UNIVERSITÀ POLITECNICA DELLE MARCHE

DIPARTIMENTO SCIENZE DELLA VITA E DELL'AMBIENTE

Corso di Laurea Magistrale
Biologia Molecolare e Applicata

ANALISI DI TRASCRITTOMI A SINGOLA CELLULA DA BIOPSIE DI TUMORE ALLA CERVICE UTERINA

SINGLE-CELL TRANSCRIPTOME ANALYSIS FORM CERVICAL CANCER BIOPSIES

Tesi di Laurea Magistrale di
Alessia Selvaggi

Relatore
Chiar.mo Prof.
Francesco Piva

Sessione di Febbraio
Anno Accademico 2022/2023

Sommario

1. INTRODUZIONE

1.1 Cancro della cervice uterina

Il cancro della cervice uterina si sviluppa nella parte inferiore dell'utero, nota come collo dell'utero o cervice. La cervice uterina è la parte più bassa dell'utero ed è una struttura cilindrica composta da stroma ed epitelio. È costituita da due parti principali: l'endocervice ed ectocervice.

Queste due parti presentano diversi tipi di cellule: l'ectocervice è rivestita da cellule squamose, mentre l'endocervice è coperta da cellule colonnari.

I due tipi cellulari si incontrano nella zona di transizione o giunzione squamo-colonnare, ed è in questa zona che hanno origine la maggior parte dei tumori della cervice [1].

1.1.1 Generalità e fattori di rischio

Tra le donne di tutto il mondo il quarto tumore più comune è proprio quello della cervice uterina ed inoltre si stima che la maggior parte dei decessi mondiali (85%) si verificano nei paesi a basso e medio reddito [2].

Questo perché diverse ricerche indicano che condizioni economiche svantaggiate, scarsa igiene personale e sessuale, fumo di sigaretta, inizio precoce dell'attività sessuale e l'avere più partner sessuali sono considerati tra i fattori di rischio associati allo sviluppo del cancro alla cervice uterina.

L'HPV (Human Papilloma Virus) costituisce il principale fattore eziologico nel processo di sviluppo del cancro. Tuttavia, non tutte le infezioni da HPV nelle donne conducono al cancro cervicale. I genotipi ad alto rischio di HPV innescano la trasformazione di una

cellula normale in una lesione precancerosa e, successivamente, in una lesione invasiva. La patogenesi dell'infezione da HPV comporta la sovra espressione di oncoproteine virali, che possono inibire diverse proteine cellulari e influenzare processi biologici come la proliferazione cellulare, il ciclo cellulare e l'apoptosi [3].

Nella carcinogenesi, la maggior parte dei tumori cervicali si sviluppano a causa di un'infezione persistente dai tipi di HPV 16 e 18 (HPV ad alto rischio più diffusi) [3].

L'HPV è identificato nel 99,7% dei casi di carcinoma a cellule squamose e adenocarcinomi. Ci sono 15 ceppi oncogeni di HPV noti.

Attualmente, più della metà degli adulti tra i 20 e i 24 anni è infettata dall'HPV, poiché la maggior parte degli individui sessualmente attivi entra in contatto con il virus [4].

Il genoma dell'HPV codifica solamente otto proteine, ognuna delle quali svolge un ruolo nel ciclo vitale dell'HPV e nella trasformazione delle cellule ospiti in cellule cancerose.

Dopo l'infiltrazione virale delle cellule epiteliali basali, le particelle virali vengono rilasciate per agevolare l'integrazione del genoma virale in quello dell'ospite, causando danni a diverse vie cellulari e contribuendo alla progressione verso il cancro.

Le oncoproteine virali (proteine precoci) E6 e E7 giocano un ruolo cruciale nell'alterare la funzione cellulare dell'ospite.

E6 interagisce con p53, una proteina soppressore dei tumori, e la inibisce, disattivando così la sua funzione. Ciò interferisce con la capacità della cellula ospite di sottoporsi a riparazione del DNA, apoptosi, arresto della crescita e angiogenesi. L'attivazione di p53, che normalmente indurrebbe l'inibitore della chinasi ciclina-dipendente p21 per costringere le cellule a rimanere in arresto nella fase G1, viene compromessa in seguito all'infezione da HPV poiché E6 degrada p53. Ciò porta le cellule a entrare nella fase S del ciclo cellulare. L' oncoproteina E7 si lega al retinoblastoma(pRb), questa interazione

provoca il rilascio di E2F fattore di trascrizione che attiva la chinasi ciclina dipendente (CDK). Quindi il ciclo cellulare perde il controllo, le cellule infette si differenziano e proliferano intensamente, favorendo lo sviluppo di cellule displastiche anomale. La comprensione di tutti i meccanismi molecolari che sottendono all'insorgenza del cancro cervicale è fondamentale per individuare possibili marcatori molecolari. Attualmente, diverse evidenze supportano l'uso di biomarcatori per rilevare lesioni precancerose e il cancro cervicale in fase iniziale. Ad esempio, la sovra espressione di CDKN2A (p16) nelle prime fasi del cancro suggerisce una risposta dell'ospite nel disattivare e rilasciare la famiglia E2F. Un altro biomarcatore è Ki-67, antigene nucleare legato alla proliferazione cellulare, è presente durante tutte le fasi attive del ciclo cellulare ma assente nelle cellule quiescenti, ed è quindi utile per determinare la crescita di una popolazione cellulare.
La co-espressione di p16 e Ki-67 migliora l'accuratezza diagnostica nello screening del cancro cervicale [3].
È rilevante notare che, nonostante l'esposizione comune all'HPV, il sistema immunitario è in grado di eliminare il virus entro sei mesi nel 50% delle donne affette. Tuttavia, quando l'infezione persiste nell'epitelio metaplastico della zona di trasformazione cervicale, possono verificarsi cambiamenti cellulari displastici. Sebbene la displasia di basso grado (CIN1) tenda a regredire, c'è il rischio che progredisca verso una displasia di alto grado (CIN2 o CIN3). Il cancro cervicale si sviluppa quando le lesioni di alto grado si estendono oltre la membrana basale dell'epitelio cervicale.
Individuare precocemente le lesioni precancerose utilizzando il test di Papanicolaou (Pap test) è la strategia principale per la prevenzione del cancro [4].

Dal punto di vista istologico, i CIN vengono categorizzati in base al loro grado di gravità. Tuttavia, nelle condizioni di CIN e cancro, le cellule infettate da HPV manifestano un'alterazione displastica.

Il termine CIN 1, conosciuto anche come CIN di basso grado (LGCIN), indica una displasia lieve in cui un terzo dell'epitelio presenta segni di displasia. Quando sono coinvolti due terzi dell'epitelio, si parla di CIN 2 o displasia moderata. La displasia grave, identificata come CIN 3, è classificata quando più di due terzi dell'intero spessore dell'epitelio sono interessati. Le lesioni CIN 2 e CIN 3 vengono collettivamente classificate come CIN di alto grado (HGCIN) [3].

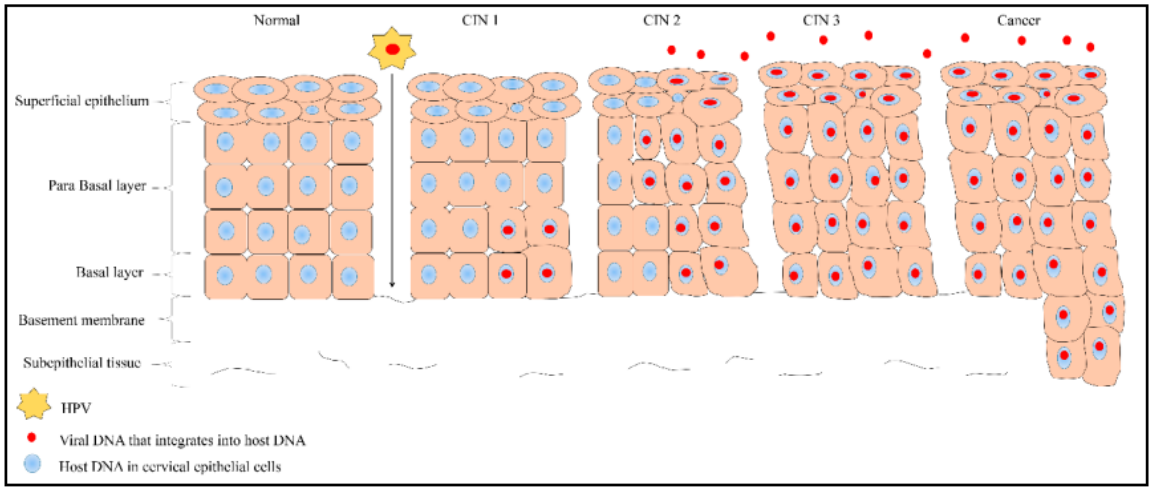

Figura 1. Ripartizione delle cellule epiteliali squamose in uno stato normale e di quelle infette dal virus del papilloma umano (HPV) durante condizioni fisiologiche, lesioni precancerose caratterizzate da displasia in vari gradi (lieve, moderata e grave, identificate come CIN 1, CIN 2 e CIN 3, rispettivamente) e nel contesto del cancro cervicale. Da Balasubramaniam, S.D., et al [3].

I tipi di tumore cervicale sono categorizzati in relazione alle cellule di origine e principalmente si distinguono in due forme: il carcinoma a cellule squamose e l'adenocarcinoma. Il carcinoma a cellule squamose ha origine dalle cellule che rivestono la

superficie dell'esocervice, mentre l'adenocarcinoma indica il cancro che si sviluppa dalle cellule ghiandolari dell'endocervice. Il cancro cervicale invasivo propagandosi può coinvolgere vagina, utero e organi limitrofi come vescica e retto. Inoltre, si diffonde attraverso i canali linfatici verso i linfonodi regionali con successiva propagazione ai nodi iliaci comuni e para-aortici. Le metastasi a distanza tramite il flusso ematogeno rappresentano un evento che si verifica in una fase più avanzata della malattia [3].

1.1.2 Stadiazione del cancro cervicale

La stadiazione del cancro cervicale ha subito una significativa revisione nel 2018. Il precedente sistema di stadiazione FIGO (Federation of Gynecology and Obstetrics) del 2009, per determinare lo stadio della malattia, si basava principalmente sull'esame clinico, la biopsia cervicale e alcuni test aggiuntivi. Il sistema di stadiazione FIGO del 2018 consente l'utilizzo, quando disponibile, di tecniche di imaging trasversale e risultati chirurgici anatomopatologici [1].

Stage	Description
I	The carcinoma is strictly confined to the cervix (extension to the uterine corpus should be disregarded)
IA	Invasive carcinoma that can be diagnosed only by microscopy, with maximum depth of invasion ≤5 mm[a]
IA1	Measured stromal invasion ≤3 mm in depth
IA2	Measured stromal invasion >3 and ≤5 mm in depth
IB	Invasive carcinoma with measured deepest invasion >5 mm (greater than Stage IA); lesion limited to the cervix uteri with size measured by maximum tumor diameter[b]
IB1	Invasive carcinoma >5 mm depth of stromal invasion and ≤2 cm in greatest dimension
IB2	Invasive carcinoma >2 and ≤4 cm in greatest dimension
IB3	Invasive carcinoma >4 cm in greatest dimension
II	The carcinoma invades beyond the uterus, but has not extended onto the lower third of the vagina or to the pelvic wall
IIA	Involvement limited to the upper two-thirds of the vagina without parametrial involvement
IIA1	Invasive carcinoma ≤4 cm in greatest dimension
IIA2	Invasive carcinoma >4 cm in greatest dimension
IIB	With parametrial involvement but not up to the pelvic wall
III	The carcinoma involves the lower third of the vagina and/or extends to the pelvic wall and/or causes hydronephrosis or nonfunctioning kidney and/or involves pelvic and/or para-aortic lymph nodes
IIIA	The carcinoma involves the lower third of the vagina, with no extension to the pelvic wall
IIIB	Extension to the pelvic wall and/or hydronephrosis or nonfunctioning kidney (unless known to be due to another cause)
IIIC	Involvement of pelvic and/or para-aortic lymph nodes (including micrometastases)[c], irrespective of tumor size and extent (with r and p notations)[d]
IIIC1	Pelvic lymph node metastasis only
IIIC2	Para-aortic lymph node metastasis
IV	The carcinoma has extended beyond the true pelvis or has involved (biopsy proven) the mucosa of the bladder or rectum. A bullous edema, as such, does not permit a case to be allotted to Stage IV
IVA	Spread of the growth to adjacent pelvic organs
IVB	Spread to distant organs

Figura2. Stadiazione FIGO cancro della cervice uterina da Bhatla, N., et al. [1].

1.1.3 Prevenzione e trattamento

La prevenzione del cancro alla cervice si realizza attraverso la diagnosi precoce ed il trattamento delle lesioni precancerose. Lo screening rappresenta una strategia fondamentale per il controllo globale del cancro cervicale, mentre la vaccinazione contro l'HPV è importante per prevenire la neoplasia impedendo l'infezione da HPV. Lo screening si concentra sulla rilevazione precoce delle lesioni precancerose come il CIN di altro grado e l'adenocarcinoma in situ, intervenendo efficacemente per prevenire lo sviluppo del cancro invasivo e ridurre i tassi di mortalità.

Tra le varie strategie di screening impiegate con successo troviamo la citologia convenzionale (Pap test), la citologia su base liquida (LBC) e il test HPV [1] .In presenza

di risultati anomali al Pap test o di lesioni visibili è indicata la colposcopia, attraverso una biopsia guidata da colposcopia o una procedura di escissione è possibile diagnosticare il cancro. Nonostante il cancro cervicale nelle fasi iniziali sia spesso asintomatico, potrebbe presentarsi con sintomi come il sanguinamento uterino. Le lesioni più avanzate invece possono provocare ostruzione del deflusso vescicale. Qualsiasi lesione sospetta visibile sulla cervice dovrebbe essere sottoposta a biopsia, indipendentemente dai risultati citologici [4].

Il trattamento del cancro cervicale è principalmente realizzato attraverso interventi chirurgici o trattamenti di radioterapia, spesso integrati con la chemioterapia.

La chirurgia è una scelta appropriata per i casi iniziali del cancro. A seconda dello stadio della malattia vengono considerate diverse opzioni chirurgiche come: isterectomia semplice o radicale. La decisione tra queste opzioni dipende dalla gravità e dall'estensione del cancro cervicale al momento della diagnosi [1]. La malattia metastatica avanzata viene trattata con la chemioterapia e si è anche osservato che l'aggiunta di Bevacizumab, un anticorpo monoclonale che agisce contro il fattore di crescita endoteliale, porta al miglioramento della sopravvivenza dei pazienti [4].

1.2 Single Cell RNA Sequencing (scRNA-seq)

Il cancro rappresenta una condizione eterogenea, è caratterizzato da una diversità genetica significativa tra i pazienti, tra i tumori primari e le metastasi, nonché all'interno dei singoli tumori.

L'scRNA-seq è una tecnica di Next Generation Sequencing (NGS) che si dimostra adatta per indagare a fondo l'eterogeneità tra tumori differenti e all'interno di singoli tumori, potrebbe consentire di effettuare diagnosi e prognosi più specifiche. In particolare, consente di studiare la complessità dei trascritti presenti all'interno di ciascuna cellula ed

inoltre permette di identificare la composizione di diversi tipi cellulari all'interno di tessuti, organi e organismi. Queste informazioni possono essere sfruttate per selezionare terapie appropriate in base al tipo specifico di tumore, permettendo un trattamento più mirato e personalizzato per il paziente [5].

Il termine "trascrittoma" si riferisce alla totalità dei trascritti presenti all'interno di una cellula. Comprende l'RNA messaggero, rRNA, tRNA e altri RNA non codificanti con funzioni regolatorie.

Utilizzando il sequenziamento dell'RNA di massa (RNA-seq) è possibile studiare in maniera accurata interi trascrittomi ma l'RNA estratto ed analizzato rappresenta una media di migliaia di trascritti cellulari presenti all'interno di un campione, per questo motivo variazioni rilevanti tra le cellule possono essere mascherate. La scRNA-seq invece offre la possibilità di esplorare nel dettaglio le proprietà biologiche di ogni singola cellula [6] fornendo informazioni uniche che contribuiscono alla comprensione di diverse malattie [5].

