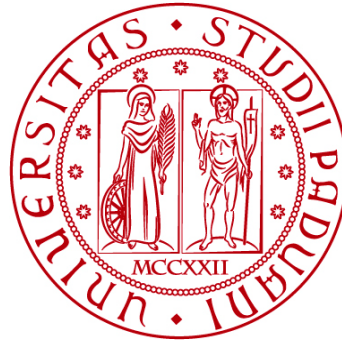


UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea in Biotecnologie



ELABORATO DI LAUREA

Analisi di dati di *single-cell* RNA sequencing derivati dall'atlante trascrittomico
del muscolo scheletrico umano pubblicato da Andrea J. De Micheli *et al.*

Tutor: Prof. Stefano Cagnin

Dipartimento di Biologia

Co-tutor: Dr. Giacomo Villa

Dipartimento di Biologia

Laureanda: Elisa Palombo

ANNO ACCADEMICO 2025/2026

Indice

1	ABSTRACT	5
2	INTRODUZIONE.....	6
2.1	INTRODUZIONE AL <i>SINGLE CELL</i> RNA <i>SEQUENCING</i>	6
2.2	ANALISI DATI.....	7
2.3	DESCRIZIONE <i>DATASET</i> UTILIZZATO PER LA METANALISI	9
3	APPROCCIO SPERIMENTALE.....	10
3.1	<i>PRE-PROCESSING</i> E <i>CLUSTERING</i> CON SEURAT	10
3.2	RIMOZIONE <i>DOUBLETS</i> CON DOUBLETFINDER	14
3.3	INTEGRAZIONE CON SEURAT	15
3.4	IDENTIFICAZIONE DEI <i>MARKER</i> CON SEURAT	16
4	RISULTATI E DISCUSSIONE.....	17
4.1	CONTROLLO QUALITÀ E <i>PRE-PROCESSING</i> DEI DATI	17
4.2	RIMOZIONE DEI <i>DOUBLETS</i>	18
4.3	INTEGRAZIONE DEI CAMPIONI	20
4.4	RIDUZIONE DELLA DIMENSIONALITÀ E <i>CLUSTERING</i>	23
4.5	ANNOTAZIONE DELLE POPOLAZIONI CELLULARI	24
4.6	CONCLUSIONI.....	25
5	BIBLIOGRAFIA.....	27

1 Abstract

I muscoli scheletrici umani sono uno dei tessuti più estesi di tutto l'organismo umano e svolgono funzioni essenziali per la sua omeostasi. Lo svolgimento di queste funzioni è permesso dalla presenza di diverse popolazioni cellulari, le quali possono essere studiate per comprenderne peculiarità e caratteristiche comuni.

Con la nascita del *single cell RNA sequencing* (scRNA-seq) è possibile studiare tali popolazioni cellulari, analizzandone l'espressione genica e ottenendo profili di espressione cellule specifici che ne consentono l'identificazione.

L'obiettivo del presente lavoro è stato quello di svolgere l'analisi trascrittomica di un *dataset* di scRNA-seq di muscolo umano, ottenuto dallo studio di Andrea J. De Micheli *et al.* (2020). Per svolgere l'analisi dei profili di espressione è stato utilizzato l'approccio bioinformatico di *clustering* cellulare di Seurat, che ha consentito l'annotazione dei cluster cellulari identificati, assegnando ad essi un'identità cellulare in base all'espressione di geni marcatori specifici.

Infine, i risultati ottenuti con Seurat sono stati confrontati con quelli dello studio di riferimento, ottenuti utilizzando Scanorama per lo step di integrazione, al fine di valutare le caratteristiche comuni identificate da entrambi gli approcci bioinformatici e quelle che invece vengono identificate in modo differenziale, e ne è emerso un discreto livello di analogia.

2 Introduzione

2.1 Introduzione al *single cell* RNA sequencing

L'analisi del trascrittoma di un campione biologico permette lo studio di tutti gli RNA, codificanti e non codificanti, presenti all'interno delle cellule analizzate. Questo approccio consente di determinare quali dei geni, tra quelli presenti all'interno del genoma, vengano trascritti in una determinata condizione biologica, fornendo informazioni sullo stato funzionale della cellula.

All'avvento degli studi di trascrittomica, quelle che venivano svolte erano analisi di RNA sequencing in *bulk*, ovvero analisi di sequenziamento di tutte le molecole di RNA estratte dai campioni di tessuto in esame. Da queste si ottengono profili di espressione genica medi, rappresentativi dell'intera popolazione cellulare presente all'interno del campione analizzato. Questo tipo di analisi può risultare utile per tessuti costituiti da popolazioni cellulari omogenee, ma quando si vogliono analizzare tessuti eterogenei, le analisi trascrittomiche in *bulk* portano ad una perdita di informazione sulla variabilità di espressione genica tra le diverse popolazioni cellulari e su quei trascritti che sono presenti in poche copie all'interno delle cellule, il cui segnale viene spesso mascherato dai trascritti che invece sono molto più abbondanti all'interno di esse.

Con il progresso delle conoscenze biomediche, approfondire le modificazioni che avvengono all'interno delle cellule durante le patologie è diventato sempre più importante, sia per la diagnosi che per la cura di esse. Data questa necessità è nato il *single cell* RNA sequencing, tecnica di sequenziamento del trascrittoma che permette di ottenere profili di espressione genica per ciascuna cellula appartenente al campione analizzato. In questo modo si ottengono profili trascrittomici più accurati, che permettono di distinguere firme trascrizionali uniche, fondamentali per l'identificazione e la catalogazione di nuovi tipi cellulari e per la distinzione tra cellule in condizioni fisiologiche o patologiche.

Nonostante questa tecnica sia notevolmente più costosa e richieda una maggiore esperienza dell'operatore rispetto ad una analisi in *bulk*, con l'evolversi di nuove tecnologie, dal 2009 a d'oggi è stato possibile ridurre notevolmente i costi e aumentarne la resa e l'efficienza, portando ad importanti avanzamenti nella ricerca.

Per ottenere profili di espressione a livello della singola cellula con il *single cell RNA sequencing* è necessario, innanzitutto, isolare le cellule dal campione di tessuto o dalla coltura in esame. Le cellule isolate vengono quindi sottoposte a delle fasi di preparazione del materiale genico, nelle quali avviene la loro digestione e la purificazione dell'RNA a partire dall'estratto cellulare.

L'RNA estratto verrà in seguito convertito in molecole di cDNA a doppio filamento, con l'aggiunta di sequenze tecniche che ne faciliteranno il sequenziamento e l'analisi. Tra questi vi è il *barcode* cellulare, una sequenza nucleotidica che serve per l'identificazione della cellula da cui si origina un trascritto. Una volta sequenziati i cDNA di una libreria trascrittomica, quelli che presentano lo stesso *barcode* vengono attribuiti alla medesima cellula di origine; raggruppandoli sarà quindi possibile creare il profilo di espressione genica specifico per essa.

