogni paziente, seppure il programma non riesca ad attribuire il tipo cellulare con una sicurezza elevata. I risultati sono stati poi confrontati con le predizioni ottenute con gli altri strumenti.

Una seconda analisi con SciBet si proponeva di scovare la diversa composizione cellulare in termini di cellule mioepiteliali e luminali epiteliali dei vari campioni, prendendo come riferimento il set di dati *Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity*.

Delle cellule predette tumorali da Cellenics è stata confrontata l'annotazione ottenuta con il database sopracitato con l'obiettivo di controllare se corrispondessero a cellule epiteliali luminali o a cellule mioepiteliali.

È emerso che delle cellule tumorali predette soltanto una piccola percentuale corrisponde a cellule mioepiteliali, a conferma del fatto che lo strato di cellule mioepiteliali nel tumore al seno, subisce lesioni che facilitano poi la progressione del tumore [14]. Infatti, tutti i pazienti

in esame presentano un tumore di tipo invasivo, caratterizzato dall'assenza dello strato di cellule mioepiteliali.

Dal confronto tra le annotazioni cellulari ottenute con i diversi software è emersa una concordanza intrasoftware, seppure non al 100%, solo per alcuni tipi cellulari.

Non tutti i software, infatti, riescono a predire gli stessi tipi cellulari e a farlo in modo univoco; in primo luogo, perché fanno affidamento ad algoritmi diversi per l'analisi e in secondo luogo perché utilizzano database di riferimento diversi.

Matlab nel caso specifico è riuscito a predire: cellule endoteliali, basali, dendritiche, cellule T, luminali epiteliali, pancreatiche, duttali e NK, fibroblasti.

Attraverso Single R sono state predette: cellule endoteliali, epiteliali, neuroni, monociti, dendritiche, HSC CD34+, muscolari lisce, fibroblasti, staminali, osteoblasti, condrociti, cellule B, Pro-B-cell CD34+, cellule T, natural killer (NK), macrofagi, Pre-B-cell CD34-, GMP, CMP.

SciBet, attraverso il database preso come riferimento riesce a predire un diverso numero e tipo di cellule; nel caso specifico il database utilizzato contiene 30 tipi cellulari diversi (Figura 17).

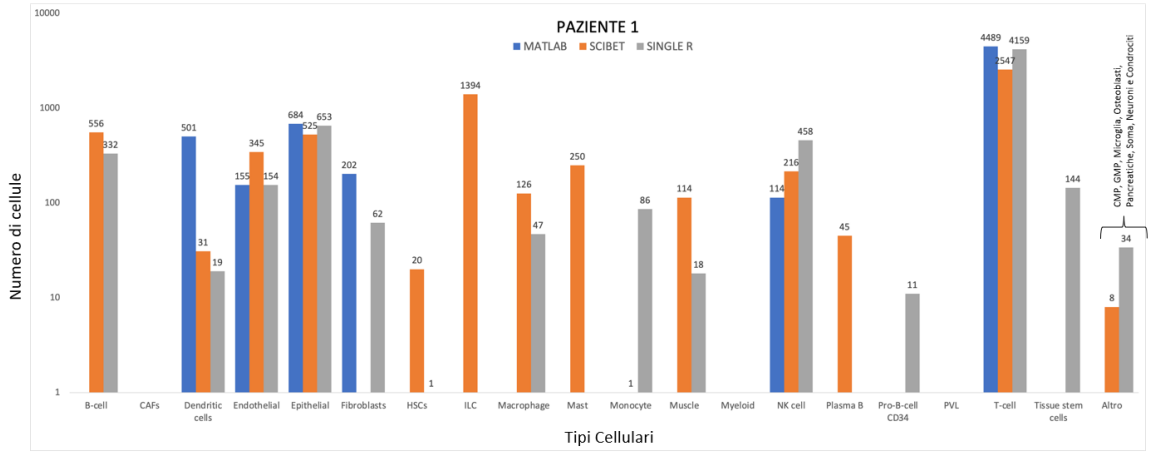


Figura 17 – Predizione dei tipi cellulari con i diversi software nel Paziente 1.

A conferma di questo, di seguito sono riportati i confronti di annotazione cellulare tra i tre software nel paziente 1 (Figura 18).