I tessuti complessi sono costituiti da una vasta gamma di tipi cellulari, e le informazioni presenti in ciascuna cellula spesso differiscono da quelle delle popolazioni cellulari adiacenti e addirittura da cellule dello stesso tipo. Proprio per questo motivo la scRNA-seq si configura come uno strumento potentissimo per analizzare l'espressione genica in ogni singola cellula. Ad esempio, nell'ambito della biologia del cancro, questa tecnica ha permesso ai ricercatori di identificare l'origine delle cellule tumorali in vari tipi di tumore, sono state individuate sottopopolazioni di cellule maligne con caratteristiche clinicamente rilevanti [6].

1.2.1 *Panoramica della metodologia*

Le fasi del protocollo di scRNA-seq includono l'isolamento e la cattura delle singole cellule, la trascrizione inversa (conversione del RNA in DNA complementare), l'amplificazione del cDNA e la preparazione della library.

I metodi di isolamento e cattura variano in base agli organismi e ai tessuti [5]. Prima di tutto, è fondamentale separare le cellule dal loro ambiente tissutale, bisogna rompere le connessioni tra le cellule stesse e la matrice extracellulare. Questo passaggio può essere effettuato attraverso la dissociazione meccanica o enzimatica del tessuto. Entrambi i procedimenti sono abbastanza complessi in quanto lo stress meccanico ed enzimatico potrebbero influenzare la vitalità delle cellule e alterare i programmi trascrizionali.

Per andare ad isolare le singole cellule possono essere utilizzate diverse tecniche, tra le più utilizzate vi sono le tecniche microfluidiche le quali consentono di lavorare con volumi liquidi molto bassi. Sono moltissime le piattaforme che permettono non solo l'isolamento automatico delle cellule, ma anche l'esecuzione automatica delle successive reazioni biochimiche, quali sintesi del cDNA e amplificazione.

Nelle piattaforme microfluidiche basate su goccioline la sospensione cellulare viene mescolata con sfere ognuna delle quali contiene un primer dotato di un codice a barre, in seguito le sfere e le singole cellule vengono racchiuse in goccioline all'interno di un'emulsione di olio.

Ad esempio, nel metodo Drop-seq la cellula all'interno della gocciolina viene lisata e il primer con il codice a barre univoco si lega all'mRNA, di conseguenza ogni cellula sarà identificata da un codice a barre univoco (chiamato cell barcode). In seguito, ogni trascritto all'interno di una cellula sarà contrassegnato da un Identificatore Molecolare Univoco (UMI). Gli UMI consistono in sequenze casuali di alcuni nucleotidi e sono

impiegati per identificare ogni molecola di mRNA all'interno di una cellula. Vengono amplificati in concomitanza con il cDNA e successivamente sequenziati. Le reads che condividono gli stessi UMI sono reads amplificate dello stesso frammento di mRNA originale; pertanto, valutando il numero di UMI associati a ciascun gene è possibile stimare il numero di trascritti di un particolare gene presente all'interno di una cellula.

Anche nella piattaforma 10xChromium Single Cell il principio utilizzato è simile, l'mRNA catturato viene retrotrascitto e il cDNA è amplificato tramite PCR.

Il protocollo inDrop rappresenta un ulteriore approccio basato su goccioline. All'interno di ogni gocciolina i frammenti di mRNA vengono amplificati linearmente con trascrizione in vitro (IVT) e retrotrascitto in cDNA [6].

La qualità della libreria scRNAseq è determinata da vari fattori, tra cui:

1 RUMORE TECNICO

-efficienza dalla cattura dell'mRNA

-tecnica utilizzata per l'amplificazione

-effetti batch sperimentali

2 RUMORE BIOLOGICO

-diversa natura dei campioni biologici (es: dimensione delle cellule, espressione genica)

-cambiamenti dinamici e casuali (es: stati del ciclo cellulare) [5]

Dato che i dati scRNA-seq contengono molta dispersione tecnica, che può essere introdotta in diverse fasi del processo, l'analisi dei dati scRNA-seq comporta un attento processo di filtraggio e controllo qualità [6].

L'analisi dei dati di scRNA-seq può essere divisa in tre fasi:

1 ELABORAZIONE DEI DATI GREZZI E CONTROLLO QUALITA'

2 ANALISI DI BASE: normalizzazione, features selection, annotazione del tipo cellulare, clustering e identificazione di geni marcatori.

3 ANALISI AVANZATA: inferenza della traiettoria, analisi comunicazione cellula-cellula (CCC), stima del flusso metabolico, previsione dell'attività dei fattori di trascrizione (TF) [7].

1.2.2 Pre-elaborazione dei dati

I processi di pre-elaborazione dei dati includono il demultiplexing, la mappatura, la quantificazione dei trascritti e il controllo qualità.

A partire dai dati FASTQ grezzi i codici a barre delle cellule e gli UMI vengono inizialmente aggiunti come tag a ciascuna reads di cDNA. Successivamente con il demultiplexing le sequenze fastaq vengono separate in base ai cell barcode e UMI assegnate alle singole cellule.

In seguito, i cDNA vengono allineati al genoma di riferimento o mappati, a seconda del flusso di lavoro adottato. In fine si ottiene una matrice grezza (UMI counts), cioè una

I ricercatori hanno la possibilità di creare i propri processi di pre-elaborazione combinando singoli metodi o seguendo percorsi predefiniti suggeriti [8].

1.2.3 Controllo qualità e normalizzazione

I livelli di espressione tra le cellule risultano non comparabili a causa di errori sistemici o interferenze tecniche. Tra le problematiche tecniche rientrano l'efficienza di cattura dell'mRNA, la trascrizione inversa, l'amplificazione del cDNA e la profondità di sequenziamento.

Tra gli effetti tecnici indesiderati vi è anche l'effetto batch, esso non può essere evitato in quanto i dati scRNA-seq possono provenire da diverse piattaforme di sequenziamento e la

creazione delle librerie può essere affidata a individui diversi i quali utilizzano differenti lotti di reagenti.

Tra i fattori biologici che influenzano, ci sono le dimensioni cellulari e le fasi del ciclo cellulare. La fase del ciclo cellulare può rappresentare un ostacolo nella caratterizzazione del segnale biologico di interesse; quindi, è di vitale importanza correggere l'effetto del ciclo cellulare.

Durante il controllo qualità è di grande importanza eliminare il segnale di fondo causato dall'RNA ambientale per migliorare le analisi successive e l'interpretazione biologica [5].

I parametri utilizzati per il controllo qualità sono:

-numero totale di geni espressi per cellula

-numero totale di UMI per cellula

-percentuali di geni mitocondriali (questa metrica consente di rilevare la presenza di una considerevole quantità di contaminazione mitocondriale derivante da cellule in fase di morte o morenti) e geni delle proteine ribosomiali in ciascuna cellula.

Non esiste un'impostazione standard assoluta per le soglie di filtro, poiché queste dipendono dal tipo di tessuto in esame e dalle condizioni patologiche. È fondamentale regolarle in modo flessibile, considerando lo stato specifico del tessuto e le sue caratteristiche [5]. Molto importante è la rimozione dei Doppietti, ovvero informazioni sul trascrittoma di singola cellula che riflettono più di una cellula. È probabile che i doppietti confondano l'analisi a valle, troppi geni rilevati potrebbero indicare la presenza di doppietti.

La matrice di espressione originale non può essere direttamente utilizzata per l'analisi a valle poiché i livelli di espressione tra le cellule non sono comparabili a causa di errori sistematici o rumori tecnici.

La normalizzazione è finalizzata a contrastare tali interferenze al fine di garantire la comparabilità tra le singole cellule [5, 7].

1.2.4 *Features selection*

È importante la selezione dei geni biologicamente rilevanti per l'analisi a valle.

I set di dati presentano una complessità dimensionale elevata.

La maggior parte dei geni individuati nelle cellule appartiene alla categoria dei geni housekeeping. Poiché questi geni non mostrano variazioni sostanziali nei livelli di espressione tra le cellule, la loro presenza può costituire un ostacolo alla corretta identificazione dei segnali biologici autentici. Pertanto, una selezione accurata dei geni da includere nell'analisi è essenziale per garantire una rappresentazione più precisa e significativa delle differenze biologiche tra le cellule.

In contrasto, i geni altamente variabili facilitano l'analisi, contribuendo a una notevole riduzione del carico computazionale. Questa classe di geni include quelli in grado di distinguere diversi tipi cellulari, e la qualità di tali geni gioca un ruolo cruciale nell'assicurare l'accuratezza del processo di clustering.

1.2.5 *Riduzione della dimensionalità*

Dopo aver identificato i geni altamente variabili, la dimensione dei dati rimane considerevole. Per mitigare ciò, si ricorre all'analisi delle componenti principali (PCA), che estrae componenti principali in base alla loro rilevanza. Viene in seguito impiegato l'incorporamento dei vicini stocastici distribuiti in t (t-SNE) per un'approssimazione efficace o l'approssimazione delle varietà uniformi (UMAP) al fine di ridurre la dimensione delle matrici di espressione [7].

1.2.6 Clustering

Uno dei passaggi chiave è l'identificazione delle sottopopolazioni cellulari attraverso processi di clustering.

La scelta delle metriche di somiglianza o distanza riveste un ruolo cruciale nel processo di clustering.

Dopo il clustering è importante l'annotazione del tipo cellulare, ovvero l'assegnazione delle identità cellulari alle sottopopolazioni cellulari. Principalmente vengono utilizzati due metodi di assegnazione del tipo cellulare, in base all'espressione genica: l'annotazione manuale e l'annotazione computazionale.

Il processo di annotazione manuale dei tipi cellulari è noto per essere laborioso e potenzialmente soggettivo, per questo motivo si è spinto lo sviluppo di strumenti computazionali per l'annotazione automatica dei tipi cellulari. Questi sono basati su geni marcatori, sulla disponibilità di marcatori specifici per un'ampia varietà di tipi cellulari nei tessuti di esseri umani e topi presenti in database pubblici come Cell Marker e PanglaoDB o nella letteratura scientifica.

I metodi computazionali si basano su un trascrittoma di riferimento e utilizzano dataset di scRNA-seq etichettati per il tipo di cellula come input per l'annotazione del tipo di cellula, cercando la migliore correlazione tra i dati interrogati e i dati di riferimento. Strumenti noti appartenenti a questa categoria includono: scmap, scMatch e SingleR [7].

1.2.7 Analisi di arricchimento funzionale

Per comprendere il significato biologico di specifiche popolazioni cellulari, è essenziale condurre analisi di arricchimento funzionale su insiemi mirati di geni differenzialmente espressi. Allo stesso tempo, strumenti quali inferenza della traiettoria, stima pseudo-

temporale e modellizzazione della velocità dell'RNA risultano utili per rilevare caratteristiche molecolari e meccanismi regolatori coinvolti in processi come differenziazione e attivazione cellulare.

Questi approcci integrati consentono di acquisire una prospettiva più dettagliata dei fenomeni biologici in corso, facilitando l'identificazione dei determinanti molecolari che guidano la diversificazione e il comportamento delle cellule in contesti specifici.

In accordo con le nozioni biologiche relative all'inizio di una traiettoria di sviluppo o di transizione di stato, è possibile organizzare le cellule in un ordine pseudo-temporale, ciò permette di dedurre la sequenza temporale delle cellule, consentendo la scoperta di tipi cellulari meno comuni e stati cellulari nascosti.

Per cogliere la dinamica del trascrittoma è fattibile sfruttare anche la velocità dell'RNA, la quale deriva dalla relazione tra trascritti maturi e non maturi all'interno della stessa cellula. Valutando il rapporto tra di essi e le variazioni nell'espressione genica durante le fasi di cambiamento di stato, è possibile stabilire la direzione delle transizioni cellulari [7].

1.2.8 *Fasi del ciclo cellulare*

Molto importante è anche comprendere in quale fase specifica del ciclo cellulare si trovi ciascuna cellula.

Per esempio, il programma Seurat (https://github.com/satijalab/seurat) per la valutazione del ciclo cellulare utilizza una funzione che assegna un punteggio numerico a ogni cellula in relazione all'espressione di geni che agiscono come marcatori per le fasi G2, M e S del ciclo cellulare [5].

1.2.9 Comunicazioni cellula-cellula (CCC)

La comunicazione tra le cellule svolge un ruolo importante per l'origine e l'evoluzione di patologie. Un esempio è rappresentato dagli ecosistemi complessi nei microambienti tumorali, dove una comunicazione anomala tra le cellule favorisce la proliferazione del tumore. Questa comunicazione cellula-cellula si basa su interazioni tra ligandi e recettori, le cui informazioni sono contenute in database dedicati. I dati derivanti dal sequenziamento a singola cellula vengono elaborati e successivamente integrati con le note interazioni ligando-recettore. Si calcolano punteggi specifici per il campione, offrendo una valutazione quantitativa del potenziale di interazione cellulare.

1.2.10 Fattori di trascrizione (TF)

I fattori di trascrizione svolgono un ruolo cruciale nella modulazione dell'espressione genica e partecipano a diversi processi fisiologici e patologici negli esseri umani. Esistono database di annotazione dedicati ai fattori di trascrizione che coprono la maggior parte dei fattori di trascrizione umani. Questi database possono essere impiegati per costruire reti specifiche di regolazione trascrizionale in base al tipo di cellula considerato, consentendo l'identificazione di fattori di trascrizione sovra-regolati o geni bersaglio dei fattori di trascrizione espressi in modo differenziale.

1.2.11 Analisi metabolica

La deregolazione metabolica emerge come caratteristica distintiva di diverse patologie, compreso il cancro. Pertanto, l'utilizzo dei dati provenienti da scRNA-seq si presentano come un mezzo per osservare e seguire le variazioni nell'espressione genica di geni metabolici cruciali durante i processi patologici [7].

Tutte le fasi di analisi riportate richiedono software specifici.

1.3 Inferenza delle variazioni del numero di copie (CNV)

La diversità genetica umana si estende da variazioni a livello di singolo nucleotide fino a eventi cromosomici di ampia portata, le differenze nei genomi umani sono spesso il risultato di variazioni strutturali.

Inizialmente, le variazioni strutturali erano definite come inserzioni, delezioni e inversioni di dimensioni superiori a 1 kb. Con il progressivo sequenziamento dei genomi umani, la definizione delle varianti strutturali (SV) e delle varianti del numero di copie (CNV) si è estesa per includere anche eventi molto più piccoli, come quelli superiori a 50 bp.

Il termine "CNV" quindi si riferisce alle variazioni del numero di copie del DNA, le quali contribuiscono allo sviluppo di malattie, compresi i tumori [9].

I CNV possono svolgere un ruolo importante nello sviluppo e nella progressione dei tumori. Queste alterazioni genomiche possono influenzare l'espressione genica, portare all'attivazione di oncogeni (geni coinvolti nello sviluppo del cancro) o alla disattivazione di geni soppressori tumorali. Questi eventi possono contribuire alla trasformazione delle cellule normali in cellule tumorali.

L’identificazione dei CNV costituisce una componente essenziale nell'analisi genomica dei tumori, con un notevole potenziale per migliorare la diagnosi e le scelte terapeutiche del cancro [10].

2. SCOPO DELLA TESI

Studio e valutazione di vari software utilizzati per l'analisi dei dati derivanti dal sequenziamento dell'RNA a singola cellula.

L'obiettivo specifico è concentrarsi sullo studio di diversi software dedicati all'analisi a livello cellulare. Per esaminare i dati derivanti dal sequenziamento a singola cellula, sono stati impiegati diversi strumenti al fine di confrontare i risultati ottenuti attraverso diversi flussi di lavoro. L'obiettivo era individuare similitudini e discrepanze tra le diverse metodologie utilizzate.

3. MATERIALI E METODI

3.1 Origine dei Dati

I set di dati utilizzati nell'analisi provengono dall' articolo: Single-cell transcriptomics reveals cellular heterogeneity and molecular stratification of cervical cancer. Nello studio, riportato nell'articolo, la libreria di RNA a cellula singola è stata creata attraverso l'utilizzo della piattaforma 10X Genomics Chromium Single-cell. I dati ottenuti mediante il sequenziamento dell'RNA a singola cellula sono stati depositati nel database ArrayExpress con il numero di accesso E-MTAB-11948.