Dopo aver ottenuto tutti i dati di sequenziamento e dopo aver allineato le sequenze al genoma di riferimento, ha inizio la fase di analisi bioinformatica, la quale porta alla formazione di *cluster* cellulari, ovvero raggruppamenti di cellule basati sulla similarità dei loro profili di espressione genica. I *cluster* ottenuti possono quindi essere annotati, ovvero nominati per identità cellulare, mediante il riconoscimento di specifici marcatori genici.

2.2 Analisi dati

L'analisi bioinformatica e il seguente *clustering* vengono solitamente effettuati mediante pacchetti sviluppati nei linguaggi di programmazione di Python o R. Nel presente lavoro, l'analisi delle cellule è stata svolta mediante il *workflow* di analisi implementato nel pacchetto Seurat, utilizzando il software RStudio, il quale supporta il linguaggio di programmazione R.

Nella *pipeline* di analisi fornita da Seurat si parte dal controllo qualità delle cellule: partendo dalla matrice di espressione, si valutano le cellule in base al numero di geni e alla percentuale di trascritti mitocondriali rilevati al loro interno, con il fine di eliminare le cellule di scarsa qualità. Segue una fase di normalizzazione e identificazione dei geni altamente variabili, rispetto alla loro espressione attesa, che risultano essere anche quelli maggiormente utili per l'analisi a valle.

I dati vengono successivamente sottoposti a *scaling* e analisi delle componenti principali (*Principal Component Analysis*, PCA), una tecnica che permette di

identificare le componenti che presentano una maggiore variabilità all'interno di tutta la matrice dei dati, le quali verranno utilizzate in seguito per rappresentare le cellule.

A partire da questa rappresentazione, grazie ad un algoritmo di *clustering*, verranno formati i *cluster* cellulari, caratterizzati dalla presenza al loro interno di cellule con profili di espressione simili.

In seguito, si usano delle tecniche di riduzione non lineare della dimensionalità per la visualizzazione dei *cluster*. Le più diffuse sono la tecnica UMAP (*Uniform Manifold Approximation and Projection*) e quella t-SNE (*t-distributed Stochastic Neighbor Embedding*). Tecniche come queste permettono la rappresentazione delle relazioni tra le cellule in uno spazio bidimensionale più semplice, facilitando la visualizzazione e l'identificazione di popolazioni cellulari caratterizzate da profili di espressione genica simili.

I *cluster* identificati vengono infine annotati mediante l'analisi di geni detti marcatori, i quali, essendo significativamente più espressi in un determinato tipo cellulare, permettono di assegnare una identità biologica ad ogni gruppo di cellule.

Spesso, analizzando diversi campioni, si ottengono dati molto eterogenei. Questo è anche dovuto a possibili variabili tecniche dei campioni, come differenti siti anatomici di origine, diverse condizioni di salute dei donatori, età, diverse procedure chirurgiche svolte per il loro ottenimento e tempo trascorso dalla raccolta al processamento. Per ridurre al minimo l'effetto di queste e altre variabili tecniche che si possono presentare, è necessaria l'applicazione di metodi bioinformatici di integrazione, che permettono il confronto dei profili di espressione genica tra campioni eterogenei.

I dati di *single cell RNA sequencing* sono inoltre, per loro natura, ricchi di rumore di fondo, dato dalla presenza di molti profili trascrittomici diversi per ciascuna cellula, ciascuno dei quali riporta l'espressione di decine di migliaia di geni. Per questo, anche la scelta del programma o *tool* utilizzato per il *processing* e dei parametri usati nell'analisi di tali dati possono andare ad influire sulle caratteristiche che vengono messe in risalto o che vengono mascherate all'interno di uno specifico set di dati.

L'obiettivo principale della presente analisi è quello di andare a confrontare i risultati ottenuti utilizzando due diversi approcci bioinformatici, quello combinatorio “*Seurat + Scanorama*” usato nello studio di riferimento e *Seurat “puro”*, usato da noi, al fine di valutare la concordanza dei risultati ottenuti e verificare la robustezza dell'identificazione delle diverse popolazioni cellulari.

2.3 Descrizione *dataset* utilizzato per la metanalisi

Per svolgere l'analisi bioinformatica, con il fine di identificare le diverse sottopopolazioni cellulari presenti all'interno del tessuto muscolare scheletrico, è stato usato l'atlante trascrittomico di scRNA-seq usato nello studio di Andrea J. De Micheli *et al.*, riportato nel database GEO (*Gene Expression Omnibus*) dell'NIH (*National Institute of Health*), sotto l'*accession number* GSE143704.

I dati compresi nel *dataset* derivano da 10 campioni biologici, provenienti da diversi distretti corporei, ciascuno ottenuto da un diverso donatore di età compresa tra i 41 e gli 81 anni, senza patologie pregresse rilevanti ai fini dello studio.

I campioni muscolari sono stati processati enzimaticamente e con il sistema di single cell RNA sequencing 10X Chromium, ottenendo i trascrittomi per 22.058 cellule (2206 ± 1961 per ciascun campione) riuniti in un unico *dataset* e analizzati utilizzando la tecnica dell'*unsupervised clustering*.

L'analisi bioinformatica condotta da De Micheli *et al.* ha permesso loro di individuare 16 popolazioni cellulari, grazie all'identificazione di specifici marcatori genici, riconducibili a cellule stromali, vascolari, immunitarie e satelliti (**Fig. 1**).

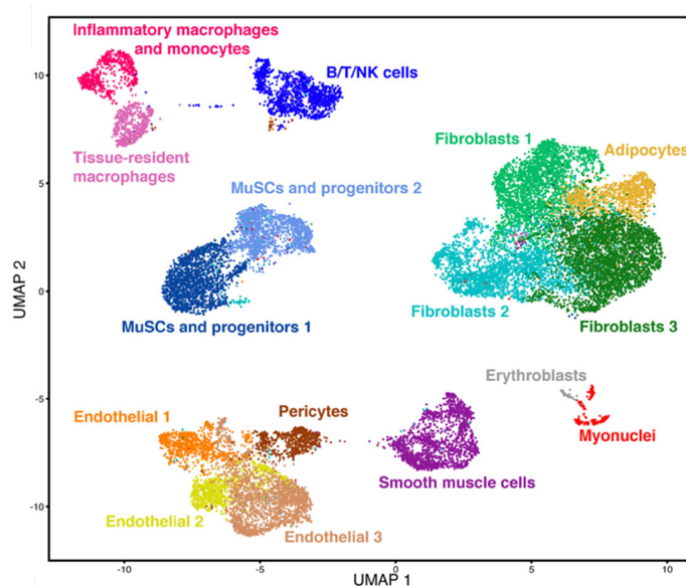


Figura 1 - Rappresentazione UMAP dei cluster cellulari identificati nello studio di De Micheli *et al.*

3 Approccio sperimentale

3.1 Pre-processing e clustering con Seurat

L'analisi bioinformatica dei dati di *single cell RNA sequencing* è stata eseguita utilizzando il linguaggio di programmazione R (versione 4.5.3) nell'ambiente di sviluppo RStudio (versione 2026.04.0+526). Le principali fasi di elaborazione dei dati, controllo qualità, normalizzazione, PCA, *clustering* e integrazione sono state svolte con il pacchetto Seurat (versione 5.5.0).