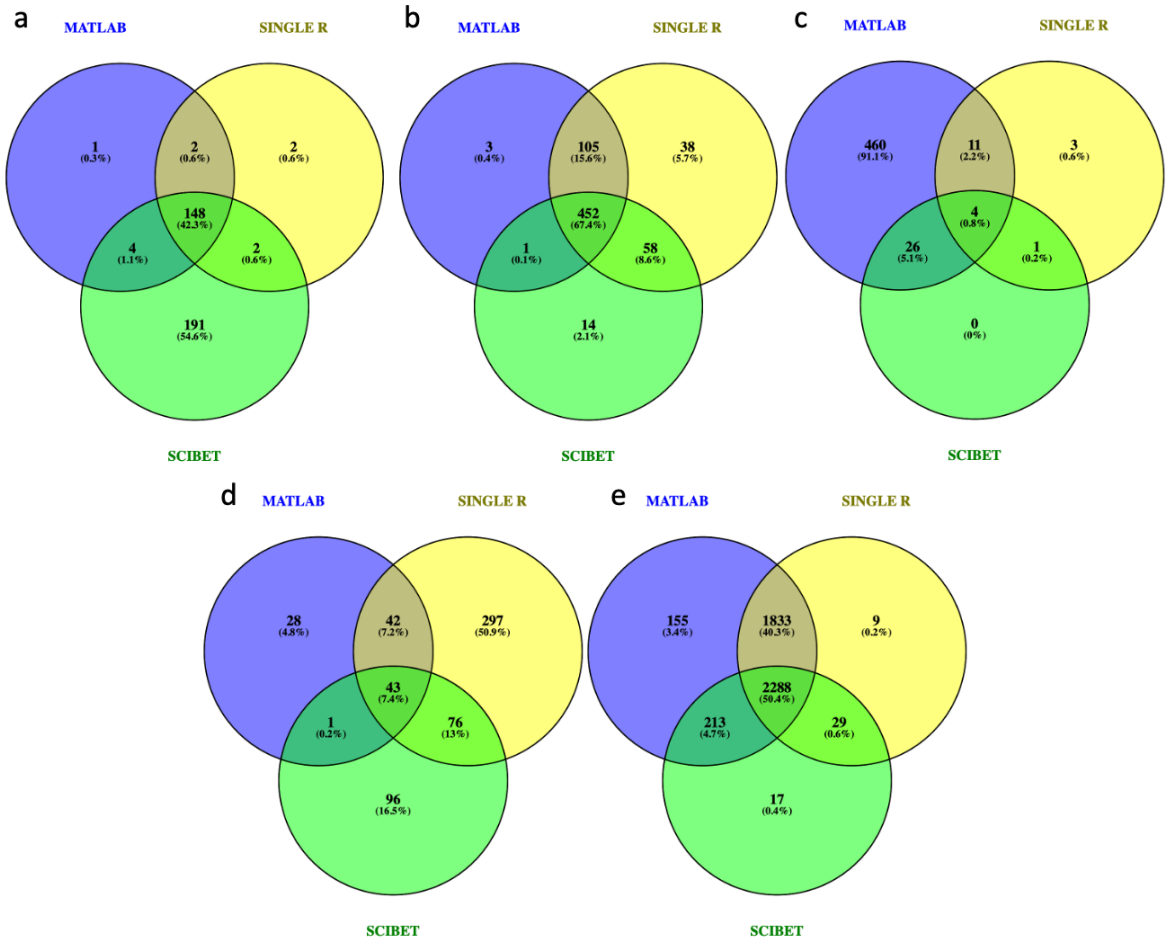


Figura 18 – Concordanza tra i programmi: a) cellule endoteliali, b) cellule epiteliali, c) cellule dendritiche, d) cellule NK, e) cellule T.

Le cellule in comune predette da tutti e 3 i programmi sono le cellule endoteliali, le cellule epiteliali, le cellule T, le cellule dendritiche e le cellule NK.

Per la predizione del tipo cellulare endoteliale, epiteliale e dei linfociti T, Matlab e Single R hanno mostrato una maggior corrispondenza compresa tra l'84,8 e 94,3% (Figura 18a, 18b e 18e).

Al contrario è stata rilevata una scarsa concordanza nella predizione delle cellule dendritiche e NK tra tutti e tre i programmi (Figura 18c e 18d).