(https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-11948)

3.2 Caratteristiche dei Dati Grezzi

Per ogni campione abbiamo scaricato tre file di dati di scRNA-seq:

1. barcodes.tsv: elenco dei codici a barre identificativi delle cellule
2. features.tsv: elenco di caratteristiche/geni (UMI) rilevati
3. matrix.mtx: una matrice sparsa contente in codice ASCII contente i valori di espressione di ogni singolo gene in ogni singola cellula.

Da questi tre file si ottiene una matrice contenente il conteggio dei trascritti rilevati, dove i geni sono disposti lungo le righe e i codici a barre/cellule sono disposti lungo le colonne. Questi file rappresentano i dati grezzi sull'espressione genica del campione.

3.3 Dettagli Clinici

Tutte le pazienti riportate nell'articolo sono state sottoposte ad un intervento di isterectomia radicale e presentano un carcinoma clinicamente visibile confinato alla cervice uterina non> di 4cm nel diametro maggiore (stadio IB1).

Le informazioni cliniche, tra cui età, stato mestruale, stadio FIGO, tipo istologico, stato HPV e trattamento, sono dettagliate nella Tabella 1 [11].

Patients	Age (year)	Menstrual status	FIGO stage	Histological type	HPV status	Treatment
P1	48	menstruating	IB1	Squamous cell carcinoma	HPV16(+)	Radical hysterectomy
P2	50	menopause	IB1	Squamous cell carcinoma	HPV16, 33(+)	Radical hysterectomy
P3	51	menopause	IB1	Squamous cell carcinoma	HPV16(+)	Radical hysterectomy

Tabella 1. Informazioni cliniche dei pazienti.

3.4 Barcodes Corrispondenti alle Cellule Tumorali

Dai materiali supplementari dell'articolo, sono stati estratti anche i barcodes corrispondenti alle cellule tumorali individuate dagli autori in base a specifici marcatori presenti in letteratura. La presenza di questi barcodes ha consentito di confrontare la concordanza tra le cellule tumorali rilevate dagli autori e quelle identificate attraverso la mia analisi.

3.5 Strumento di Elaborazione Dati: Cellenics

Per elaborare i dati grezzi ricavati dall'articolo abbiamo utilizzato un programma in rete: Cellenics.

Dopo aver caricato i dati abbiamo iniziato l'elaborazione, la sequenza di elaborazione dati di Cellenics si compone di 7 passaggi nei quali vengono impiegati filtri per eliminare informazioni indesiderate o di scarsa qualità da ciascun campione. I primi cinque passaggi

sono dedicati all'analisi dei singoli campioni, mentre il sesto passo mira a combinare diversi set di dati per eliminare variazioni dovute a batch e a ridurre la complessità dimensionale. Infine, nel settimo passaggio, vengono configurati metodi di incorporamento, come UMAP o t-SNE, seguiti dall'applicazione di tecniche di clustering per ottenere una visione più chiara e strutturata dei dati.

Una volta che l'elaborazione è stata completata con successo, i dati che sono stati sottoposti a filtraggio e integrazione, unitamente ai risultati derivanti dal clustering, sono pronti per essere esaminati e rappresentati visivamente nei moduli successivi di Cellenics, come quelli dedicati all'esplorazione dei dati e alla creazione di grafici e tabelle.

3.5.1 Passaggi di controllo qualità

Nelle analisi effettuate, dopo il caricamento dei dati abbiamo iniziato l'elaborazione:

STEP 1: Filtro classificatore

Il filtro classificatore mira a escludere le goccioline vuote e a trattenere quelle che contengono cellule. Per raggiungere questo obiettivo, sfrutta il metodo "emptydrops" per calcolare il tasso di falsi positivi (FDR), un parametro statistico che indica la probabilità che una goccia sia vuota. Il valore predefinito del FDR è 0,01 per tutti i campioni, e vengono conservate solo le goccioline con un FDR inferiore a 0,01. Di conseguenza, in questa fase, vengono mantenute per l'analisi successiva le goccioline con un basso FDR, mentre quelle con un FDR elevato vengono escluse dall'analisi successiva.

STEP 2: Filtro del contenuto mitocondriale

Il filtro del contenuto mitocondriale ha lo scopo di eliminare le goccioline che contengono cellule morte o sono di scarsa qualità, valutando la percentuale di trascritti mitocondriali in ciascuna gocciolina e stabilendo una soglia appropriata. Le goccioline con un livello di contenuto mitocondriale che supera questa soglia vengono escluse dall'analisi successiva.

La soglia predefinita per la proporzione di geni mitocondriali è calcolata individualmente per ogni campione. Tipicamente, l'intervallo di cut-off si situa tra il 10% e il 50% delle letture mitocondriali per cellula, con il cut-off predefinito in Cellenics fissato a 3 deviazioni assolute mediane al di sopra della mediana.

STEP 3: Numero di geni vs UMI

Il filtro Numero di geni vs UMI si basa sull'assunto che il numero di trascrizioni uniche, rappresentato dal conteggio UMI, mostri una relazione lineare con il numero di geni. Le goccioline che deviano da questa relazione lineare possono essere suddivise in due categorie:

Goccioline con un elevato numero di geni ma pochi UMI, indicando un'efficienza di amplificazione delle trascrizioni non ottimale.

Goccioline con pochi geni ma molti UMI, suggeriscono un'eccessiva amplificazione delle poche trascrizioni presenti.

Questo filtro visualizza i dati attraverso un grafico a dispersione, dove il numero di conteggi genici è rappresentato su una scala logaritmica rispetto al numero di UMI anch'esso su scala logaritmica.

STEP 4: Rimozione doppietti

Il filtro Doppietto valuta la probabilità che una gocciolina sia un duplicato e successivamente elimina le cellule con una probabilità significativamente alta di essere effettivamente un duplicato. L'analisi della probabilità si avvale dell'algoritmo scDblFinder. In questo processo, viene applicata una soglia rigida: tutte le goccioline con una probabilità superiore a tale soglia vengono escluse mediante il filtro.

STEP 5: Integrazione dei dati

La fase di integrazione dei dati mira a eliminare gli effetti batch e a ridurre la complessità dei dati. Gli effetti batch rappresentano variazioni causate da differenze nelle condizioni sperimentali, introducendo rumore che può distorcere la vera variazione tra i campioni. Senza affrontare questi effetti batch, il confronto diretto tra campioni potrebbe essere influenzato da rumori diversi. La rimozione degli effetti batch consente il confronto e l'analisi combinata di campioni provenienti da diverse sessioni, minimizzando gli errori. In sostanza, correggendo gli effetti batch, ci si assicura che l'analisi successiva si focalizzi sulle differenze biologiche reali tra i campioni, escludendo variazioni irrilevanti da campione a campione o da lotto a lotto.

Prima dell'integrazione, si applica una normalizzazione specifica a ciascun campione. Esistono diverse modalità per effettuare questa normalizzazione, e in Cellenics il metodo di default adottato è il LogNormalize.

La riduzione della dimensionalità semplifica la complessità del dataset mantenendo al contempo la variazione.

La PCA viene utilizzata per effettuare una riduzione dimensionale dei dati grezzi, agendo come una fase preliminare. I risultati ottenuti vengono poi sottoposti ad altri algoritmi di riduzione dimensionale, come UMAP o t-SNE, per proiettare i dati in due dimensioni.

STEP 6: Configura l'incorporamento

Nella fase finale del modulo di Elaborazione dati, l'incorporamento bidimensionale dei dati integrati subisce una riduzione ulteriore.

Cellenics offre due approcci per visualizzare l'incorporamento dei dati: UMAP e t-SNE. Dopo la creazione dell'incorporamento, i punti dati incorporati vengono organizzati in gruppi e colorati in base alle annotazioni dei cluster. Il clustering, che rappresenta il processo di aggregazione di celle con somiglianze significative, è eseguito utilizzando il

metodo Louvain come impostazione predefinita in Cellenics. I cluster sono distinti mediante colori e numeri, semplificando l'identificazione e l'esplorazione durante le fasi successive dell'analisi in Cellenics.

L'output del clustering può essere personalizzato regolando la risoluzione del clustering tramite il menu delle impostazioni dedicate al clustering.

3.5.2 *Dot plot*

In Cellenics, il dot plot rappresenta la percentuale di cellule che esprimono i geni selezionati. La dimensione del punto indica la percentuale di espressione genica in tutte le cellule di un cluster specifico: punti più piccoli indicano una minore percentuale di espressione, mentre punti più grandi indicano una maggiore espressione del gene in quel cluster. Il colore del punto riflette il livello di espressione del gene.

Permette di esaminare l'espressione di geni marcatori personalizzati identificati, e in base all'espressione di questi marcatori, abbiamo individuato specifici cluster. Attraverso questa funzione di Cellenics, abbiamo integrato i marcatori tumorali da noi correlati ai tumori.

3.6 Marcatori tumorali

I marcatori impiegati nell'analisi sono specifici marcatori utilizzati in immunoistochimica, oppure sono marcatori ricercati e individuati in letteratura scientifica.

3.7 Analisi dei Doppietti: DoubletFinder e ICARUS

L'analisi dei dati di scRNA-seq risulta complessa a causa della presenza di artefatti tecnici noti come doppietti, che consistono in informazioni sul trascrittoma di singola cellula che riflettono più di una cellula. Per superare questa sfida, DoubletFinder identifica le cellule che mostrano caratteristiche di espressione simili a quelle presenti nei doppietti simulati, consentendo così di individuare numerosi doppietti.

DoubletFinder è uno strumento avanzato progettato per riconoscere i doppietti esclusivamente attraverso l'analisi dei dati sull'espressione genica. La sua capacità predittiva si basa sulla vicinanza nello spazio dell'espressione genica di ciascuna cellula reale rispetto ai doppietti simulati, i quali vengono creati mediante la combinazione dei profili trascrizionali di coppie di cellule selezionate casualmente [12].

Per predire i doppietti abbiamo utilizzato ICARUS, un'applicazione online per l'analisi di RNA-seq a singola cellula (scRNA-seq) accessibile direttamente attraverso il browser web. La capacità operativa di ICARUS si fonda sul pacchetto R chiamato Seurat, il quale costituisce un insieme di strumenti estremamente completo per l'analisi dei dati di RNA-seq a singola cellula (scRNA-seq), integrando le metodologie più avanzate del settore. Fornisce all'utente la possibilità di gestire ogni singolo passo per personalizzare l'analisi in base al dataset di interesse. Le rappresentazioni grafiche durante ciascuna fase assicurano un'interpretazione agevole e coerente.

Per la rimozione dei doppietti ICARUS utilizza DOUBLET FINDER.

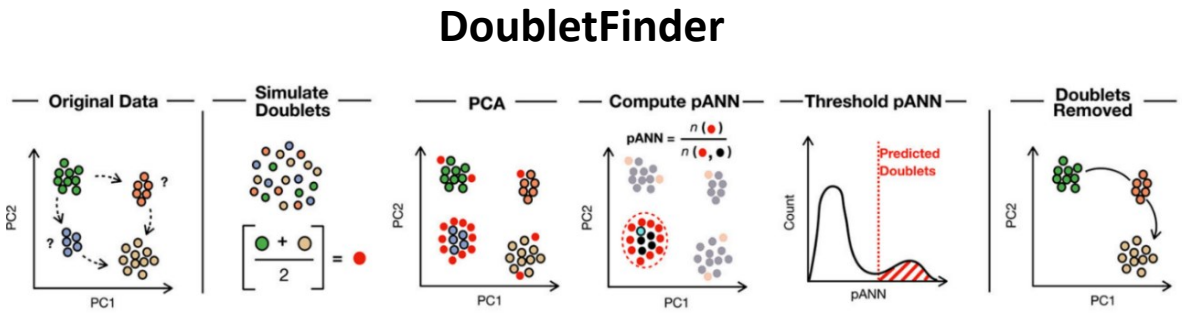

Figura 3. Doublet Finder da Christopher S. McGinnis et al., 2019 [13]

3.8 Confronto con Barcodes delle Cellule Tumorali

Per valutare l'accuratezza del programma di predizione dei doppietti ICARUS, è stato condotto un confronto tra i doppietti predetti dal software e i barcodes corrispondenti alle cellule tumorali. Per valutare se il programma ICARUS ha eliminato erroneamente cellule

tumorali considerandole doppietti, è stato eseguito un confronto tra i barcodes delle cellule tumorali noti dall'articolo di riferimento e i doppietti predetti dal programma. I barcodes delle cellule tumorali che corrispondevano ai doppietti predetti sono stati identificati e analizzati. Successivamente, ho ampliato l'analisi includendo un confronto tra i doppietti predetti da ICARUS e i barcodes delle cellule tumorali da me individuati utilizzando Cellenics.

3.9 Annotazione cellulare

Per l'identificazione dei tipi cellulari sono stati utilizzati diversi strumenti.

3.9.1 Strumenti di Annotazione Cellulare

SINGLE R

Disponibile pubblicamente come pacchetto R, dimostra un'elevata efficienza nella predizione di tipi cellulari rari, nella gestione di un ampio numero di classi di tipi cellulari e nel distinguere tra tipi cellulari che presentano notevole somiglianza.
Sfrutta dataset trascrittomici di riferimento di tipi cellulari, non confronta direttamente i profili di espressione genica delle singole cellule, crea un profilo "pseudo-bulk" rappresentativo per ciascun tipo cellulare. Questo profilo è ottenuto dalla media delle espressioni geniche delle cellule appartenenti allo stesso tipo cellulare.
L'utilizzo di un riferimento pseudo-bulk semplifica l'analisi e riduce il rumore associato a singole cellule.
SingleR fa uso di un riferimento pseudo-bulk di RNA-seq al fine di correlare i profili di espressione medi dei tipi cellulari con le singole cellule presenti nei dati di query (dati delle singole cellule che vengono analizzati o confrontati con un riferimento per identificare o annotare i tipi cellulari). Durante questo processo, si avvale di geni altamente

variabili per trovare la migliore corrispondenza tra i profili di espressione delle singole cellule e i profili medi dei tipi cellulari [14].

SCIBET

SciBet online può essere utilizzato dagli utenti che hanno la possibilità di caricare i loro insiemi di dati al fine di effettuare la classificazione.

SciBet emerge come uno strumento di grande utilità nel campo della trascrittomica a singola cellula, rispondendo in modo efficace alla necessità di analizzare con precisione e efficienza set di dati eterogenei e estesi. La sua abilità nel considerare sia la similarità relativa che quella assoluta si traduce in un notevole potenziamento delle performance globali del metodo.

Riesce a mantenere un equilibrio tra elevata precisione e basso tasso di falsi positivi introducendo un set di dati nullo come punto di riferimento alternativo per le cellule che presentano tipologie non ancora rappresentate nei dati esistenti [15].

SCGEATOOL(PANGLAODB)

ScGEAToolbox è un tool di MATLAB per l'analisi dei dati di scRNA-seq, che contiene un insieme completo di funzioni.

La funzione che permette l'annotazione cellulare si basa sulla disponibilità di marcatori specifici del tipo cellulare in un database pubblico: PanglaoDB.

E' un database che comprende analisi derivanti da oltre 1054 esperimenti su singole cellule contenente 6000 marcatori che permettono di distinguere diversi tipi cellulari [16].

3.9.2 Coerenza nell'Annotazione

Successivamente, è stata creata una tabella contenente i barcodes delle cellule e l'annotazione cellulare assegnata da ciascun programma. Per ogni barcodes, è stato

registrato il tipo cellulare assegnato da ciascun programma partecipante. La tabella risultante è stata analizzata per determinare se esiste una concordanza tra i programmi nella definizione del tipo cellulare per ciascun barcodes.

Questo approccio mira a identificare eventuali divergenze nelle annotazioni cellulari fornite dai programmi, consentendo una valutazione critica della consistenza delle assegnazioni cellulari su un set di barcodes cellulare già esistente.