Per poter svolgere le analisi bioinformatiche sulla matrice di espressione genica, è necessario innanzitutto convertire tale matrice in un oggetto Seurat impiegando la funzione `CreateSeuratObject()`. Un oggetto Seurat è un elemento in grado di riportare in modo organizzato le informazioni principali relative ai nostri dati.

Nella prima fase di *quality control*, le cellule vengono selezionate in modo da cercare di eliminare quelle danneggiate o morenti, definite di bassa qualità, che possono andare ad influenzare negativamente i risultati dell'analisi. Visualizzando il *dataset* attraverso la funzione `VlnPlot()`, la quale crea un plot “a violino” (Fig. 2), con la funzione `subset()`, si selezionano le cellule in base al numero di “*features*” (geni espressi) e alla percentuale di trascritti mitocondriali che presentano.

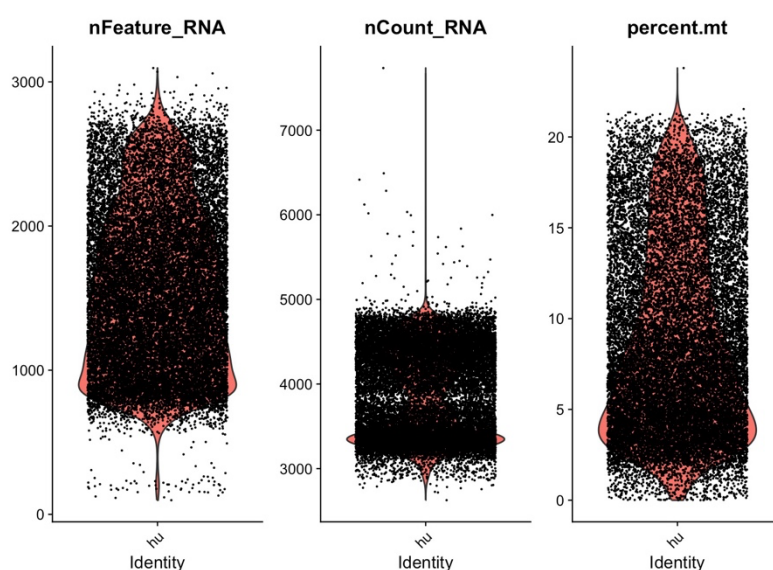


Figura 2 – Violin plot utilizzato per rappresentare i parametri di trascritti (*nCount*), di geni (*nFeature*) e percentuale di trascritti mitocondriali (*percent.mt*) prima del *quality control*.

La normalizzazione è stata svolta utilizzando la funzione implementata in Seurat `NormalizeData()`, la quale svolge una normalizzazione logaritmica; l'espressione genica di ciascuna cellula viene normalizzata rispetto al numero di trascritti totali rilevati nella stessa, poi moltiplicata per uno “*scale factor*” pari a 10.000 e infine sottoposta ad una trasformazione logaritmica. In questo modo, anche cellule che presentano un numero molto diverso di trascritti identificati al loro interno possono essere paragonate.

Dopo la normalizzazione si procede con l'identificazione delle “*variable features*” all'interno delle cellule, ovvero geni che presentano una variabilità di espressione maggiore rispetto a quella attesa, in funzione dei loro livelli medi di espressione.

I geni con elevata variabilità di espressione possono essere altamente informativi per la distinzione delle diverse popolazioni cellulari, in quanto questi possono essere parte di una specifica funzione biologica, tipica di un particolare tipo cellulare.

La funzione utilizzata per la ricerca delle *variable features* è `FindVariableFeatures()` ed è stata impostata per la ricerca dei 2000 geni maggiormente variabili all'interno dei dati (**Fig. 3**), i quali vengono memorizzati nell'oggetto Seurat per essere utilizzati in seguito nelle fasi di riduzione della dimensionalità.

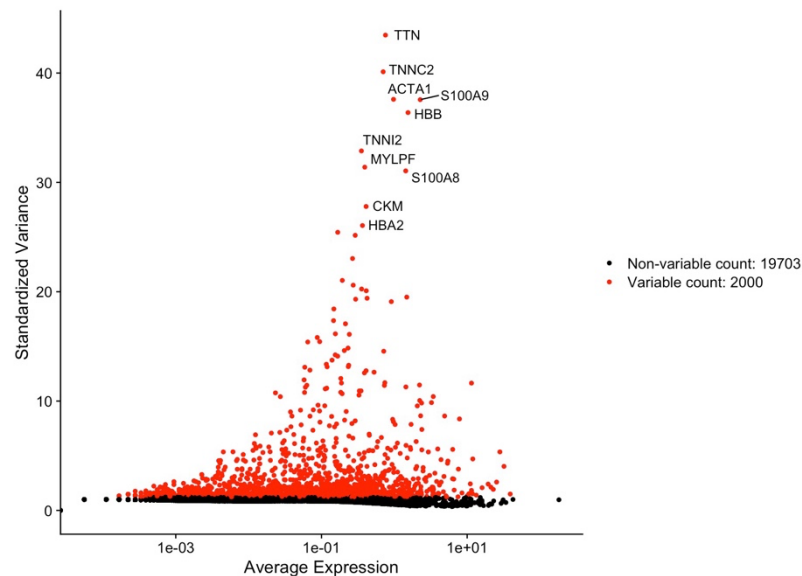


Figura 3 - Rappresentazione dei 2000 geni più variabili.

Con la funzione `ScaleData()` è stato svolto lo “*scaling*” dei dati, necessario per la standardizzazione dell’espressione di ogni gene. Con questo passaggio si riducono le differenze di scala nell’espressione dei diversi geni e si rendono possibili le successive analisi di riduzione della dimensionalità.

Con la *Principal Component Analysis* l’obiettivo è ridurre la dimensionalità del *dataset*, riunendo i geni che contribuiscono maggiormente alla variabilità biologica e tecnica in un numero più ridotto di variabili, dette componenti principali (*Principal Components*, PCs).