La ragione di questo può essere attribuita al fatto che SciBet riesce a predire più tipi cellulari ma questo non è sempre un vantaggio in quanto espone ad un rischio maggiore di commettere errore.

Scibet e Single R hanno predetto tipi cellulari non appartenenti al tessuto mammario come osteoblasti, neuroni, cellule pancreatiche ecc. Questo suggerisce che non sempre un software che fa affidamento a database molto vasti, e che quindi riesce a predire una grande varietà di tipi cellulari sia la scelta migliore.

Al contrario, un programma che si basa su database poco completi, potrebbe incorrere in errori nell'annotazione poiché, non essendo in grado di riconoscere i tipi cellulari a causa dell'assenza di informazioni nei riferimenti, potrebbe attribuirlo ad un tipo cellulare noto anche se non corretto. *In medio stat virtus*.

Inoltre, nessuno dei tre programmi riesce a identificare cellule tumorali a causa della scarsità dei database di riferimento con cui elaborano l'analisi.

È quindi evidente che ad oggi, non esiste un programma di predizione univoco per l'annotazione cellulare di dati scRNA-seq né delle regole universali per ottenere dati quanto più attendibili.

È essenziale selezionare gli strumenti in base alle specifiche esigenze ed interpretare i risultati con attenzione, evitando di accettarli come verità assolute. Con l'ampliamento e il miglioramento dei database, sarà possibile ottenere risultati più concordanti tra i diversi programmi.

4.3 Icarus

Dall'analisi è emerso che solo una piccola percentuale di cellule tumorali sono state predette come doppietti. Nel caso del paziente 1, ad esempio, su 615 cellule tumorali predette da Cellenics, solo 174 di queste sono state considerate doppietti (28,3%).

Al contrario, quasi tutte le cellule predette come doppietti sono tumorali (85,2%).

4.4 Scevan

Il programma SCEVAN restituisce una lunga serie di output. Tra questi è significativa la rappresentazione dell'Heatmap dei CNV inferiti (Figura 19, 21e 23), in cui le cellule sono disposte sulle righe, mentre ogni colonna rappresenta un cromosoma. Attraverso la Heatmap è possibile visualizzare a colpo d'occhio, di ogni cromosoma, le duplicazioni (in rosso), le delezioni (in blu) e gli stati di neutralità (in bianco) per ogni singola cellula.

Sulla base dei CNV inferiti il programma per ogni campione raggruppa le cellule in tumorali (arancione) e non tumorali (verde); successivamente le cellule predette tumorali sono clusterizzate in sottogruppi differenti in funzione dei CNV in comune (Figura 20, 22 e 24).