3.10 Analisi del Ciclo Cellulare con scGEATool di MATLAB

3.10.1 Strumento di Analisi: scGEATool

È uno strumento per l'analisi dei dati scRNA-seq implementato in MATLAB. Comprende un set completo di funzioni per normalizzare i dati, per il filtraggio di geni e cellule, l'identificazione di geni altamente variabili (HVGs), la correzione degli effetti di batch, la riduzione della dimensionalità, la visualizzazione dei dati, il raggruppamento delle cellule, l'analisi della traiettoria e la costruzione di reti [17].

Il tool di Matlab ha consentito di determinare la posizione delle cellule nel ciclo cellulare.

3.11 Analisi delle Variazioni del Numero di Copie di DNA nelle Cellule Tumorali

3.11.1 Software Utilizzato

L'analisi per inferire i CNV a partire dalla matrice grezza è stata condotta utilizzando il programma SCEVAN (Single-Cell Estimation of Variations in Allelic Copy Number), un approccio avanzato basato su clustering, segmentazione variazionale e classificazione per identificare sottocloni tumorali e analizzare la eterogeneità clonale nei tumori.

L'obiettivo principale dell'algoritmo è distinguere tra cellule maligne e non maligne, nonché identificare sottocloni tumorali in base alle variazioni del numero di copie inferite. L'algoritmo sfrutta l'idea che le cellule appartenenti allo stesso sottoclone condividono gli stessi punti di interruzione di copie genomiche. Utilizzando un approccio variazionale e un algoritmo di segmentazione, SCEVAN analizza i cluster di cellule tumorali e identifica i sottocloni, classificando le loro alterazioni genomiche specifiche e condivise.

3.11.2 *Flusso di lavoro di SCEVAN:*

1. *Preelaborazione dei dati*: Inizia con la matrice grezza di conteggio delle cellule ottenute dai dati di scRNA-seq. Questa matrice viene sottoposta a una fase di preelaborazione per la rimozione di rumore e la normalizzazione dei dati.

2.Identificazione di cluster: SCEVAN utilizza un approccio basato su clustering per raggruppare le cellule simili in cluster. Questa fase è fondamentale per identificare e isolare le cellule tumorali dal resto delle cellule.

3.Segmentazione variazionale: L'algoritmo applica un approccio variazionale per la segmentazione, cercando di identificare punti di interruzione di copie genomiche all'interno dei cluster di cellule tumorali. Questo passo mira a individuare variazioni nel numero di copie del DNA nelle cellule.

4.Classificazione di sottocloni: SCEVAN classifica i sottocloni tumorali basandosi sui punti di interruzione di copie genomiche identificati nella fase di segmentazione. Questo

consente di distinguere tra diverse popolazioni di cellule tumorali con specifiche alterazioni genomiche.

5.Analisi e interpretazione: Una volta completata la segmentazione e la classificazione, SCEVAN fornisce risultati dettagliati sull'eterogeneità clonale all'interno del tumore. Questi risultati includono la composizione clonale e le specifiche alterazioni genomiche associate a ciascun sottoclone [18].

4. RISULTATI

4.1 Analisi dei dati con Cellenics

La fase iniziale dell'indagine ha coinvolto l'inserimento dei set di dati grezzi provenienti da ciascun paziente nel sistema di analisi Cellenics, il quale ha permesso un'analisi dettagliata seguendo una sequenza di sette passaggi finalizzati a filtrare, integrare e visualizzare in modo efficace le informazioni delle singole cellule.

È importante sottolineare che questi specifici set di dati, relativi a tre pazienti con tumore alla cervice uterina, sono stati scaricati direttamente dall'articolo di riferimento.

-Filtro classificatore

il primo passo ha coinvolto l'applicazione di un filtro classificatore per escludere goccioline vuote e trattenere quelle contenenti cellule, fornendo così una panoramica focalizzata delle entità biologiche rilevanti in ciascun campione.

-Eliminazione delle Goccioline con Contenuto Mitocondriale Elevato

Il secondo passaggio ha visto l'implementazione di un filtro del contenuto mitocondriale, mirato a rimuovere goccioline con cellule morte o di bassa qualità. La soglia di cut-off, calcolata individualmente per ogni campione, ha garantito l'esclusione di goccioline con livelli di contenuto mitocondriale superiori alla norma, preservando così la qualità e la rappresentatività dei dati analizzati.

-Analisi del Numero di Geni vs UMI

Il terzo passo ha esaminato la relazione tra il numero di geni e le UMI, identificando deviazioni significative e categorizzando le goccioline in base a comportamenti anomali.

-Rimozione dei Doppietti

Il quarto passaggio ha applicato un filtro basato sull'algoritmo scDblFinder per eliminare goccioline con alta probabilità di essere duplicati, garantendo l'integrità delle singole cellule nei campioni.

- Integrazione dei Dati

Il quinto passo ha svolto un ruolo cruciale nell'eliminare gli effetti batch e ridurre la complessità dimensionale, facilitando l'analisi delle variazioni biologiche tra i campioni.

-Configurazione dell'Incorporamento

Il sesto passo ha configurato l'incorporamento bidimensionale attraverso t-SNE, semplificando l'identificazione dei cluster.

Il clustering, che rappresenta il processo di aggregazione di cellule con somiglianze significative, è stato eseguito utilizzando il metodo Louvain come impostazione predefinita in Cellenics.

In seguito, mi sono focalizzata sulla sezione Dot Plot di Cellenics per analizzare l'espressione di geni selezionati, in particolare quelli identificati come marcatori tumorali (specifici marcatori utilizzati in immunoistochimica, oppure sono marcatori ricercati e individuati in letteratura scientifica).

4.1.1 *Analisi tramite DOT PLOT*

La sezione Dot Plot di Cellenics è stata esplorata per ogni paziente al fine di valutare l'espressione dei marcatori tumorali personalizzati. Questo grafico permette di visualizzare chiaramente la percentuale di cellule che esprimono specifici geni all'interno di ciascun cluster. La dimensione dei punti riflette la percentuale di espressione genica, mentre il colore indica il livello di espressione del gene.

4.1.2 *Esplorazione dei Marcatori Tumorali*

Attraverso la funzione Dot Plot, sono stati integrati i marcatori tumorali precedentemente identificati in letteratura. Questi marcatori, secondo la letteratura, sono geni che mostrano un'elevata espressione nelle cellule tumorali della cervice uterina. Ciò ha consentito di individuare specifici cluster correlati ai tumori.

Le figure seguenti presentano i Dot Plot per ciascun paziente, evidenziando le differenze nell'espressione genica all'interno dei cluster.

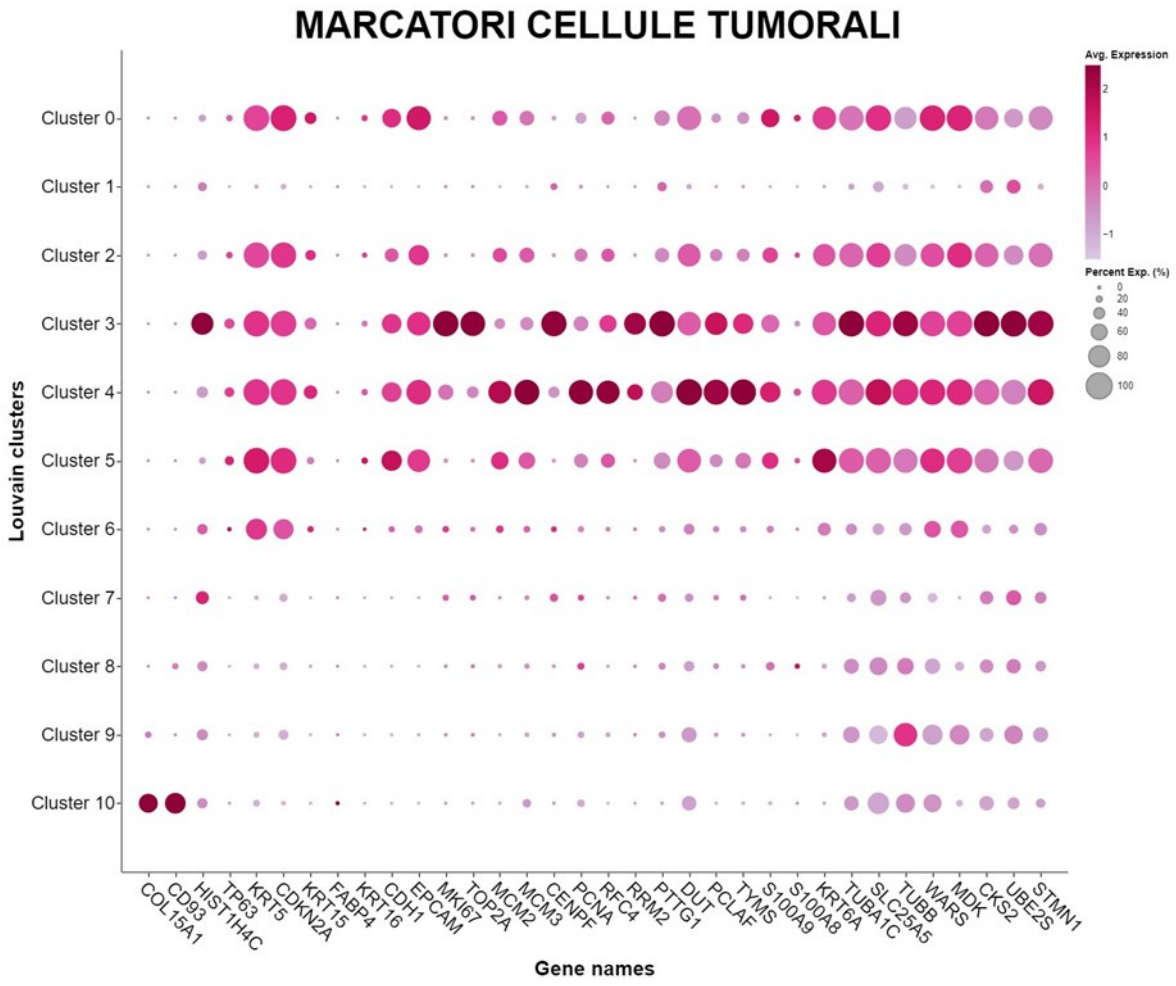

Figura 4. Dot Plot del Paziente 1 - Marcatori Tumorali.

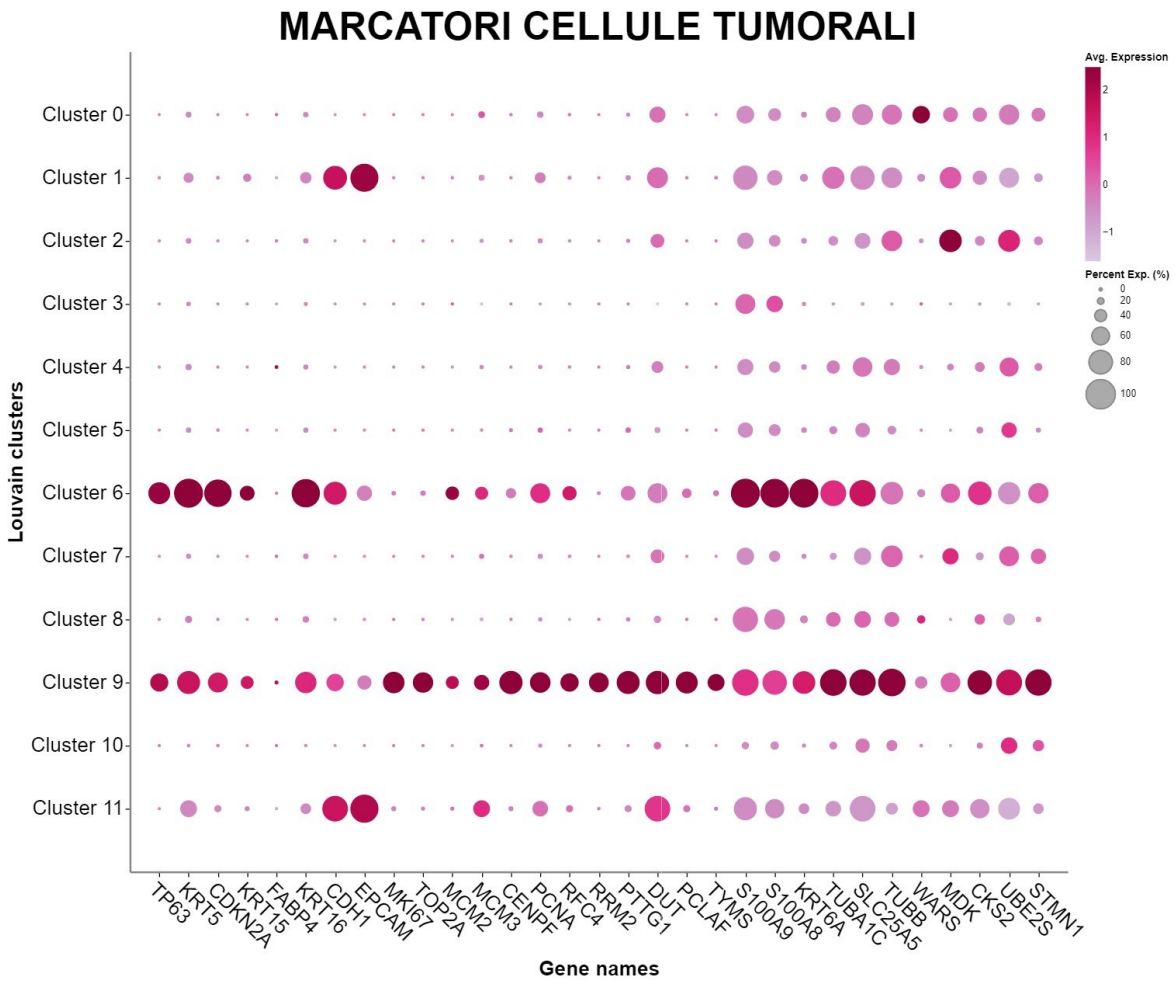

Figura 5: Dot Plot del Paziente 2 - Marcatori Tumorali.

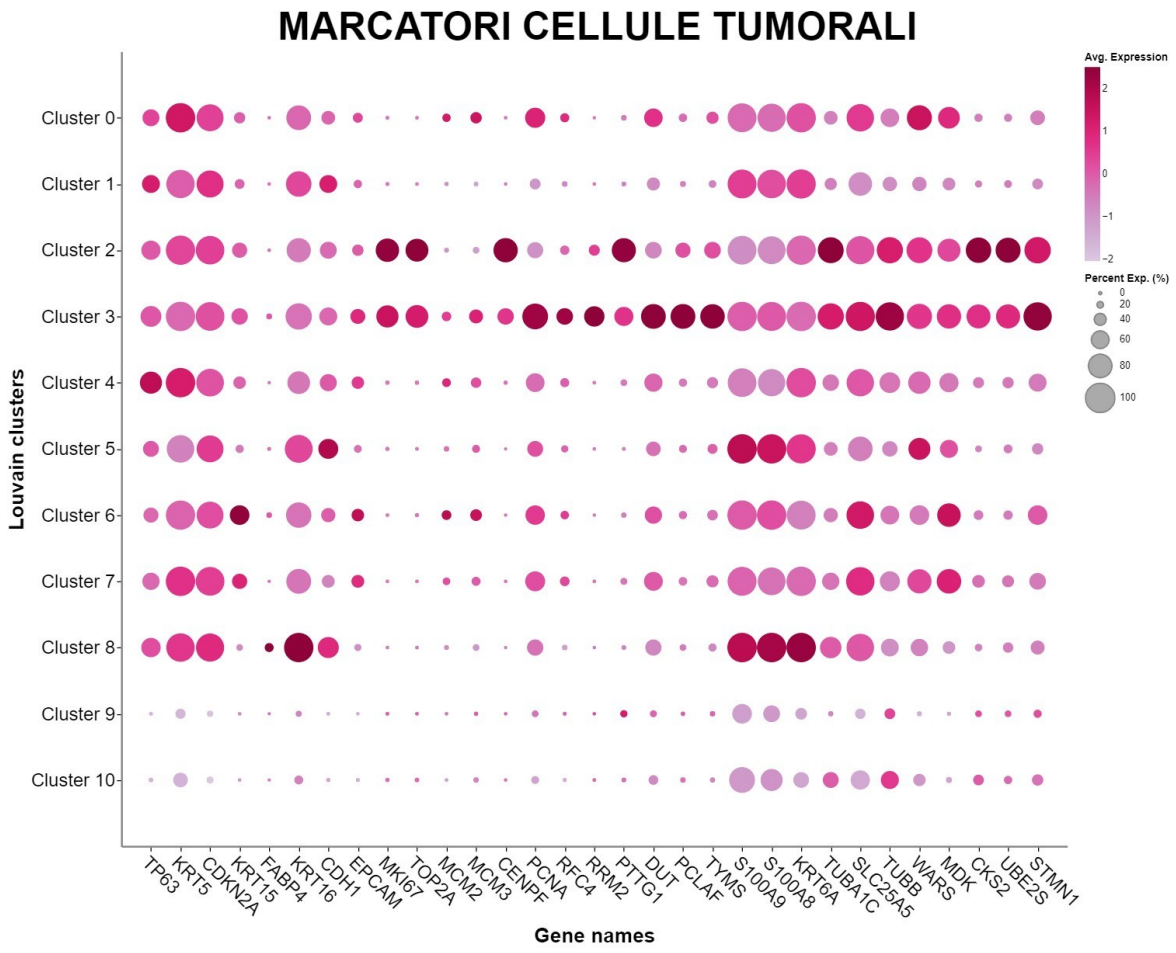

Figura 6. Dot Plot del Paziente 3 - Marcatori Tumorali.