Il numero massimo di PCs calcolato mediante la funzione `RunPCA()` è pari a 50. Per la scelta del numero di componenti principali da utilizzare nelle fasi successive di analisi è stato esaminato il grafico “*a gomito*” (**Fig. 4**) creato con la funzione `ElbowPlot()`. Tale grafico riporta le componenti principali in funzione della quota di variabilità dei dati che esse spiegano. Il gomito del grafico rappresenta il punto in cui la varianza spiegata dalle PCs diventa meno pronunciata e di conseguenza meno informativa per le successive analisi. Per evitare di includere nelle analisi informazioni poco significative o riconducibili al rumore tecnico, solitamente si selezionano le PCs precedenti al gomito, ovvero quelle più importanti per spiegare la variabilità dei dati.

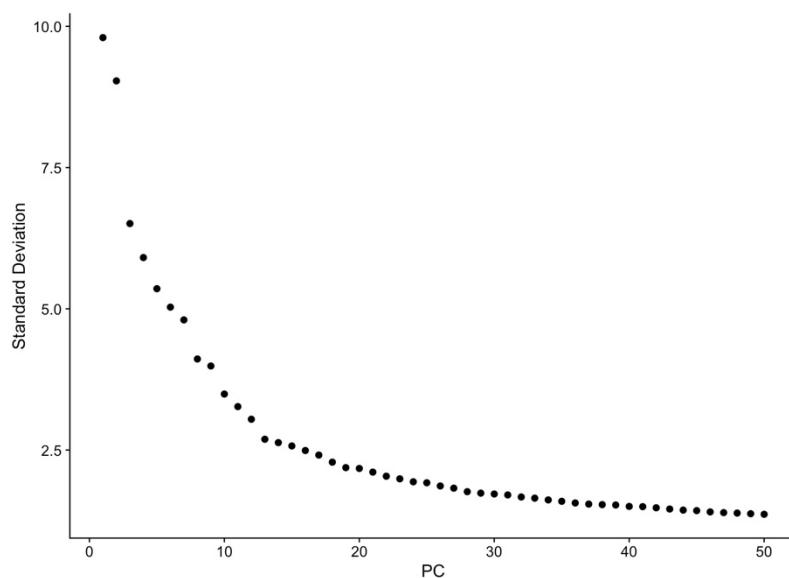


Figura 4 - Elbow plot che mostra le 50 PCs calcolate.

In seguito alla definizione del numero di PCs da utilizzare si passa alla fase vera e propria di “*clustering*”, che prevede l’applicazione di due diverse funzioni.

La prima funzione è `FindNeighbors()`, la quale costruisce un grafo di vicinanza tra le cellule in base alle componenti principali selezionate. La seconda funzione è `FindClusters()`, la quale applica l’algoritmo di Louvain, un metodo di “*community detection*” utilizzato per l’identificazione dei *cluster*; l’algoritmo identifica comunità di cellule strettamente connesse tra loro nel grafo di vicinanza sulla base dei loro profili di espressione genica. In questa seconda funzione è presente un parametro molto importante, quello della “*resolution*”, che determina la granularità dei *cluster* che si vengono a formare: all’aumentare del valore del parametro aumenta generalmente il numero di comunità identificate, viceversa, al diminuire del parametro diminuisce generalmente il numero di comunità identificate. Il valore della *resolution* è generalmente compreso tra 0,3 e 1,5.

Per la visualizzazione dei risultati è stata impiegata la tecnica di riduzione non lineare della dimensionalità UMAP, tramite la funzione `RunUMAP()`, la quale, in base al numero di PCs selezionate in precedenza, permette di rappresentare graficamente, in uno spazio bidimensionale, le relazioni di similarità di espressione genica identificate tra le cellule (**Fig. 5**).

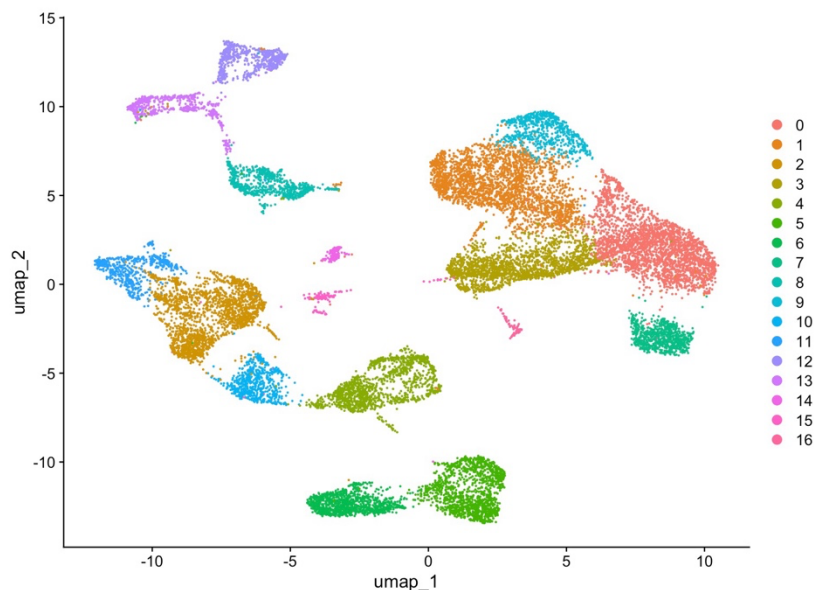


Figura 5 - UMAP rappresentante i cluster trovati prima della rimozione dei doublets e prima dell’integrazione.

3.2 Rimozione *doublets* con DoubletFinder

Uno dei problemi principali che si possono riscontrare durante la preparazione delle librerie di scRNA-seq, in modo particolare in quelle basate sulla microfluidica come 10X Chromium, è la creazione di “*doublets*”. Questi sono degli eventi in cui, all’interno della gocciolina, nell’emulsione acqua-in-olio creata durante la preparazione della libreria, ricadono due cellule distinte; il profilo di espressione genica che si ottiene dal sequenziamento di quella goccia rappresenta la combinazione dei profili delle due diverse cellule e questo profilo artificiale può andare a compromettere le seguenti fasi di analisi.

Per ridurre la presenza di questi artefatti tecnici è stata svolta l’identificazione dei *doublets* mediante il pacchetto DoubletFinder (versione 2.0.6).

Per svolgere questa analisi è stato necessario separare i 10 diversi campioni che costituiscono il *dataset* e svolgere per ognuno di essi la *pipeline* prevista dal pacchetto.