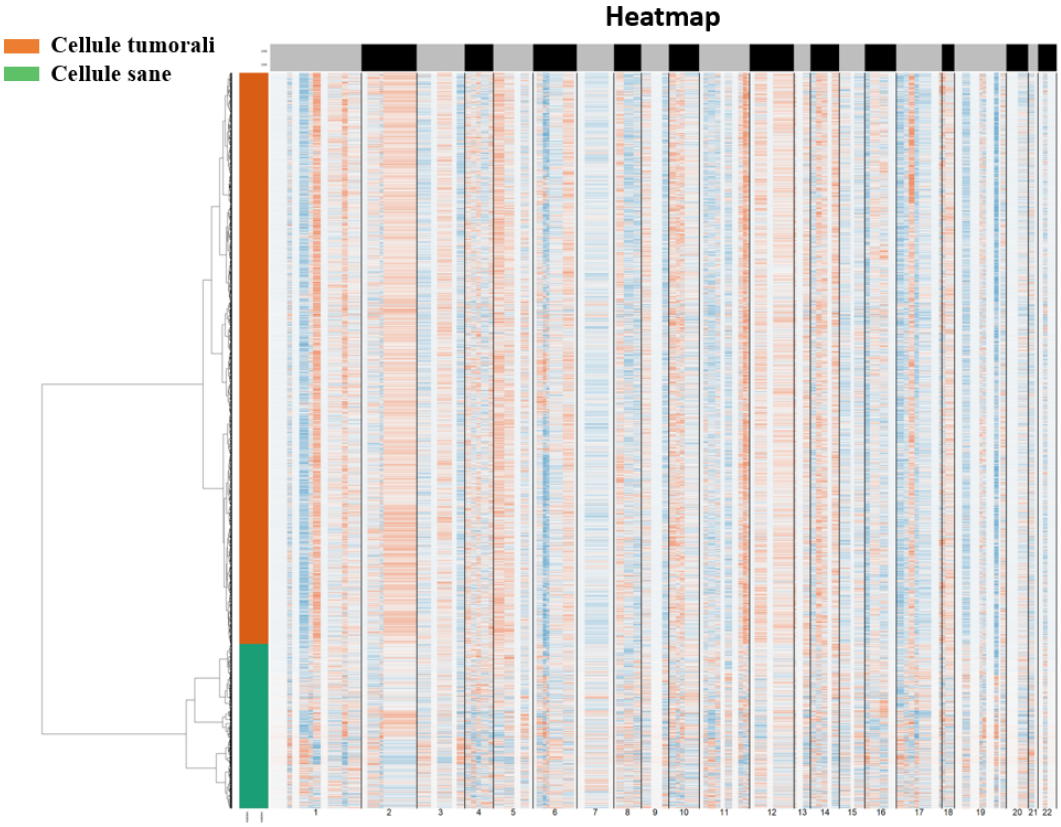


Figura 19 - Paziente 1

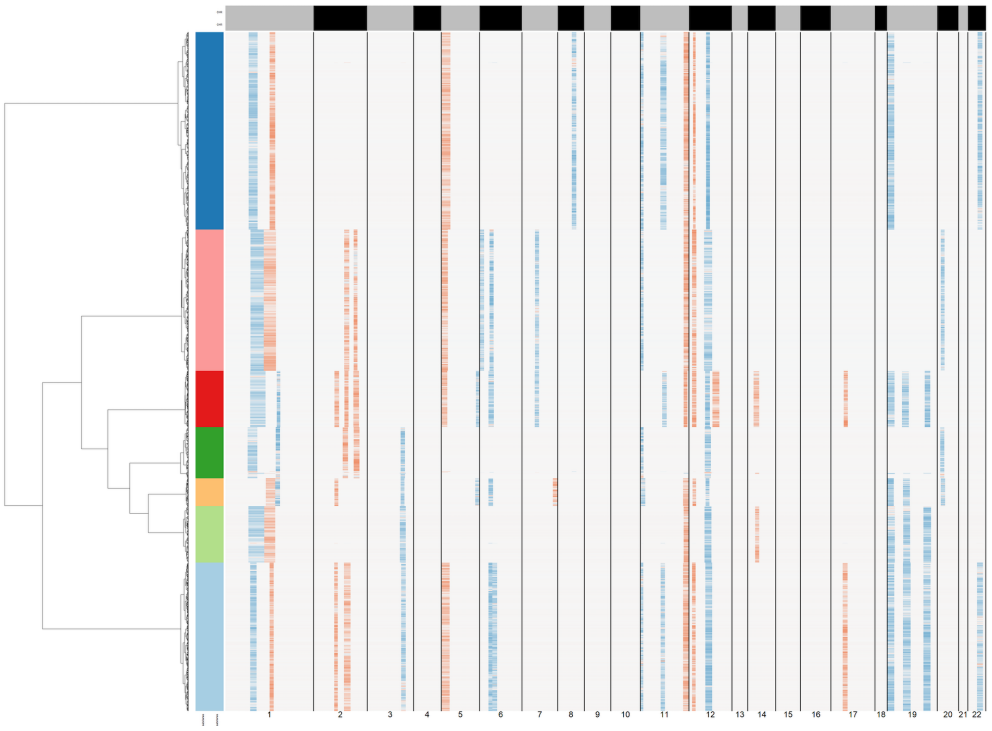


Figura 20 – Heatmap dei 7 subcloni tumorali nel Paziente 1

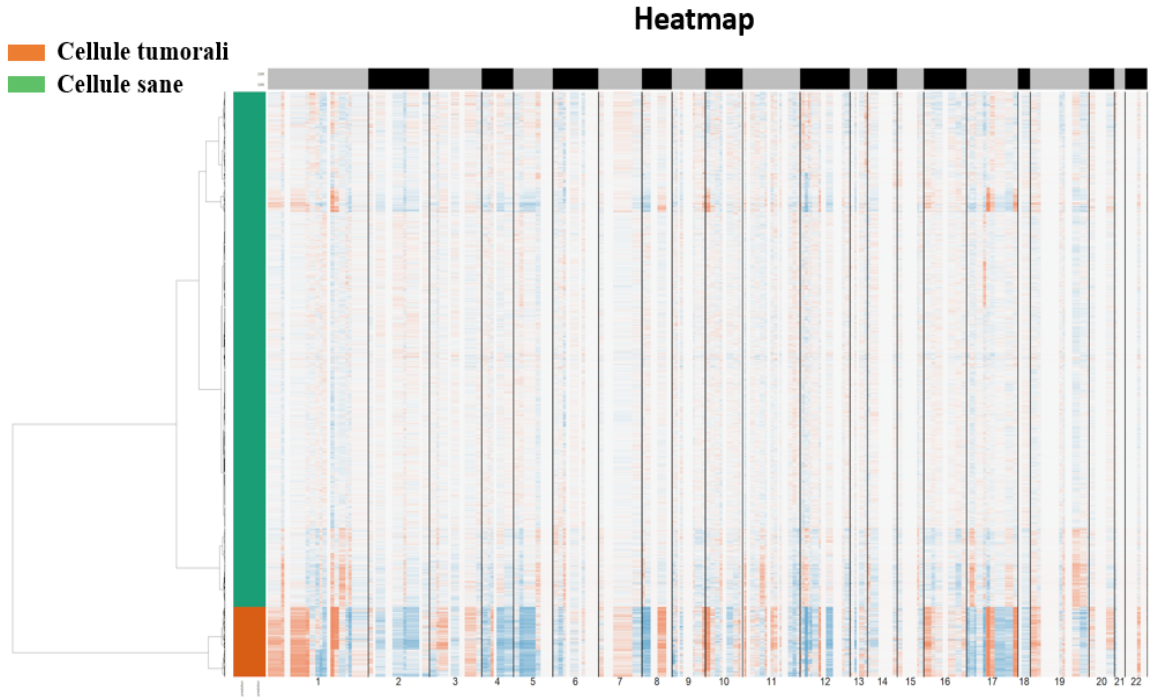


Figura 21 - Paziente 3

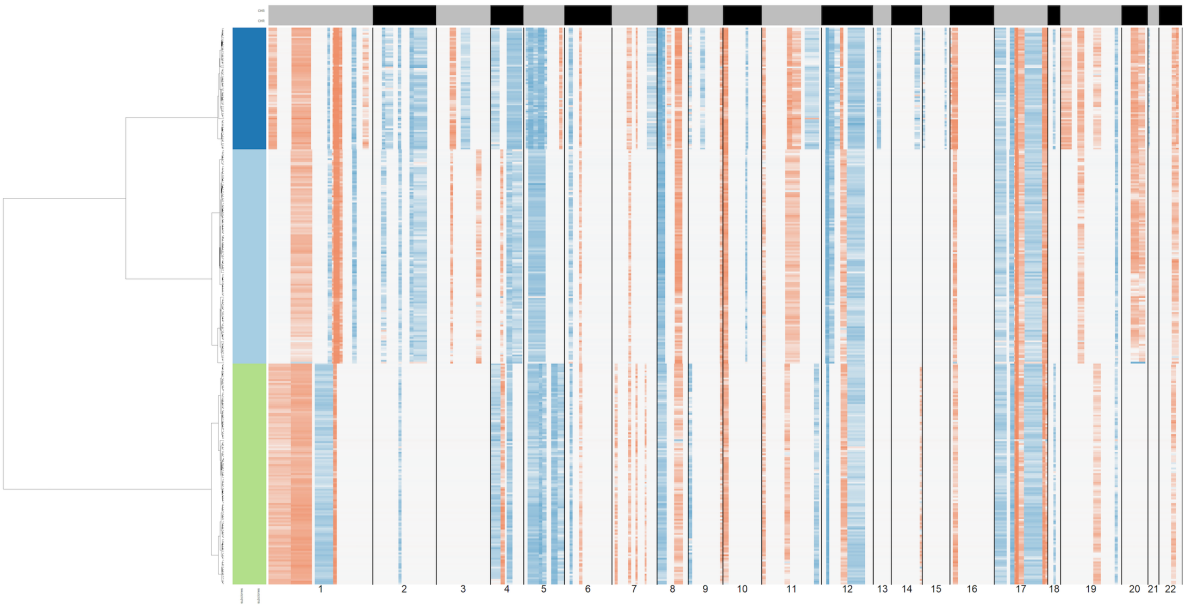


Figura 22 - Heatmap dei 3 subcloni tumorali nel Paziente 3

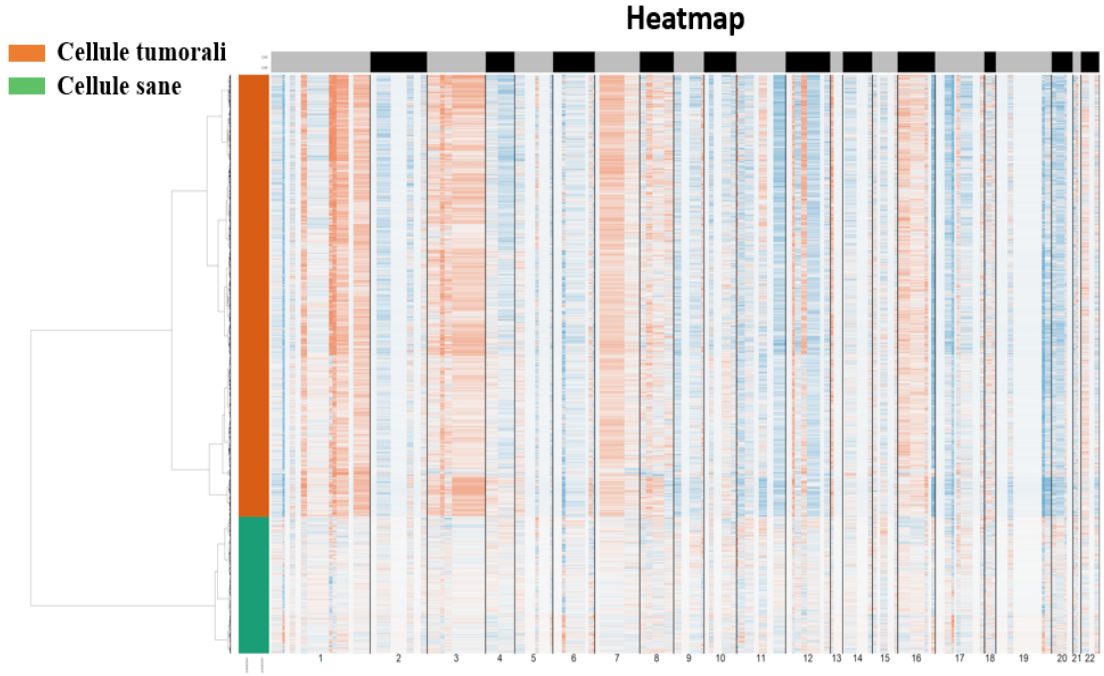


Figura 23 - Paziente 10

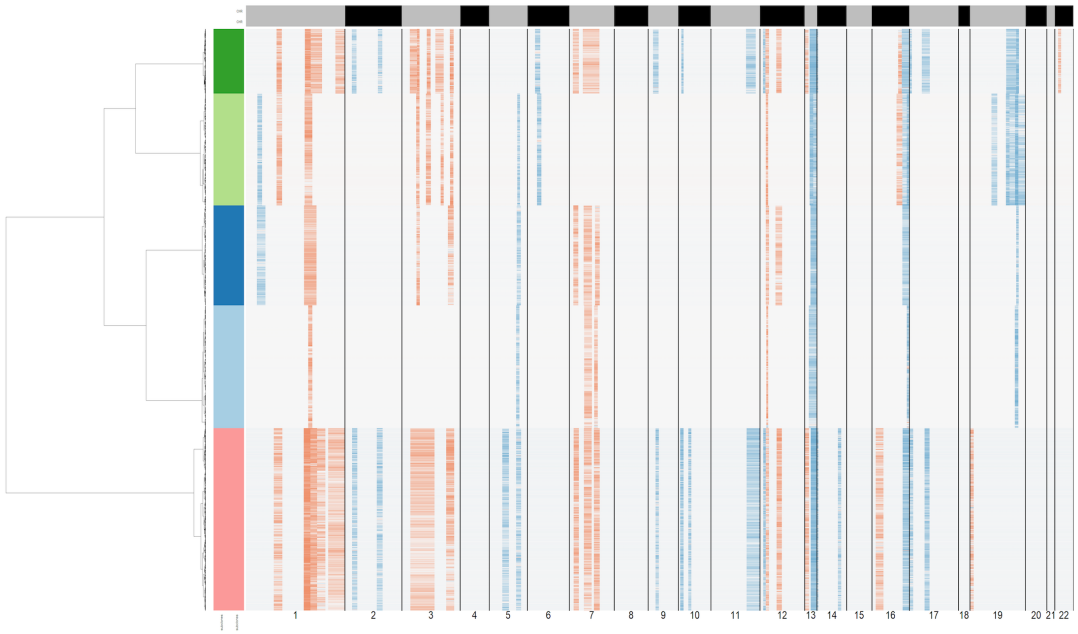


Figura 24 - Heatmap dei 5 subcloni tumorali nel Paziente 10

La predizione di cellule tumorali attraverso l'inferenza dei CNV non è risultata sempre concordante con quella ottenuta attraverso i marcatori tumorali in Cellenics.

Nel Paziente 3 è stata osservata una corrispondenza del 75% tra i due metodi di predizione (Figura 21).

Al contrario, nel Paziente 1 delle 615 cellule tumorali predette attraverso Cellenics, soltanto 29 (4,71%) di queste sono state predette come cellule tumorali da SCEVAN.