In base all'espressione dei marcatori tumorali, per quanto riguarda il paziente 1 ho considerato tumorali le cellule appartenenti ai cluster 3-4, per il paziente 2 i cluster 6-9 mentre per quanto riguarda il paziente 3 i cluster 2-3.

Parallelamente, ho eseguito lo stesso processo con i dati provenienti dai controlli sani per confermare l'assenza di espressione dei marcatori tumorali nei cluster in questo contesto.

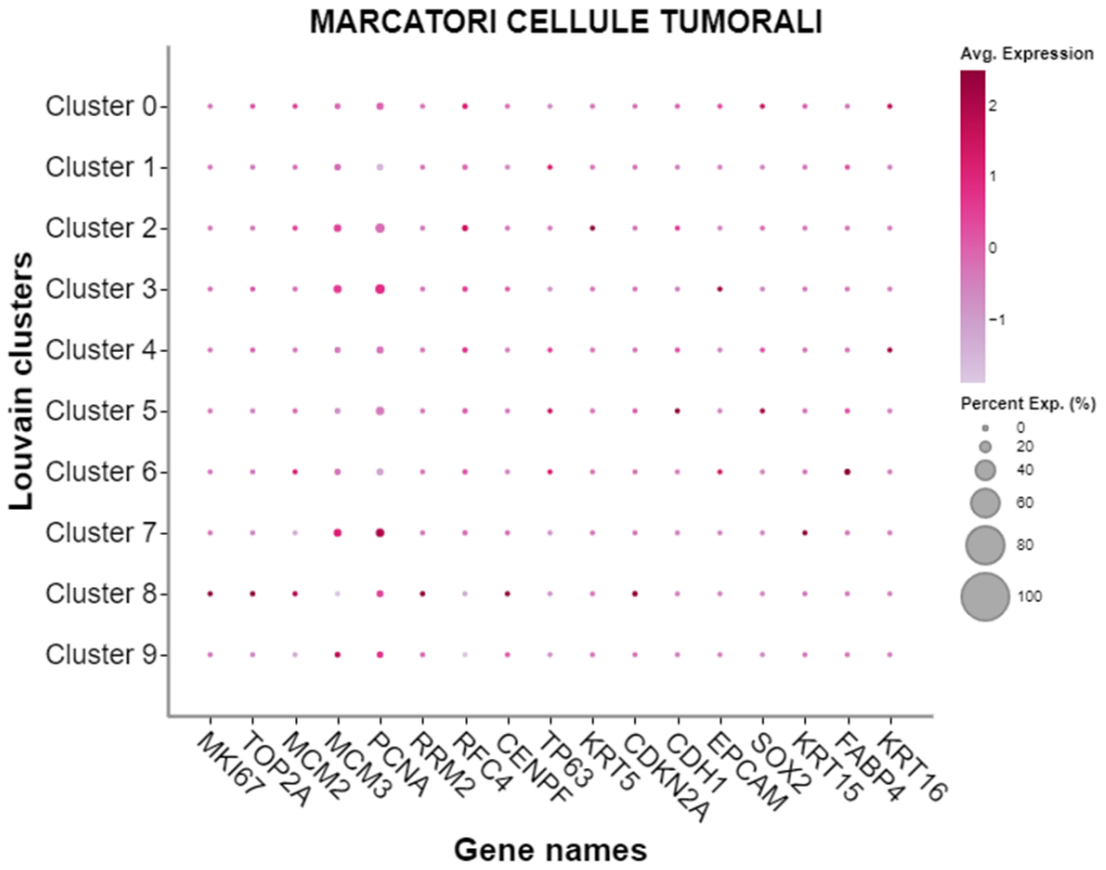

Figura 7. Dot Plot del Paziente 3 (Tessuto Sano) - Marcatori Tumorali.

Dall'esempio, riguardante il controllo sano del paziente 3, riportato nella figura 7 si evince che in nessun cluster sono espressi i principali marcatori tumorali.

4.2 Confronto dei Barcodes delle Cellule Tumorali tra Dati Interni e Articolo di Riferimento:

Una volta individuati i cluster tumorali per ogni paziente, in base all'espressione dei vari marcatori tumorali, sono state scaricate le matrici di espressione di ogni cluster tumorale. L'obiettivo principale di questa fase dell'analisi era confrontare i barcodes delle cellule tumorali identificati nel mio studio con Cellenics con quelli forniti dagli autori nell'articolo di riferimento. I risultati del confronto sono i seguenti:

Per il Paziente 1, c'è una corrispondenza del 73,1% tra i barcodes tumorali.

Per il Paziente 2, la corrispondenza è più alta, pari al 89,1%.

Per il Paziente 3, la corrispondenza è significativamente più bassa, con il 27,3%.

Questo approccio ha permesso di stabilire una corrispondenza tra le cellule tumorali identificate nel mio studio con Cellenics e quelle descritte dagli autori.

Le differenze potrebbero essere attribuite a variazioni nelle piattaforme tecnologiche utilizzate o ad altri fattori che potrebbero influire sulla rappresentazione dei dati.

Tale confronto è fondamentale per comprendere l'aderenza dei miei risultati a quelli esistenti in letteratura e per fornire una base solida alle conclusioni del presente studio.

4.3 Risultati dell'Analisi dei Doppietti tramite ICARUS

Per predire i doppietti è stato utilizzato anche un programma di predizione dei doppietti chiamato ICARUS. Nella figura riportata (Figura 8), sono rappresentati i risultati della predizione dei doppietti effettuata con ICARUS per i pazienti 1, 2 e 3.

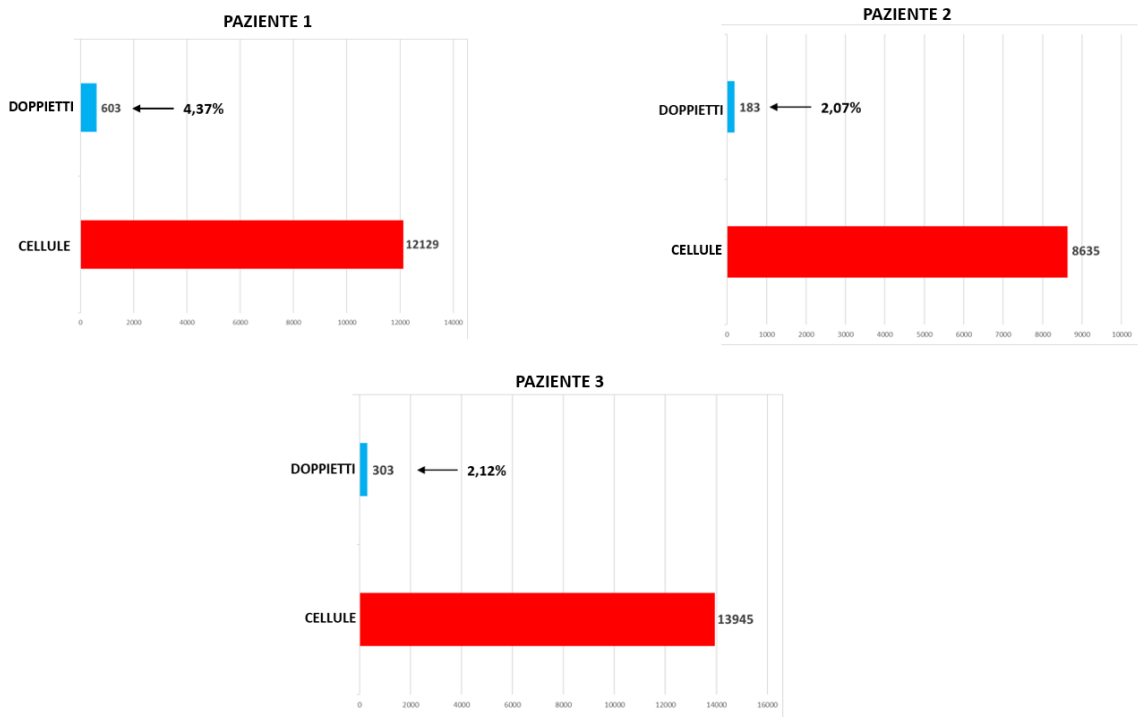

Figura 8. Predizione doppietti con ICARUS.

Si osserva che:

Per il Paziente 1, il 4,37% delle cellule sono state predette come doppietti.

Nel caso del Paziente 2, il 2,07% e per il Paziente 3, la percentuale di cellule predette come doppietti è pari al 2,12%.

4.3.1 Confronto tra Barcodes Tumorali e Doppietti Predetti

In seguito, è stata condotta un'approfondita analisi dei risultati ottenuti, focalizzandosi sul confronto tra i doppietti predetti dal software ICARUS e i barcodes corrispondenti alle cellule tumorali.

È importante sottolineare che per condurre questo confronto, sono stati impiegati sia i barcodes tumorali forniti dagli autori nel contesto dell'articolo di riferimento, sia quelli da me individuati attraverso l'utilizzo di Cellenics.

Questa fase di valutazione è stata fondamentale per valutare l'accuratezza del programma nella distinzione tra doppietti e barcodes effettivi delle cellule tumorali.

È stato eseguito un dettagliato confronto tra i barcodes delle cellule tumorali noti dall'articolo di riferimento e i doppietti predetti dal programma ICARUS.

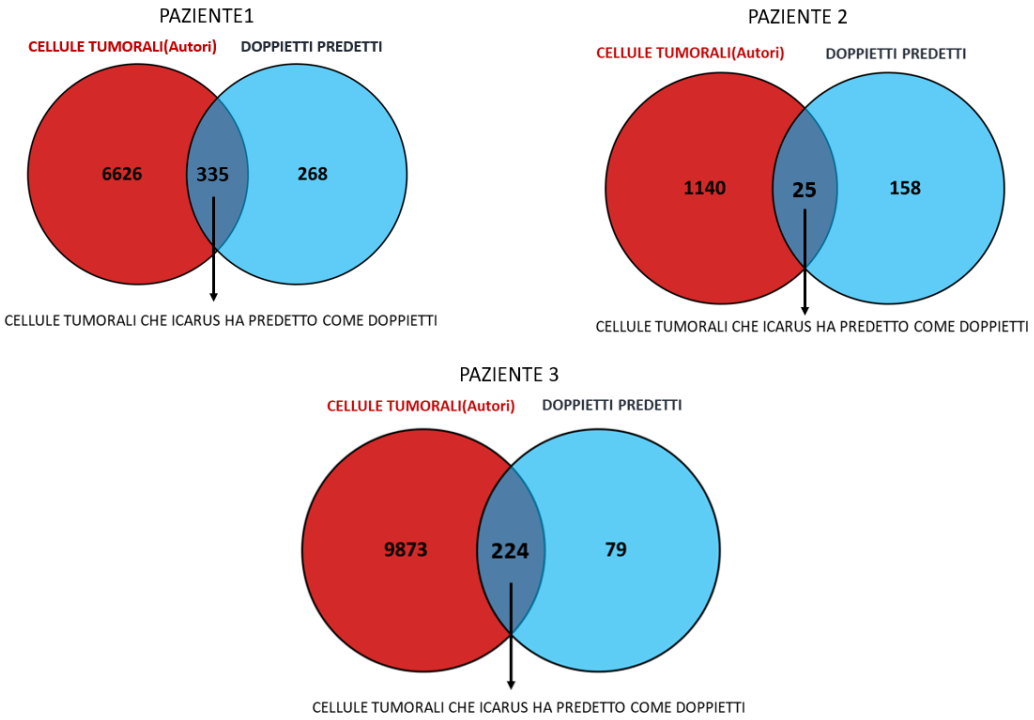

Figura 9. Cellule tumorali note dall'articolo di riferimento e doppietti predetti da ICARUS.

Nella figura 9 si osserva che per il paziente 1, ICARUS ha previsto la presenza di 335 cellule tumorali sotto forma di doppietti (4.8%). Nel caso del paziente 2, la previsione è stata di 25 cellule tumorali come doppietti (2.14%). Per quanto riguarda il paziente 3, ICARUS ha predetto 224 cellule tumorali come doppietti (2.21%).

Allo stesso modo, è stata ripetuta la procedura utilizzando i barcodes delle cellule tumorali da me individuate con Cellenics.

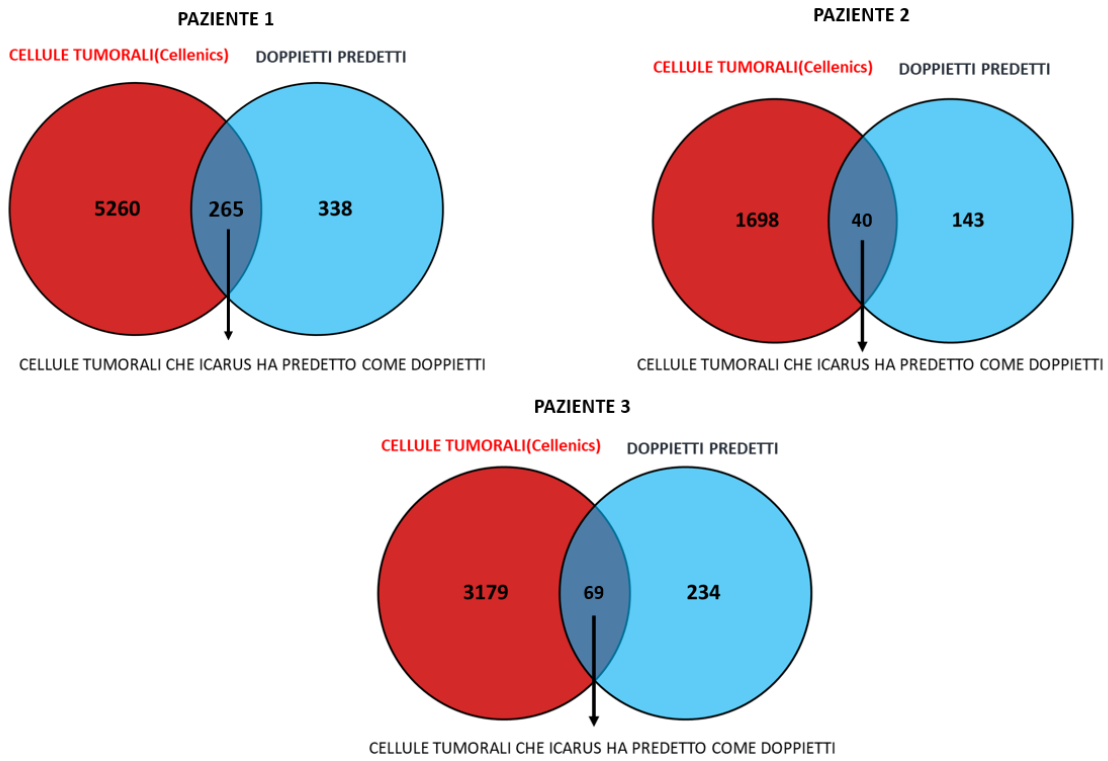

Figura 10. Cellule tumorali individuate tramite l'analisi con Cellenics e doppietti predetti da ICARUS.