Per ogni campione vengono eseguite le funzioni `paramSweep()`, `summarizeSweep()` e `find.pK()` per definire il parametro `pK`, il quale non ha un significato biologico diretto, ma rappresenta un parametro computazionale utilizzato da DoubletFinder per distinguere tra le cellule reali e i *doublets* artificiali. Viene anche calcolato il numero di *doublets* attesi all’interno del *dataset*, calcolo basato sul numero di cellule totali e sul tasso di presenza di *doublets* previsto dalla piattaforma di sequenziamento utilizzata (5% nel 10X Chromium). Il valore ottenuto viene in seguito corretto per la possibilità di presenza di *doublets* omotipici, ovvero formati da due cellule appartenenti alla stessa popolazione cellulare.

Dopo aver ottenuto questi due valori è possibile eseguire la funzione `doubletFinder()` per ogni campione, la quale definisce ogni cellula come “*Singlet*” o “*Doublet*”, così da rimuovere queste ultime per le successive analisi.

3.3 Integrazione con Seurat

Poiché i dati ottenuti per la creazione dell’atlante trascrittomico provengono da campioni diversi, è stata eseguita l’integrazione mediante le funzioni implementate nel pacchetto Seurat.

L’integrazione serve per ridurre al minimo gli effetti dovuti alla variabilità tecnica (*batch effect*), pur mantenendo le differenze biologiche reali dei dati. In Seurat, per svolgere l’integrazione bisogna separare i dati per campione e svolgere su ciascuno di essi le fasi di normalizzazione e ricerca delle *variable features*. La prima funzione utilizzata per l’integrazione è `SelectIntegrationFeatures()`, la quale identifica i geni altamente variabili condivisi tra i diversi campioni, così da poterli utilizzare come riferimento per identificare popolazioni cellulari simili all’interno di ciascun campione.

La seconda funzione utilizzata, `FindIntegrationAnchors()`, individua le “*integration anchors*”, ovvero coppie di cellule provenienti da *dataset* diversi, le quali presentano profili di espressione altamente simili, presumibilmente appartenenti alla stessa popolazione cellulare.

Infine, con la funzione `IntegrateData()`, si vanno ad utilizzare le *anchors* per allineare i diversi *dataset* e riunirli in un unico oggetto Seurat da utilizzare per le seguenti analisi di *clustering* e annotazione cellulare.

L’oggetto Seurat integrato ottenuto è stato successivamente utilizzato in un nuovo ciclo di *pre-processing*, di riduzione della dimensionalità, *clustering* e identificazione dei geni *marker* senza *batch effect*.

Rispetto alla fase iniziale di *pre-processing* e *clustering*, sono stati utilizzati parametri leggermente differenti, con l’obiettivo di migliorare la separazione delle popolazioni cellulari dopo l’integrazione dei dati.

3.4 Identificazione dei *marker* con Seurat

Per l'identificazione dei geni *marker* all'interno dei *cluster* identificati è stata utilizzata la funzione `FindAllMarkers()` del pacchetto Seurat. Questa funzione svolge il confronto di ciascun *cluster* con tutte le altre cellule del *dataset* e restituisce i geni che risultano significativamente più espressi nel *cluster* di riferimento rispetto agli altri e che quindi possono essere utilizzati per l'annotazione delle differenti popolazioni cellulari.

Ad ogni gene *marker* identificato viene applicato il *Receiver Operating Characteristic* (ROC) test, il quale valuta la capacità di un gene di discriminare una cellula appartenente ad un gruppo rispetto ad un altro. Il test restituisce un valore AUC (*Area Under the Curve*) compreso tra 0 e 1; un valore prossimo ad 1 indica che il gene in questione è espresso quasi esclusivamente nelle cellule appartenenti al *cluster* di interesse, 0,5 rappresenta un gene che non distingue un *cluster* da un altro, mentre 0 rappresenta un gene che è maggiormente espresso negli altri *cluster*, escluso quello di interesse.

4 Risultati e discussione

L'obiettivo della presente analisi è stato quello di caratterizzare le principali popolazioni cellulari presenti all'interno del muscolo scheletrico umano mediante l'utilizzo dei dati di *single cell RNA sequencing* presenti nell'atlante trascrittomico pubblicato da De Micheli *et al.* L'analisi è stata condotta mediante l'utilizzo del software Seurat ed ha compreso diverse fasi, quali controllo qualità, rimozione dei *doublers*, integrazione dei campioni, riduzione della dimensionalità, *clustering* e annotazione delle popolazioni cellulari.

I risultati ottenuti sono stati confrontati con quelli riportati nello studio di riferimento, al fine di valutare la concordanza tra i risultati ottenuti e la robustezza dell'approccio bioinformatico utilizzato per il loro ottenimento.

4.1 Controllo qualità e *pre-processing* dei dati

La fase di controllo qualità rappresenta un passaggio fondamentale per l'analisi dei dati di *single cell RNA sequencing*, poiché consente di individuare ed eliminare le cellule che, durante la preparazione della libreria trascrittomico risultano essere danneggiate oppure morte, le quali presentano dei profili di espressione genica alterati che possono andare ad influenzare le analisi a valle.

Osservando la distribuzione dei parametri mediante *violin plot* (**Fig. 6**), sono state selezionate le cellule che presentano un numero di geni espressi compreso tra 700 e 2800 e una percentuale di trascritti mitocondriali inferiore al 15%, ottenendo un sottogruppo di 18.033 cellule del *dataset* iniziale.

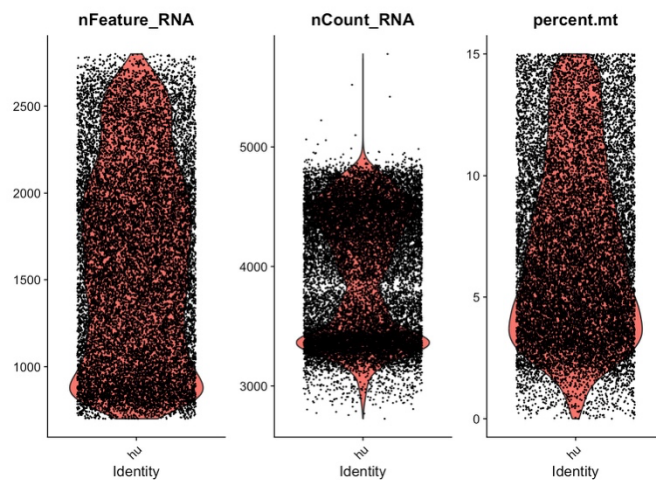


Figura 6 – Violin plot mostra i valori soglia usati per escludere le cellule di bassa qualità durante la fase di rimozione dei *doublers*.