Osservando le Heatmap dei Pazienti 1, 3 e 10 (Figura 19, 21, 23) è possibile notare che, mentre nel Paziente 3 e nel Paziente 10 vi è una netta distinzione tra le cellule con CNV, quindi presunte tumorali, e le sane, nel Paziente 1 si ha una distribuzione omogenea dei CNV predetti che mascherano una chiara distinzione tra la popolazione sana e quella tumorale.

Questi risultati evidenziano che con la predizione dei CNV non sempre si riesce a distinguere in modo chiaro le cellule tumorali dalle sane e che, in questo caso, Cellenics si è rivelato il metodo più attendibile seppure il più laborioso.

Cellenics risulta infatti uno strumento "soggettivo" che dipende dai marker utilizzati, che non sempre si sono rilevati univoci, e dalla precisione dell'utente nello scovare le differenze tra un cluster e l'altro.

5. CONCLUSIONI

Il mondo della scRNA-seq è tanto vasto quanto ancora inesplorato; gli strumenti a disposizione necessitano di un costante affinamento e nuove integrazioni al fine di premettere un'analisi più esaustiva ed affidabile.

I limiti della tecnica sono legati principalmente all'enorme volume dei dati che derivano dagli esperimenti di scRNA-seq. Per l'elaborazione è necessario disporre di strumenti potenti (computer con una grande RAM) che ne permettano l'intera elaborazione.

In secondo luogo, a livello laboratoristico, è importante operare con la massima cautela e osservare i protocolli nel modo più rigido possibile. L'isolamento delle cellule, ad esempio, è tra i passaggi più delicati del processo: lo stress meccanico a cui queste sono sottoposte per il loro isolamento e per la successiva lisi, potrebbe portare alla morte cellulare o all'attivazione di geni dello stress che falserebbero l'analisi.

Un altro limite risiede nella scarsa profondità di lettura del sequenziamento dei dati di scRNA-seq: questo potrebbe mascherare la bassa espressione di alcuni geni che non vengono rilevati e la cui espressione viene erroneamente ritenuta uguale a zero.

Un altro problema che è stato riscontrato durante questo lavoro è che non sempre i diversi software utilizzano lo stesso input di dati per l'analisi o lo stesso formato; la conversione tra un formato e l'altro e la creazione del giusto file richiesto, oltre che allungare l'analisi, potrebbero essere fonte di errori a causa della perdita di alcuni dati o di una conversione imprecisa.

Dall'inizio del lavoro di tesi ad oggi, in un arco di tempo di circa un anno, diversi strumenti sono stati migliorati, incorporando nuove funzioni e consentendo analisi più sofisticate.

Anche se la strada potrebbe sembrare ancora lunga, questo dinamismo nel miglioramento dei software evidenzia l'impegno continuo della comunità scientifica nel perfezionare le

tecnologie esistenti e nell'introdurre nuovi strumenti, offrendo prospettive incoraggianti per il futuro.