Nella Figura 10 si evidenzia che, per il primo paziente, ICARUS ha predetto 265 cellule tumorali come doppietti (4.7%). Per il secondo paziente, ha predetto 40 cellule tumorali come doppietti (2.3%). Per il terzo paziente, ha predetto 69 cellule tumorali come doppietti (2.12%). In tutti i casi, si osserva che la percentuale di cellule tumorali predette da ICARUS come doppietti è relativamente bassa.

L'obiettivo principale era valutare se il software avesse eliminato cellule tumorali considerandole doppietti. Dai risultati presentati, sembra che ICARUS abbia mantenuto una buona precisione nel distinguere tra cellule tumorali e doppietti.

4.4 Annotazione cellulare

4.4.1 Risultati dell'Annotazione Cellulare con Diversi Metodi

Dopo aver utilizzato diversi strumenti di annotazione cellulare - SingleR, SciBet e scGEATool (PanglaoDB) - per identificare i tipi cellulari nei dati di scRNA-seq, sono stati ottenuti risultati dettagliati. Ciascun metodo si basa su approcci specifici per interpretare le informazioni trascrittomiche e assegnare etichette ai singoli barcodes delle cellule. Di seguito, vengono presentati i risultati di questa fase dell'analisi.

Sono riportate, per ciascun paziente, le tabelle contenenti le annotazioni cellulari assegnate dai diversi strumenti.

ANNOTAZIONE CELLULARE-PAZIENTE 1

TIPO CELLULARE	SINGLE R	SCIBET	MATLAB
'Epithelial_cells'	10261	9503	7154
'T_cells'	1287	1398	1383
'NK_cell'	534	374	393
'Tissue_stem_cells'	62	0	0
'Endothelial_cells'	50	254	40
'Fibroblasts'	75	14	123
'Monocyte'	35	4	0
'HSC_CD34+'	13	10	0
'Macrophage'	125	258	0
'B_cell'	87	15	72
'DC'	81	18	465
'Pre-B_cell_CD34-'	23	0	0
'CMP'	20	0	0
'Neutrophils'	17	0	0
'MSC'	7	0	0
'Smooth_muscle_cells'	7	225	0
'Neurons'	14	0	56
'Astrocyte'	2	0	0
'Keratinocytes'	15	0	0
'Hepatocytes'	1	0	0
'Pro-B_cell_CD34+'	4	0	0
'GMP'	5	0	0
'Chondrocytes'	3	0	0
'iPS_cells'	1	0	0
'Osteoblasts'	1	0	0
'BM & Prog.'	1	0	0
'Gametocytes'	1	4	0
Plasma cells	0	109	60
'Mast'	0	80	0
'Skeletal Muscle Myoblasts'	0	16	0
'ILC'	0	21	0
'Pancreatic islets'	0	1	0
'H9 cells'	0	1	0
'HEK and 3T3 mix'	0	4	0
'Microglia'	0	2	0
'FGC'	0	1	0
assente'	0	0	3042

Tabella 2. Annotazione cellulare PAZIENTE1.

ANNOTAZIONE CELLULARE-PAZIENTE 2

TIPO CELLULARE	SINGLE R	SCIBET	MATLAB
'Endothelial_cells'	1594	2892	1555
'Smooth_muscle_cells'	638	0	318
'Epithelial_cells'	2223	2041	1929
'Tissue_stem_cells'	1081	0	0
'Fibroblasts'	971	126	2113
'T_cells'	647	432	857
'Neutrophils'	911	0	0
'Monocyte'	172	134	655
'HSC_CD34+'	20	25	0
'Neurons'	27	1	0
'GMP'	4	0	0
'MSC'	24	0	0
'NK_cell'	117	19	0
'Chondrocytes'	103	0	0
'Macrophage'	97	1181	0
'B_cell'	59	32	0
'CMP'	32	0	0
'Osteoblasts'	16	0	0
'DC'	44	23	415
'Keratinocytes'	20	0	0
'Pro-B_cell_CD34+'	4	0	0
'Pre-B_cell_CD34-'	14	0	0
Mast cell'	0	159	75
'Pancreatic islets'	0	4	0
'preimplantation blastomer	0	2	0
'Microglia'	0	2	0
'HEK and 3T3 mix'	0	2	0
'Muscle'	1372	0	0
'ILC'	206	0	0
'Skeletal Muscle Myoblasts'	77	0	0
'Plasma B'	0	70	116
'Pancreatic stellate cells'	0	0	78
assente'	0	0	706

Tabella 3. Annotazione cellulare PAZIENTE 2.

ANNOTAZIONE CELLULARE-PAZIENTE 3

TIPO CELLULARE	SINGLE R	SCIBET	MATLAB
'Epithelial_cells'	13573	8855	2729
'T_cells'	255	267	345
'Keratinocytes'	193	0	4837
'Macrophage'	72	474	0
'NK_cell'	68	267	0
'Monocyte'	26	150	0
'DC'	33	11	134
'CMP'	5	0	0
'GMP'	2	0	0
'B_cell'	10	0	0
'Pre-B_cell_CD34-'	1	0	0
'Neutrophils'	3	0	0
'Endothelial_cells'	1	18	0
'BM'	1	0	0
'Fibroblasts'	2	0	0
'Pro-Myelocyte'	1	0	0
'Tissue_stem_cells'	1	0	0
basal cells'	0	0	6129
'ILC'	0	14	0
'HSCs'	0	1	0
'Microglia'	0	3	0
'preimplantation blastomere'	0	1	0
'H9 cells'	0	1	0
Plasma B'	0	16	0
'Soma'	0	4	0
'Mast'	0	21	0
assente'	0	0	2729

Tabella 4. Annotazione cellulare PAZIENTE 3.

4.4.2 Confronto tra Annotazioni Cellulari

In seguito, è stato effettuato un confronto tra le annotazioni cellulari fornite dai programmi per identificare eventuali divergenze.

Nelle seguenti figure sono riportati i risultati dei confronti tra le annotazioni cellulari effettuate dai vari programmi: Single R, SciBet e Matlab.

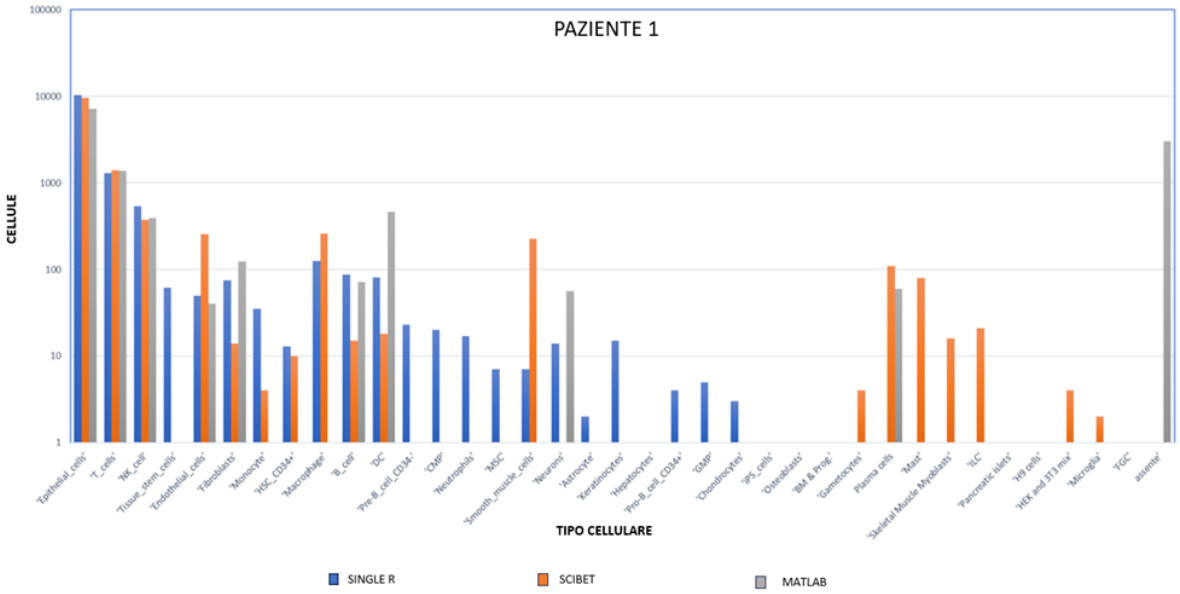

Figura 11. Confronto annotazione cellulare-PAZIENTE 1.

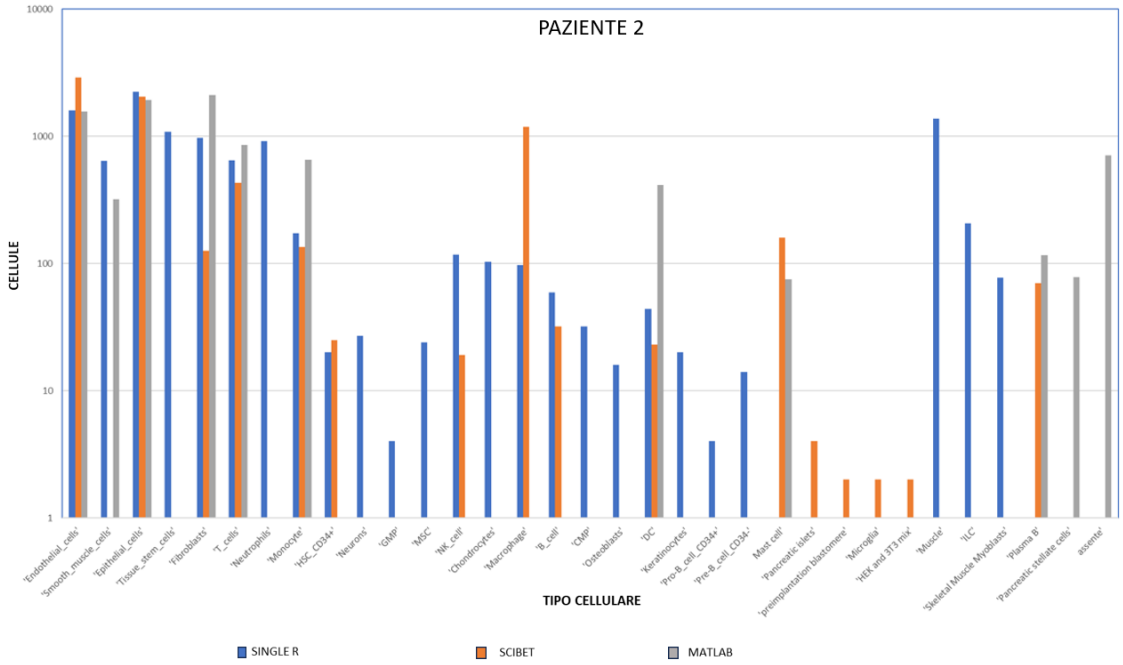

Figura 12. Confronto annotazione cellulare-PAZIENTE 2.

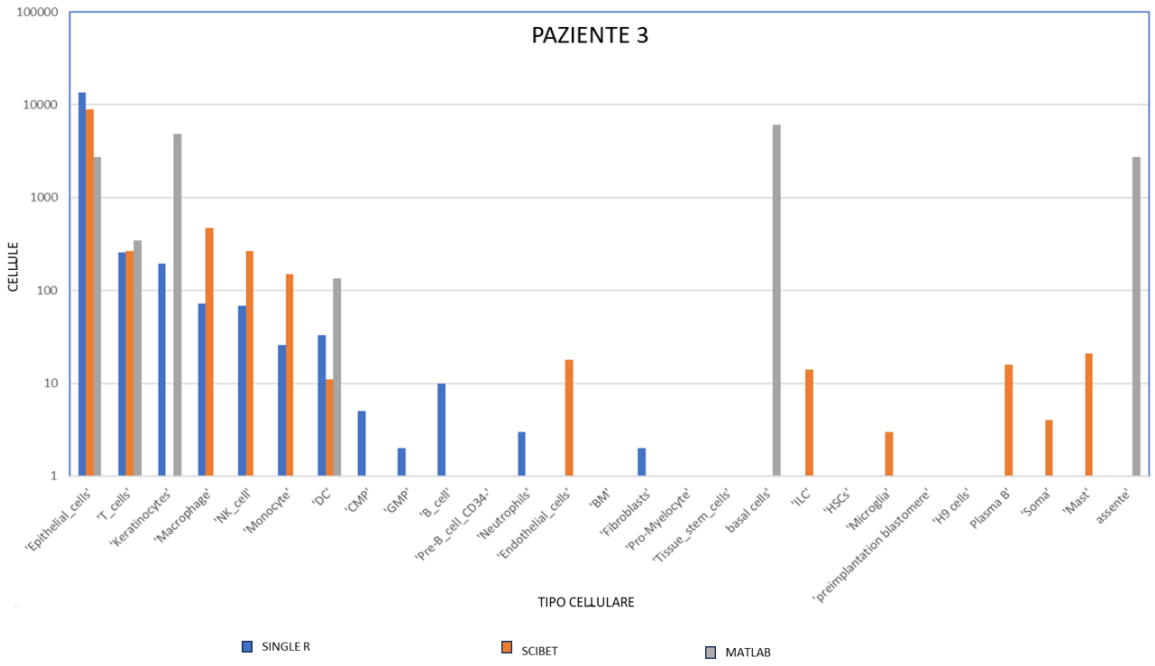

Figura 13. Confronto annotazione cellulare-PAZIENTE 3.

Il Tool di Matlab non è in grando di riconoscere una grande varietà di cellule, mentre SciBet e Single R riescono a riconoscere più tipi cellulari. Tuttavia, va sottolineato che non siamo certi della correttezza del riconoscimento cellulare ottenuto da nessuno dei software utilizzati.

Successivamente, per condurre un approfondito confronto tra le annotazioni cellulari effettuate dai tre software, sono stati analizzati i risultati relativi al Paziente 1. Ad esempio, per quanto riguarda l'identificazione delle cellule epiteliali, Single R ne ha individuate 10.261, SciBet 9.503 e Matlab 7.154. Tra questi, 3.087 cellule epiteliali sono state identificate in comune dai tre software. Inoltre, SciBet e Single R hanno individuato 9,401 cellule epiteliali in comune, superando il numero di cellule comuni individuate da SciBet e Matlab (3,115), e da Matlab e Single R (3,382).

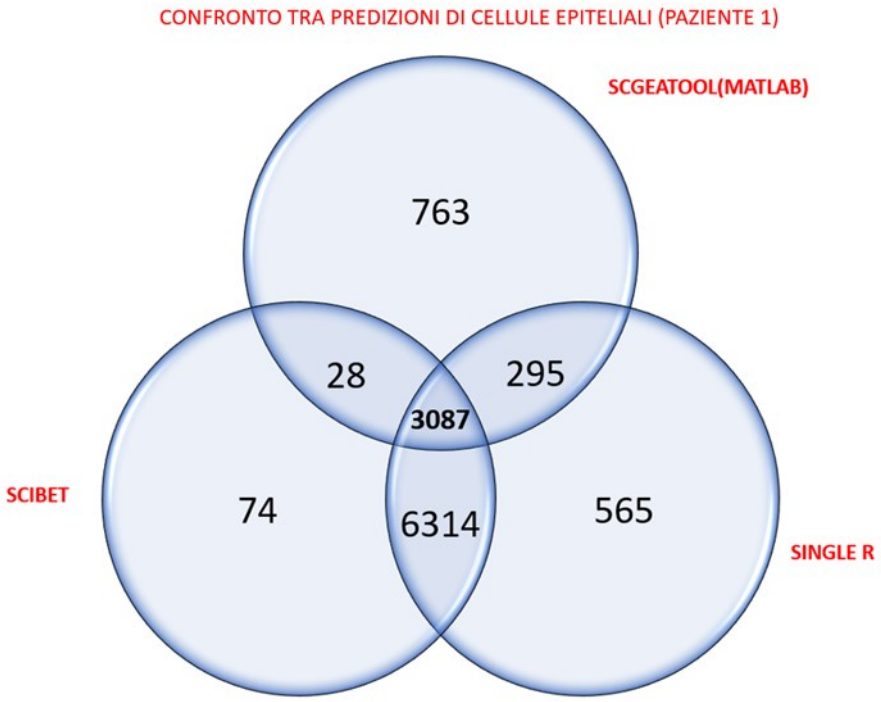

Figura 14. Confronto tra predizioni di cellule epiteliali con i programmi: Single R, SciBet, scGEATOOL (Matlab)-Paziente 1.