Cellule con un numero elevato di geni espressi vengono escluse in quanto si presume che possano rappresentare dei *doublets*, ovvero goccioline di sequenziamento che contengono più di una cellula. Al contrario, cellule che presentano un numero di geni espressi molto basso o con una elevata percentuale di trascritti mitocondriali sono indicative di cellule danneggiate o morte.

In entrambi i casi, quelli che si ottengono sono profili di espressione alterati, non rappresentativi, che pertanto vengono rimossi dall'analisi.

Dopo aver svolto normalizzazione, identificazione delle *variable features*, lo *scaling* dei dati e la *Principal Component Analysis* avviene la fase di *clustering* delle cellule. In questa fase vengono definiti i parametri *dims* e *resolution*, necessari per la definizione del numero e la composizione dei *cluster* ottenuti.

La scelta del parametro *dims* viene fatta sulla base del grafico a gomito (**Fig. 4**), che consente di individuare il numero di componenti principali che spiegano la maggior parte della variabilità presente nei dati. Tuttavia, questo parametro può essere aggiustato empiricamente valutando la qualità della separazione tra i *cluster* ottenuti. Analogamente, anche il parametro *resolution* viene definito empiricamente, in funzione al numero di *cluster* ottenuti.

Nella presente analisi, dopo aver valutato diverse combinazioni di parametri, la combinazione di *dims* pari a 12 e *resolution* pari a 0,3 è risultata quella più appropriata per separazione delle popolazioni cellulari, senza generare *cluster* eccessivamente frammentati, portando alla formazione di 17 cluster cellulari (**Fig. 5**).

4.2 Rimozione dei *doublets*

Per la rimozione dei *doublets*, il *dataset* è stato diviso in base al campione di provenienza, ottenendo così 10 diversi campioni riuniti in una lista di oggetti. Questa sarà la lista sulla quale verranno svolte le seguenti fasi per la rimozione dei *doublets*.

Per ciascun campione sono state ripetute le fasi di normalizzazione, identificazione delle *variable features*, *scaling* dei dati e PCA, al fine di individuare per ognuno il numero ottimale di componenti principali da utilizzare nelle successive analisi. Per ogni campione, inoltre, è stato svolto anche il *clustering*, con parametri specifici per ciascuno.

Dopo il *pre-processing* comincia la vera e propria fase per la rimozione dei *doublets*. Con le PCs selezionate in precedenza vengono eseguite le funzioni `paramSweep()`, `summarizeSweep()` e `find.pK()`, le quali restituiscono, per ogni campione, una lista di valori pK, i quali vengono visualizzati sotto forma di grafico. Per ogni campione dovrà essere selezionato il valore di pK più alto (**Tab. 1**).

In seguito, per ogni campione viene calcolato il numero atteso di *doublets* al loro interno (nExp) (**Tab. 1**). Il calcolo viene fatto moltiplicando la percentuale di formazione di *doublets*, specifica del metodo di sequenziamento utilizzato (5% nel 10X Chromium), per il numero di cellule presenti all'interno del campione. Il numero ottenuto viene poi corretto per tenere conto della presenza di possibili “*doublets omotipici*”, ovvero formati da cellule appartenenti alla stessa popolazione cellulare.

Campione	Sample ID	PCs	Resolution	pK	nExp
1	092618	15	0,7	0,22	138
2	092712	15	0,5	0,02	41
3	101018	15	0,5	0,005	129
4	111418	17	0,5	0,08	80
5	120518	10	0,5	0,005	206
6	021219	15	0,5	0,03	30
7	061419	10	0,5	0,28	21
8	062419	10	0,5	0,02	28
9	062719	10	0,5	0,11	8
10	080619	20	0,7	0,17	28

Tabella 1 – Riporta per ogni campione appartenente al dataset il Sample ID e i valori di Principal Components (PCs), resolution utilizzata, pK e numero di doublets attesi (nExp).

Infine, è stata eseguita la funzione `doubletFinder()` per ogni campione (**Fig. 7**), la quale richiede l'inserimento dei valori di PCs, pK e nExp specifici per ogni campione. Alla fine dell'esecuzione, all'interno dei metadati di ogni campione, le cellule sono state classificate come *Singlet* o *Doublet*.

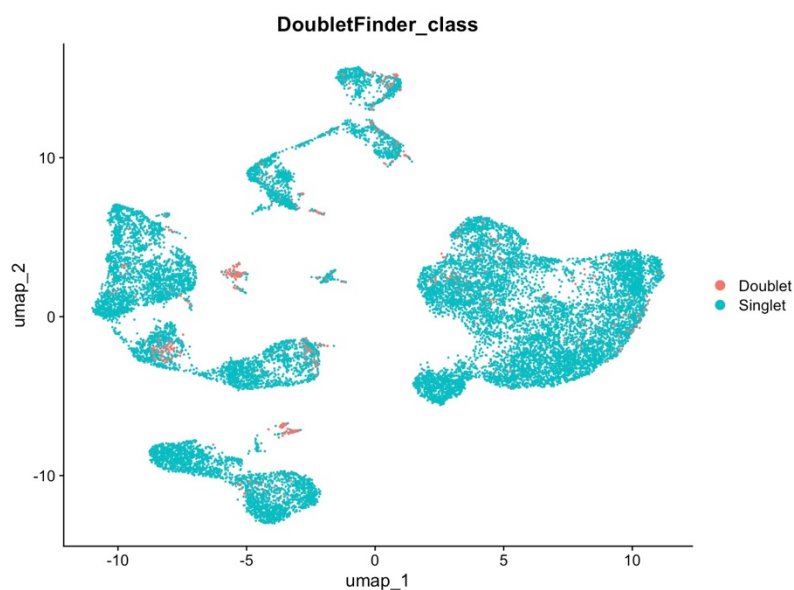


Figura 7 - UMAP evidenzia la distinzione tra singlets e doublets.

Successivamente, i campioni sono stati riuniti all'interno di un unico oggetto Seurat e le cellule classificate come *Doublet* sono state rimosse. Questa procedura ha portato all'eliminazione di 709 cellule, ottenendo quindi un totale di 17.324 cellule classificate come *Singlet*.

4.3 Integrazione dei campioni

Partendo dal nuovo oggetto Seurat contenente esclusivamente i *singlet* vengono svolte nuovamente le fasi di *quality control*, normalizzazione, identificazione delle *variable features*, *scaling* dei dati e *Principal Components Analysis*. Per la costruzione dell'oggetto Seurat integrato sono stati usati criteri di qualità più stringenti, con l'obiettivo di aumentare l'affidabilità delle cellule mantenute. Sono state selezionate le cellule che presentano un numero di geni espressi compreso tra 700 e 2700, si è quindi diminuita la soglia massima per diminuire la probabilità di mantenere *doublets* residui, e, parallelamente, cellule che presentano una percentuale di trascritti mitocondriali inferiore al 10%, valore soglia appropriato per il controllo qualità di *dataset* umani di *single cell RNA sequencing*. In conclusione, si è ottenuto un sottogruppo di 12.993 cellule (**Fig. 8**).