Nel campo dell'oncologia, in particolare nel tumore al seno, la scRNA-seq potrebbe permettere l'individuazione di sottotipi tumorali nuovi che riflettono una diversa espressione genica e una prognosi altrettanto differente gettando le basi per una terapia personalizzata. Conoscere tutti i dettagli molecolari dell'intera cellula permetterebbe inoltre di operare più tempestivamente sia nella diagnosi che nella scelta del trattamento migliorando ulteriormente l'aspettativa di vita dei pazienti.

6. RIFERIMENTI

1. Jesinger, R.A., *Breast anatomy for the interventionalist.* Tech Vasc Interv Radiol, 2014. **17**(1): p. 3-9.
2. Khan, Y.S., A.O. Fakoya, and H. Sajjad, *Anatomy, Thorax, Mammary Gland*, in *StatPearls*. 2024: Treasure Island (FL) ineligible companies. Disclosure: Adegbenro Fakoya declares no relevant financial relationships with ineligible companies. Disclosure: Hussain Sajjad declares no relevant financial relationships with ineligible companies.
3. Shams, A., *Re-evaluation of the myoepithelial cells roles in the breast cancer progression.* Cancer Cell Int, 2022. **22**(1): p. 403.
4. Tsang, J.Y.S. and G.M. Tse, *Molecular Classification of Breast Cancer.* Adv Anat Pathol, 2020. **27**(1): p. 27-35.
5. Viale, G., *The current state of breast cancer classification.* Ann Oncol, 2012. **23 Suppl 10**: p. x207-10.
6. Watkins, E.J., *Overview of breast cancer.* JAAPA, 2019. **32**(10): p. 13-17.
7. Jovic, D., et al., *Single-cell RNA sequencing technologies and applications: A brief overview.* Clin Transl Med, 2022. **12**(3): p. e694.
8. Olsen, T.K. and N. Baryawno, *Introduction to Single-Cell RNA Sequencing.* Curr Protoc Mol Biol, 2018. **122**(1): p. e57.
9. Tang, F., et al., *mRNA-Seq whole-transcriptome analysis of a single cell.* Nat Methods, 2009. **6**(5): p. 377-82.
10. Su, M., et al., *Data analysis guidelines for single-cell RNA-seq in biomedical studies and clinical applications.* Mil Med Res, 2022. **9**(1): p. 68.

11. Huttenhower, C., et al., *Nearest Neighbor Networks: clustering expression data based on gene neighborhoods.* BMC Bioinformatics, 2007. **8**: p. 250.

12. Wu, S.Z., et al., *A single-cell and spatially resolved atlas of human breast cancers.* Nat Genet, 2021. **53**(9): p. 1334-1347.

13. Li, C., et al., *SciBet as a portable and fast single cell type identifier.* Nat Commun, 2020. **11**(1): p. 1818.

14. Pandey, P.R., J. Saidou, and K. Watabe, *Role of myoepithelial cells in breast tumor progression.* Front Biosci (Landmark Ed), 2010. **15**(1): p. 226-36.

15. Huang, Q., et al., *Evaluation of Cell Type Annotation R Packages on Single-cell RNA-seq Data.* Genomics Proteomics Bioinformatics, 2021. **19**(2): p. 267-281.

16. Jiang, A., et al., *ICARUS, an interactive web server for single cell RNA-seq analysis.* Nucleic Acids Res, 2022. **50**(W1): p. W427-W433.

17. Jiang, A., et al., *Delineation of complex gene expression patterns in single cell RNA-seq data with ICARUS v2.0.* NAR Genom Bioinform, 2023. **5**(2): p. lqad032.

18. De Falco, A., et al., *A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data.* Nat Commun, 2023. **14**(1): p. 1074.

19. Saha, S.K., et al., *KRT19 directly interacts with beta-catenin/RAC1 complex to regulate NUMB-dependent NOTCH signaling pathway and breast cancer properties.* Oncogene, 2017. **36**(3): p. 332-349.

20. Kwon, M.J., et al., *CD24 Overexpression Is Associated with Poor Prognosis in Luminal A and Triple-Negative Breast Cancer.* PLoS One, 2015. **10**(10): p. e0139112.