4.5 Fasi del ciclo cellulare

L'analisi del ciclo cellulare condotta utilizzando scGEATool (MATLAB) ha rivelato una distribuzione eterogenea delle fasi del ciclo cellulare nelle cellule esaminate per quanto riguarda il Paziente 1.

Per quanto riguarda i Pazienti 2 e 3, un numero più elevato di cellule si trova nella fase G1 del ciclo cellulare. Ciò è raffigurato nella figura seguente:

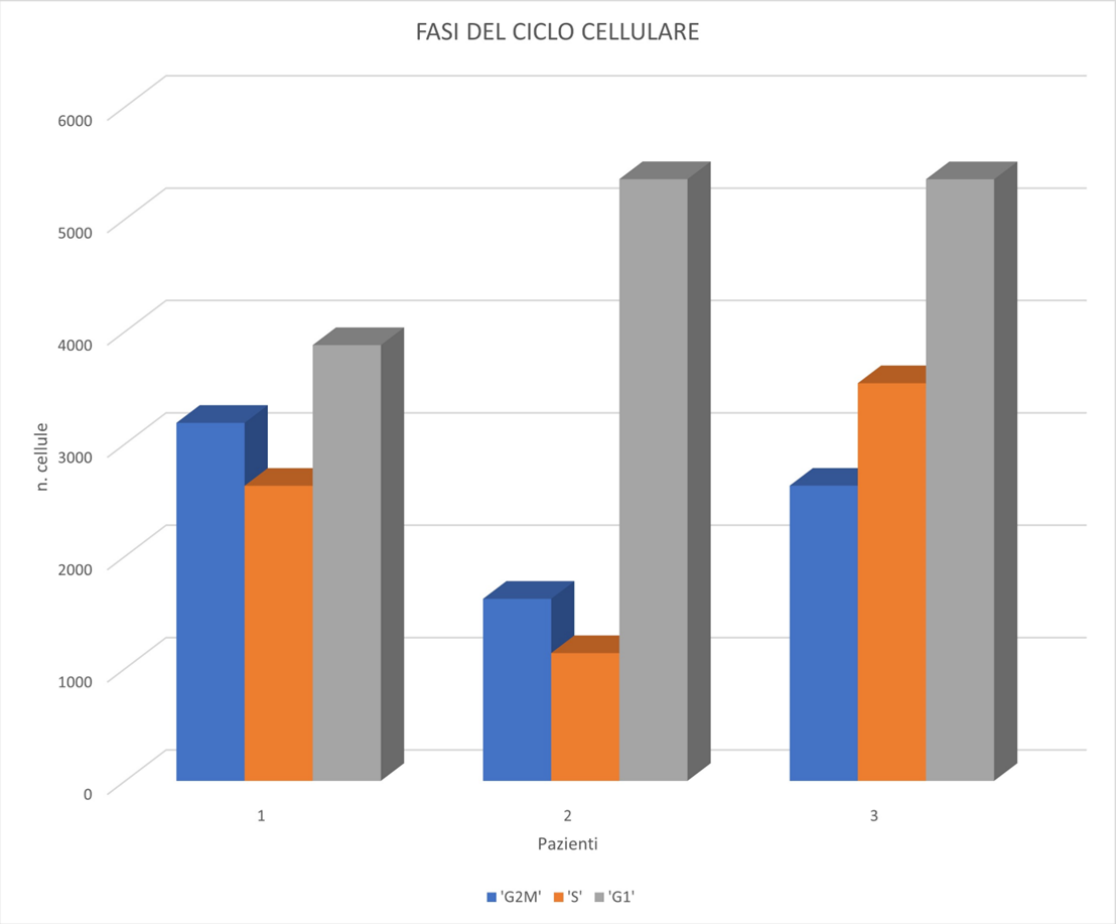

Figura 15. Distribuzione delle fasi del ciclo cellulare nelle cellule esaminate.

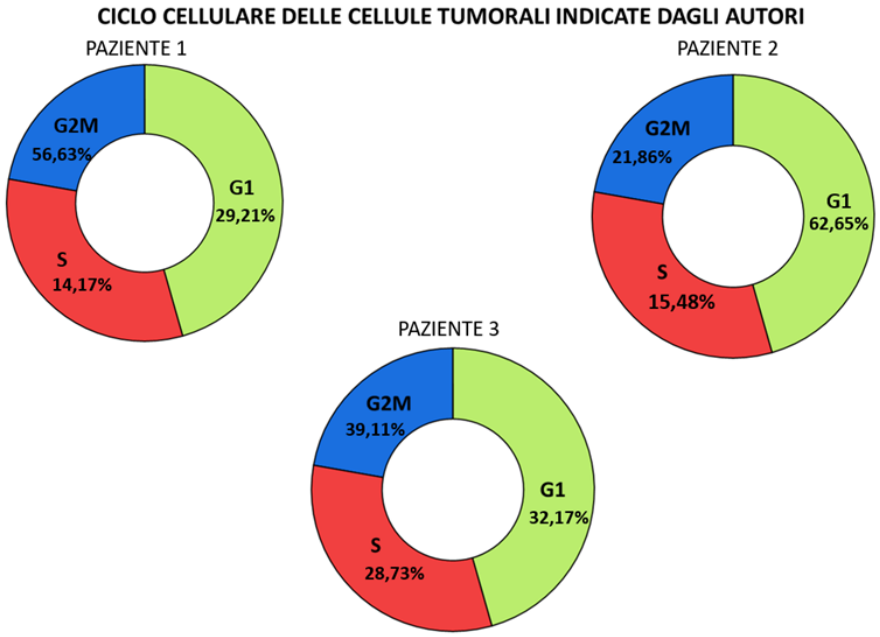

Figura 16. Ciclo cellulare delle cellule tumorali (indicate dagli autori).

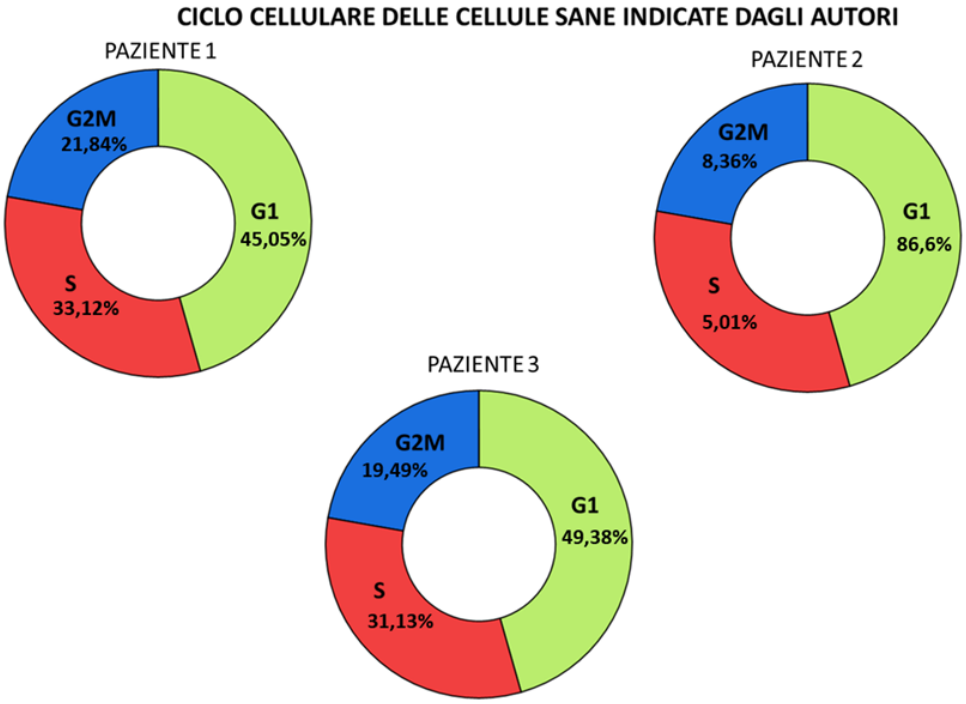

Figura 17. Ciclo cellulare delle cellule sane (indicate dagli autori).

Nelle figure 16 e 17 sono rappresentate le fasi del ciclo cellulare in cui si trovano le cellule tumorali e sane (individuate dagli autori) dei pazienti 1, 2 e 3.

4.6 Inferenza CNV

Tramite inferenza dei CNV, effettuata con SCEVAN, sono state predette le cellule tumorali per ciascun paziente. Nelle figure 18, 19, 20 sono rappresentate le Heatmap di ciascun paziente e vengono raffigurate amplificazioni (in blu), delezioni (in rosso) e stati di neutralità nei loci di tutti i cromosomi di ciascuna cellula. Nelle Heatmap ogni riga rappresenta una cellula mentre i numeri indicano i cromosomi.

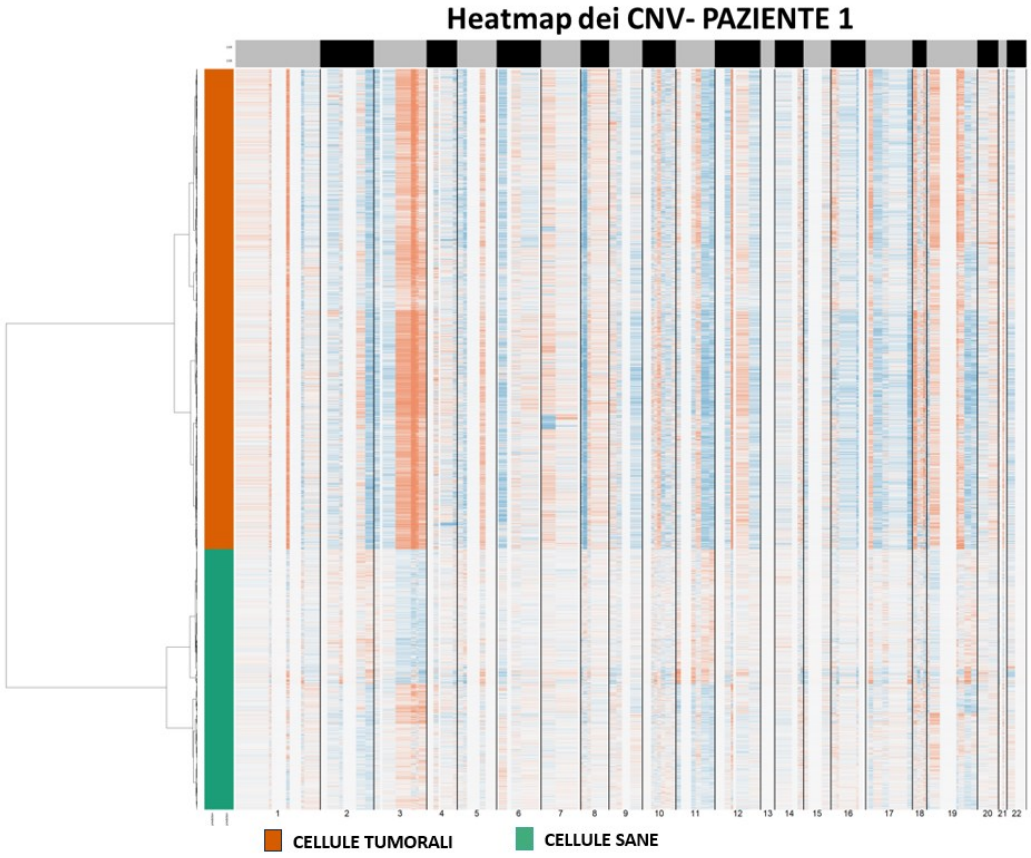

Figura 18. Heatmap dei CNV inferiti- Paziente 1.

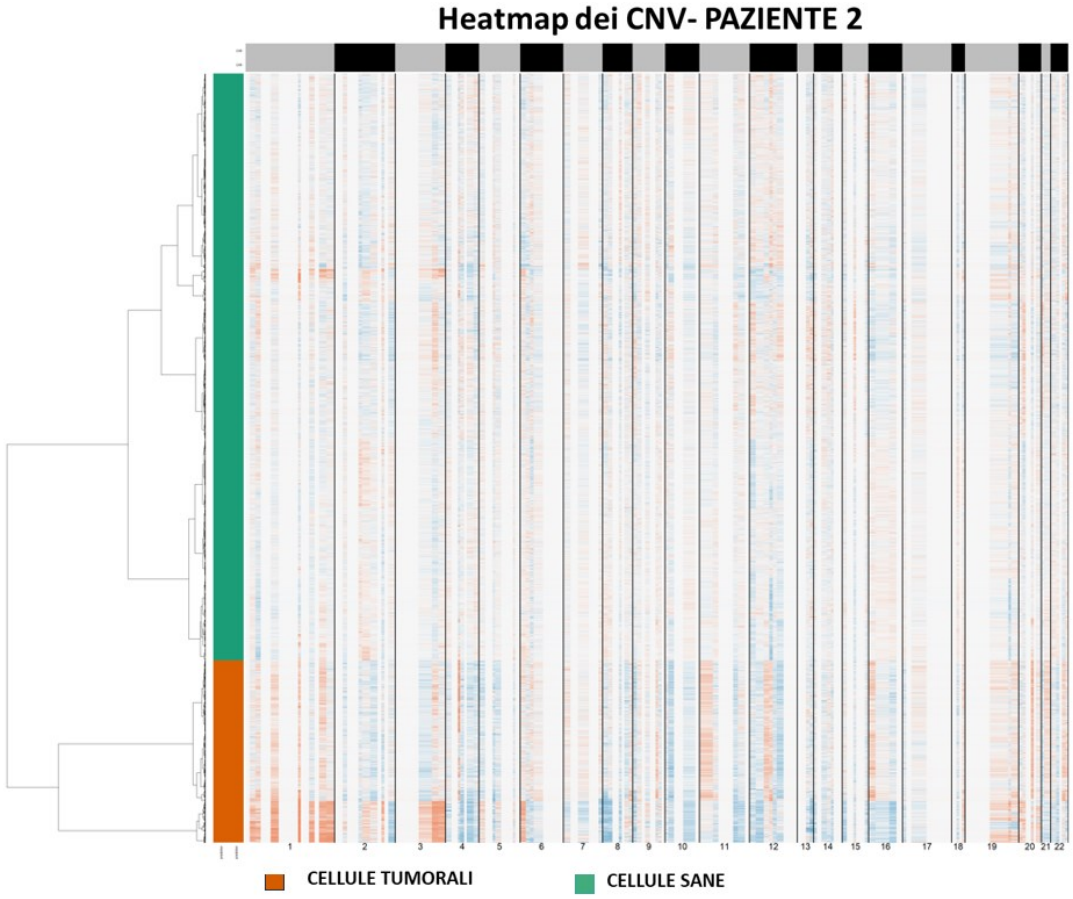

Figura 19. Heatmap dei CNV inferiti- Paziente 2.