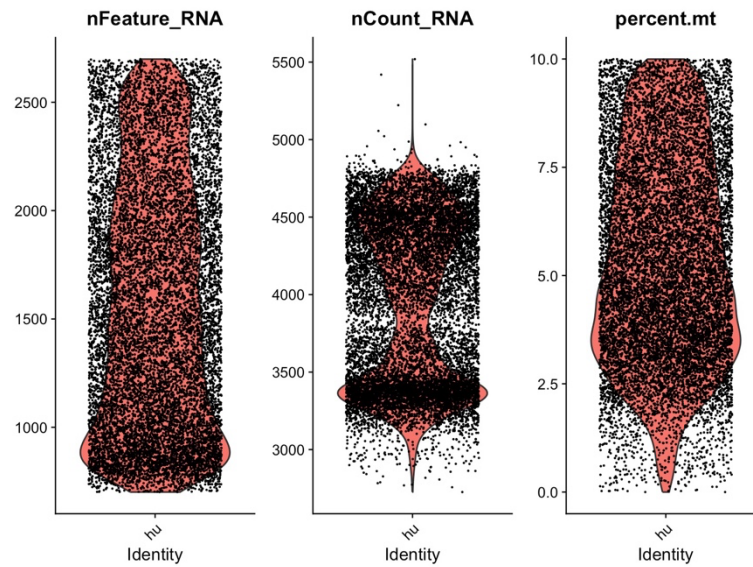


Figura 8 - Violin plot mostra i valori soglia utilizzati per escludere le cellule di bassa qualità in fase di integrazione.

Anche per la riduzione della dimensionalità e il *clustering* i parametri sono stati variati. Il numero di componenti principali considerate è passato a 25, mentre la *resolution* per la formazione dei *cluster* è rimasta sempre a 0,3.

Creando una rappresentazione mediante *Uniform Manifold Approximation and Projection* (UMAP), dove le cellule sono state colorate in base al campione di origine per verificare la presenza di *effetti batch* all'interno del *dataset*, è possibile osservare la presenza di *cluster* cellulari formati, per la maggior parte, da cellule provenienti da un unico campione (**Fig. 9**). La presenza di questi gruppi cellulari caratterizzati da cellule derivanti da un principale campione suggerisce che le differenze osservate non riflettano le differenze biologiche tra le cellule, ma bensì quelle tecniche presenti tra i diversi campioni.

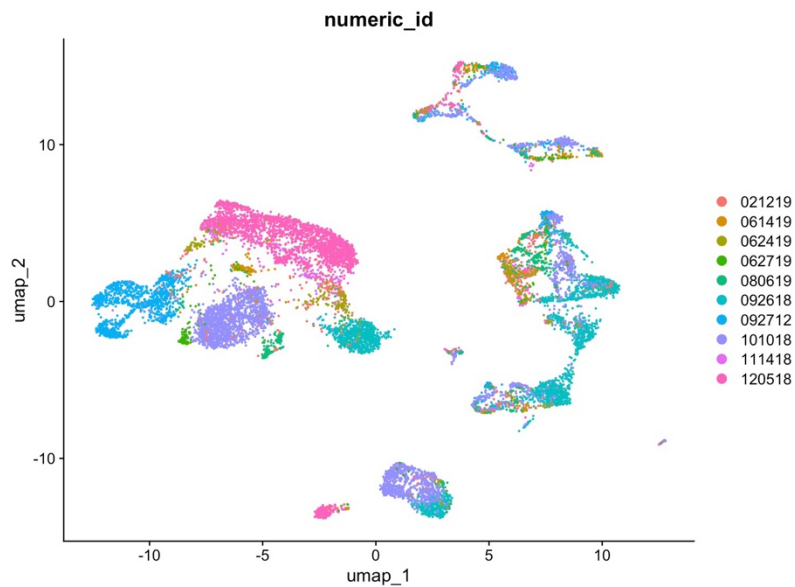


Figura 9 - UMAP mostra le cellule organizzate in cluster e colorate per campione di origine prima dell'integrazione.

Per svolgere l'integrazione, il *dataset* è stato nuovamente diviso in diversi oggetti in base al campione di provenienza e per ogni nuovo oggetto vengono svolte le fasi di normalizzazione e ricerca delle *variable features*.

Dopodiché vengono svolte le funzioni `SelectIntegrationFeatures()` e `FindIntegrationAnchors()` per l'identificazione delle “*integration anchors*” e con la funzione `IntegrateData()` si vanno a riunire i dati integrati all'interno di un unico oggetto Seurat.

Al termine dell'integrazione è stata creata una nuova UMAP (**Fig. 10**) ed è possibile osservare una mescolanza notevolmente migliore tra cellule provenienti da campioni diversi, il che indica una riduzione degli *effetti batch* all'interno del *dataset*, consentendo il confronto delle cellule sulla base delle loro differenze biologiche.

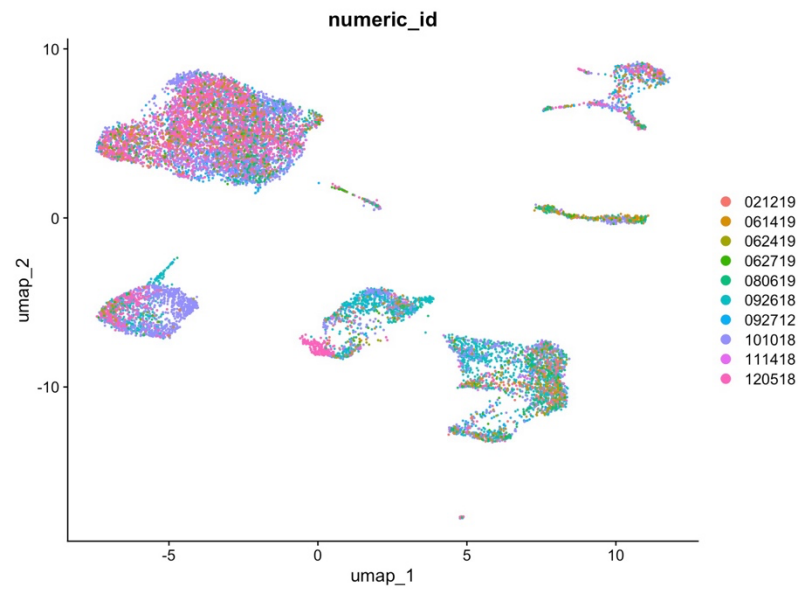


Figura 10 - UMAP mostra le cellule organizzate in cluster e colorate per campione di origine dopo l'integrazione.