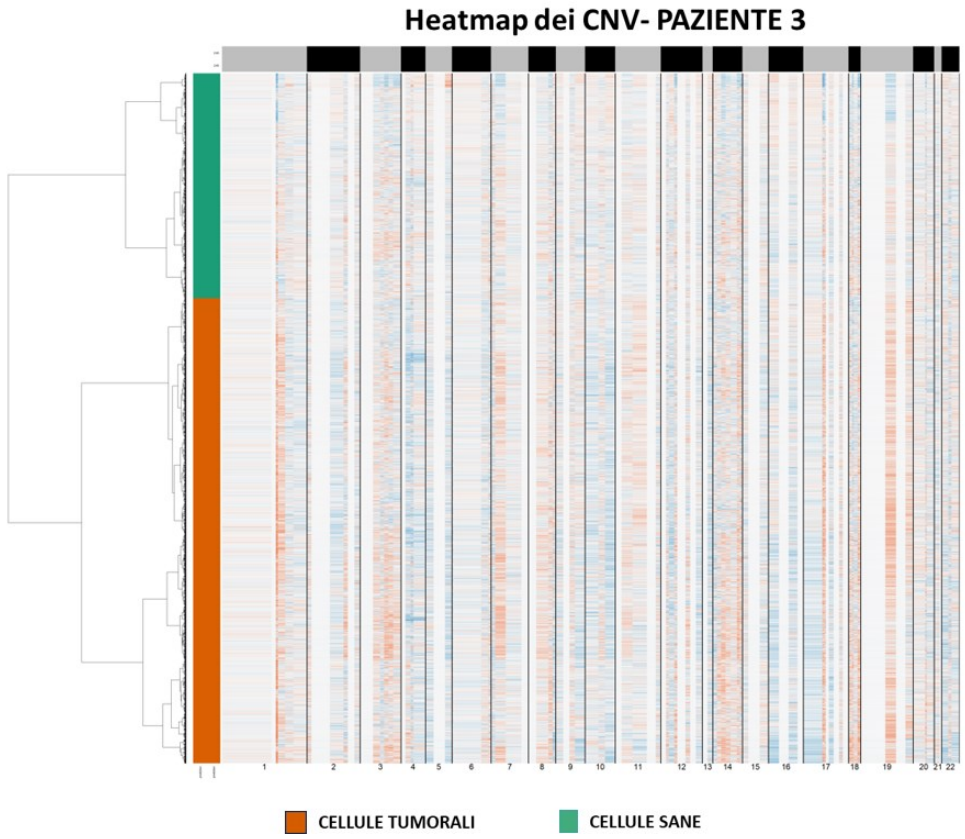

Figura 20. Heatmap dei CNV inferiti- Paziente 3.

In particolare, nel paziente 1 e 2 è visibile una chiara distinzione tra cellule sane e tumorali, infatti, le cellule tumorali predette presentano un maggior numero di CNV rispetto alle cellule sane. Invece, per quanto riguarda il paziente 3, la distinzione non è molto chiara e la distribuzione dei CNV è simile tra cellule maligne e sane.

Questo è coerente con i risultati ottenuti confrontando le cellule tumorali di Cellenics con quelle predette da SCEVAN.

Infatti, per quanto riguarda il Paziente 1 c'è una corrispondenza del 97,2% tra i barcodes tumorali, per il Paziente 2 dell'83% mentre per il Paziente 3 la corrispondenza è più bassa, ovvero del 61,7%.

5. DISCUSSIONE

L'eterogeneità del cancro si riferisce alla diversità genetica e fenotipica che può essere presente tra pazienti con la stessa tipologia di tumore o anche all'interno dello stesso tumore. Questa diversità può influenzare la risposta al trattamento e la prognosi del paziente.

Comprendere e caratterizzare questa diversità è importante per lo sviluppo di terapie più mirate e personalizzate, migliorando l'efficacia dei trattamenti e riducendo gli effetti collaterali.

La scRNA-seq è una tecnica che consente di analizzare il profilo genetico di singole cellule, consentendo una visione dettagliata della diversità cellulare all'interno di un tessuto tumorale. Offre la possibilità di esplorare dettagliatamente le proprietà biologiche di ogni singola cellula. Questa tecnica avanzata consente di analizzare il profilo genetico di singole cellule, consentendo una visione dettagliata della diversità cellulare all'interno di un tessuto tumorale. È particolarmente utile per comprendere le differenze genetiche tra tumori primari e metastasi, nonché per identificare sottopopolazioni cellulari che possono essere responsabili della resistenza al trattamento.

In questo studio sono stati valutati dei software utilizzati per l'analisi dei dati derivanti dal sequenziamento dell'RNA a singola cellula per comprendere la correttezza e l'affidabilità delle informazioni ottenute da questa tecnologia avanzata. Studiare questi software è molto importante in quanto diversi strumenti possono produrre risultati leggermente diversi a causa delle loro implementazioni algoritmiche e delle modalità di elaborazione dei dati. L'analisi dettagliata delle differenze e delle convergenze tra questi software permettere di comprendere le limitazioni e le potenzialità dei programmi impiegati nello studio

dell'eterogeneità tumorale mediante scRNA-seq. Questo tipo di analisi può fornire indicazioni preziose per migliorare la precisione e l'interpretazione dei risultati.

Nello studio sono stati analizzati dati derivanti dal sequenziamento dall'RNA a singola cellula da biopsie di tumori della cervice uterina. Il tumore della cervice uterina è molto diffuso ma poco studiato.

I dati sono stati inizialmente analizzati utilizzando il software Cellenics, il quale ha permesso di identificare le cellule tumorali in base all' espressione dei geni marcatori del tumore della cervice, individuati in letteratura.

In seguito, è stato effettuato il confronto dei barcodes delle cellule tumorali identificate tramite espressione genica con quelli riportati nell'articolo di riferimento.

Le percentuali di corrispondenza dei barcodes tumorali ottenute per ciascun paziente offrono informazioni preziose sulla corrispondenza tra i risultati dello studio e quelli presentati nell'articolo di riferimento.

Per il Paziente 1, c'è una corrispondenza del 73,1%, ciò indica una sovrapposizione significativa nei barcodes tumorali rispetto all'articolo di riferimento.

Per il Paziente 2, la corrispondenza è più alta, pari al 89,1%.

Un'alta corrispondenza suggerisce una consistenza notevole tra i barcodes delle cellule tumorali individuate con Cellenics e quelli dell'articolo di riferimento.

Questo può indicare una maggiore affidabilità e precisione nelle analisi effettuate su questo campione specifico.

Per il Paziente 3, la corrispondenza è significativamente più bassa, con il 27,3%. Le ragioni di questa discrepanza, ad esempio, potrebbero essere dovute a variazioni tecniche che hanno influenzato i risultati.

L'analisi dei dati di scRNA-seq risulta complessa a causa della presenza di artefatti tecnici noti come doppietti, che consistono in informazioni sul trascrittoma di singola cellula che riflettono più di una cellula. Per questo motivo è stata condotta un'analisi utilizzando un software per predire questi artefatti.

Una volta ottenuti i risultati è stato effettuato un confronto tra i doppietti predetti dal software ICARUS e i barcodes corrispondenti alle cellule tumorali, sia quelle fornite dagli autori nell'articolo di riferimento sia quelle individuate tramite Cellenics.

Dai risultati ottenuti, sembra che ICARUS abbia mantenuto una buona precisione nel distinguere tra cellule tumorali e doppietti.

Per identificare i tipi cellulari nei dati di scRNA-seq sono stati utilizzati diversi strumenti di annotazione cellulare e sono state poi confrontate tra loro le annotazioni cellulari fornite dai programmi per identificare eventuali divergenze.

Sono state osservate molte divergenze tra i vari programmi ed inoltre nessun programma è stato in grado di predire le cellule tumorali. I motivi delle divergenze tra i programmi sono molteplici. È importante capire quali sono i criteri per classificare i tipi di cellule utilizzati dalle diverse piattaforme che identificano i tipi cellulari basandosi su dati scRNA-seq.

Alcuni identificano i tipi cellulari basandosi sulla presenza di biomarcatori, ma abbiamo visto che piattaforme diverse utilizzano biomarcatori diversi per identificare lo stesso tipo di cellula.

Non esiste un marker univoco per le cellule tumorali in quanto esistono sottotipi di cellule tumorali, biomarcatori diversi possono essere utilizzati per identificare lo stesso tipo di cellula e ciò sottolinea la complessità nell'interpretare i risultati di scRNA-seq.

Tutte queste incertezze nell'annotazione cellulare possono influenzare la comprensione della biologia cellulare e le implicazioni cliniche dei risultati.

Per quanto riguarda l'analisi dei CNV attraverso SCEVAN i risultati sono interessanti, evidenziano differenze significative tra le cellule sane e tumorali nei pazienti 1 e 2, ma mostrano una distinzione meno chiara nel paziente 3.

È stato eseguito un confronto tra le cellule tumorali di Cellenics (individuate utilizzando marcatori tumorali) con quelle predette da SCEVAN (individuate in base al numero di CNV). Nei pazienti 1 e 2 c'è una corrispondenza elevata tra i barcodes tumorali di Cellenics e quelli predetti da SCEVAN.

Nel paziente 3, la corrispondenza più bassa può essere interpretata come una discrepanza nei risultati tra Cellenics e SCEVAN per quanto riguarda l'identificazione delle cellule tumorali, in questo caso è bene considerare che ci sono specifiche caratteristiche biologiche o tecniche che influenzano i risultati.

Nel paziente 3 si era osservata anche una scarsa corrispondenza tra i barcodes tumorali individuati con Cellenics e quelli forniti dagli autori, ciò aggiunge un ulteriore livello di complessità all'interpretazione dei risultati.

La discrepanza potrebbe essere dovuta a differenze nelle tecniche di analisi e alla specificità dei marcatori utilizzati.

6. CONCLUSIONI

Questo lavoro di tesi ha permesso di esplorare una tecnica promettente ma complessa. L'analisi critica dei software utilizzati nell'ambito della scRNA-seq rivela che, nonostante i progressi, esistono ancora limitazioni nella correttezza e nell'affidabilità delle informazioni ottenute. I programmi per filtrare i dati sono vari ma si registra solo una parziale coerenza tra di essi. È importante sottolineare che le versioni aggiornate di tali software vengono rilasciate con alta frequenza. Durante l'analisi, si è notato che l'elaborazione dei dati con programmi diversi produce risultati poco sovrapponibili, e spesso mancano dati di riferimento per attribuire un valore alle predizioni. Questo rende difficile indicare quale programma elabori i dati con maggiore precisione, impedendo la formulazione di consigli utili. È sorprendente la diversità di programmi disponibili per identificare i tipi cellulari, ma è altrettanto notevole il limitato accordo tra di essi. Si è osservato che in alcuni casi i risultati di un programma di riconoscimento cellulare dipendono notevolmente dai dati con cui è stato addestrato. Si auspica che in futuro possano essere sviluppati dataset completi che includano trascrittomica e proteomica di tutti i tipi cellulari, contribuendo così alla creazione di programmi più robusti. Anche il riconoscimento delle cellule tumorali è tutt'altro che univoco e può avvenire attraverso firme trascrittomiche o la presenza di CNV. Diversi autori suggeriscono identikit trascrittomici diversi per riconoscere le cellule tumorali, portando a risultati solo parzialmente sovrapponibili. Inoltre, la stringenza di queste firme trascrittomiche non è ben definita. Ad esempio, se l'identikit di una cellula tumorale è costituito dall'espressione di 6 geni e una cellula ne esprime 4, come dovremmo considerare questa cellula? È anche degno di nota che, nonostante la maggior parte degli autori di articoli sulle single cell affermi che le cellule tumorali possono essere identificate

in modo univoco attraverso CNV, i CNV possono solo essere predetti dai dati single cell, e quindi stupisce constatare affermazioni di discriminazione certa basata su predizioni. In questo lavoro di tesi, è stato utilizzato uno di questi programmi di predizione di CNV, ma sorprende il fatto che i risultati sono fortemente influenzati dai parametri del software, senza che siano forniti parametri consigliati a priori. Tutto questo fa attribuire un peso limitato agli articoli che descrivono esperimenti di trascrittomica su singola cellula. Infatti, a partire dagli stessi dati autori diversi ottengono risultati e quindi conclusioni biologiche diverse. Naturalmente si prevede che tutte queste difficoltà siano temporanee in quanto la tecnica migliorerà velocemente e in conseguenza di questo gli algoritmi di elaborazione potranno migliorare rendendo i risultati meno dipendenti dagli operatori. Occorrerebbero esperimenti creati apposta per verificare l'affidabilità della Single Cell, in cui si mescolano rapporti determinati di tipi cellulari diversi per verificare i risultati di un software. Questo studio sottolinea la necessità di miglioramenti nei software e nella standardizzazione per ottenere risultati più affidabili e clinicamente rilevanti.

7. RIFERIMENTI

1. Bhatla, N.; Aoki, D.; Sharma, D. N.; Sankaranarayanan, R., Cancer of the cervix uteri: 2021 update. *Int J Gynaecol Obstet* **2021,** 155 Suppl 1, (Suppl 1), 28-44.
2. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R. L.; Torre, L. A.; Jemal, A., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **2018,** 68, (6), 394-424.
3. Balasubramaniam, S. D.; Balakrishnan, V.; Oon, C. E.; Kaur, G., Key Molecular Events in Cervical Cancer Development. *Medicina (Kaunas)* **2019,** 55, (7).
4. Wipperman, J.; Neil, T.; Williams, T., Cervical Cancer: Evaluation and Management. *Am Fam Physician* **2018,** 97, (7), 449-454.
5. Jovic, D.; Liang, X.; Zeng, H.; Lin, L.; Xu, F.; Luo, Y., Single-cell RNA sequencing technologies and applications: A brief overview. *Clin Transl Med* **2022,** 12, (3), e694.
6. Olsen, T. K.; Baryawno, N., Introduction to Single-Cell RNA Sequencing. *Curr Protoc Mol Biol* **2018,** 122, (1), e57.
7. Su, M.; Pan, T.; Chen, Q. Z.; Zhou, W. W.; Gong, Y.; Xu, G.; Yan, H. Y.; Li, S.; Shi, Q. Z.; Zhang, Y.; He, X.; Jiang, C. J.; Fan, S. C.; Li, X.; Cairns, M. J.; Wang, X.; Li, Y. S., Data analysis guidelines for single-cell RNA-seq in biomedical studies and clinical applications. *Mil Med Res* **2022,** 9, (1), 68.
8. You, Y.; Tian, L.; Su, S.; Dong, X.; Jabbari, J. S.; Hickey, P. F.; Ritchie, M. E., Benchmarking UMI-based single-cell RNA-seq preprocessing workflows. *Genome Biol* **2021,** 22, (1), 339.

9. Alkan, C.; Coe, B. P.; Eichler, E. E., Genome structural variation discovery and genotyping. *Nat Rev Genet* **2011,** 12, (5), 363-76.

10. Liu, B.; Morrison, C. D.; Johnson, C. S.; Trump, D. L.; Qin, M.; Conroy, J. C.; Wang, J.; Liu, S., Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. *Oncotarget* **2013,** 4, (11), 1868-81.

11. Li, C.; Wu, H.; Guo, L.; Liu, D.; Yang, S.; Li, S.; Hua, K., Single-cell transcriptomics reveals cellular heterogeneity and molecular stratification of cervical cancer. *Commun Biol* **2022,** 5, (1), 1208.

12. McGinnis, C. S.; Murrow, L. M.; Gartner, Z. J., DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **2019,** 8, (4), 329-337 e4.

13. McGinnis, C. S.; Patterson, D. M.; Winkler, J.; Conrad, D. N.; Hein, M. Y.; Srivastava, V.; Hu, J. L.; Murrow, L. M.; Weissman, J. S.; Werb, Z.; Chow, E. D.; Gartner, Z. J., MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat Methods* **2019,** 16, (7), 619-626.

14. Huang, Q.; Liu, Y.; Du, Y.; Garmire, L. X., Evaluation of Cell Type Annotation R Packages on Single-cell RNA-seq Data. *Genomics Proteomics Bioinformatics* **2021,** 19, (2), 267-281.

15. Li, C.; Liu, B.; Kang, B.; Liu, Z.; Liu, Y.; Chen, C.; Ren, X.; Zhang, Z., SciBet as a portable and fast single cell type identifier. *Nat Commun* **2020,** 11, (1), 1818.

16. Franzen, O.; Gan, L. M.; Bjorkegren, J. L. M., PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)* **2019,** 2019.

17. Cai, J. J., scGEAToolbox: a Matlab toolbox for single-cell RNA sequencing data analysis. *Bioinformatics* **2019**.

18. De Falco, A.; Caruso, F.; Su, X. D.; Iavarone, A.; Ceccarelli, M., A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. *Nat Commun* **2023,** 14, (1), 1074.