4.4 Riduzione della dimensionalità e *clustering*

Con il nuovo oggetto Seurat integrato viene svolto lo *scaling* dei dati, la *Principal Components Analysis* e il *clustering*. Per la riduzione della dimensionalità e il *clustering* i parametri sono stati nuovamente modificati per meglio adattarsi ai dati integrati. Il numero di PCs selezionate è passato a 30 e il parametro della *resolution* è aumentato a 0,6. Questi nuovi parametri hanno permesso di ottenere, a partire dai dati integrati, 16 *cluster* cellulari distinti (**Fig. 11**).

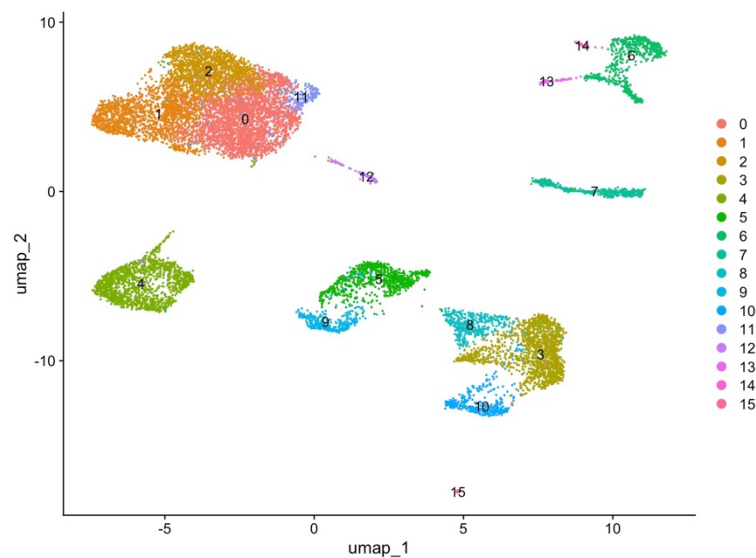


Figura 11 - UMAP rappresenta i 16 cluster trovati dopo la rimozione dei doublets e l'integrazione.

4.5 Annotazione delle popolazioni cellulari

Dopo aver ottenuto i *cluster* cellulari dall'analisi dell'oggetto Seurat integrato si procede con l'identificazione dei geni *marker* specifici per ciascuno di essi. L'identità biologica dei gruppi cellulari è stata assegnata confrontando i marcatori ottenuti con quelli riportati nell'atlante trascrittomico del muscolo scheletrico umano di De Micheli *et al.* e all'interno dei database CellMarker 2.0 (Hu *et al.*, 2023) e PanglaoDB (Franzén *et al.*, 2019).

Le popolazioni cellulari identificate appartengono ai principali compartimenti del muscolo scheletrico umano.

I *cluster* cellulari 0, 1, 2 e 11 sono stati attribuiti al compartimento stromale. In modo specifico, le cellule appartenenti a questi cluster vengono identificate come cellule progenitrici fibro-adipogeniche (FAPs), una popolazione cellulare mesenchimale specifica del muscolo scheletrico che svolge un ruolo fondamentale nella regolazione della rigenerazione del tessuto muscolare e che presenta la capacità di differenziarsi verso fenotipi adipogenici e fibrogenici. I principali *marker* che hanno permesso la loro identificazione sono stati PDGFRA (*Platelet-Derived Growth Factor Receptor Alpha*), DCN (*Decorin*), APOD (*Apolipoprotein D*) e CXCL14 (*C-X-C motif Chemokine 14*).

I *cluster* 4 e 12 sono stati attribuiti al compartimento muscolare. Il *cluster* 4 è stato ricondotto alle cellule staminali muscolari (MuSC), anche dette cellule satellite, grazie alla presenza del *marker* MYF5 (*Myogenic Factor 5*), uno dei principali *marker* per la loro identificazione. Il *cluster* 12 è stato invece associato a cellule muscolari di tipo II, in base all'espressione di geni *marker* come ATP2A1 (*Sarcoplasmic/endoplasmic reticulum calcium ATPase 1*), MYH2 (*Myosin-2*), ACTA1 (*Actin, alpha skeletal muscle*) e TNNT3 (*Troponin T, fast skeletal muscle*).

I *cluster* 6, 7, 13 e 14 sono stati attribuiti al compartimento immunitario. I *cluster* 6 e 14 sono stati associati a cellule appartenenti alla linea mieloide, comprendente macrofagi e monociti, in base alla presenza di *marker* come C1QA (*Complement C1q subcomponent subunit A*), CD74 (*HLA class II histocompatibility antigen gamma chain*), S100A9 (*Protein S100-A9*) e LYZ (*Lysozyme C*). Invece, i *cluster* 7 e 13 sono stati ricondotti a cellule appartenenti alla linea linfoide, come linfociti B, linfociti T e cellule Natural Killer. Questi sono stati identificati grazie alla presenza

di *marker* quali PTPRC (*Receptor-Type Tyrosine-Protein Phosphatase C*), SRGN (*Serglycin*) e GZMB (*Granzyme B*).

Infine, i *cluster* 3, 5, 8, 9, 10 e 15 sono stati attribuiti al compartimento vascolare, comprendente principalmente cellule endoteliali, periciti e cellule muscolari lisce. I *cluster* di cellule endoteliali e periciti sono stati identificati dalla presenza di *marker* quali PECAM1 (*Platelet Endothelial Cell Adhesion Molecule*), VWF (*von Willebrand factor*) e RGS5 (*Regulator of G protein Signaling 5*). Il *cluster* 9 è stato invece ricondotto alle cellule muscolari lisce (*Smooth Muscle Cells*, SMCs). Sebbene queste presentino molti *marker* in comune con i periciti, la presenza nel *cluster* di geni *marker* come ACTA2 (*Actin, aortic smooth muscle*), TAGLN (*Transgelin*) e MYH11 (*Myosin-11*) ha permesso di distinguerle dalla popolazione pericitica.

4.6 Conclusioni

Dal confronto tra i due approcci di analisi dei dati di *single cell RNA sequencing* ne emerge che, nel caso del presente lavoro, nonostante le principali popolazioni cellulari identificate (stromali, vascolari, immunitarie e muscolari) siano concordanti, la distinzione più fine tra diversi tipi cellulari appartenenti alla stessa categoria non risulta sempre semplice.

Nello studio di riferimento, dove è previsto l'uso dello strumento di integrazione fornito da Scanorama, è stato possibile distinguere in modo accurato alcune popolazioni cellulari. De Micheli *et al.* hanno evidenziato la presenza di due sottopopolazioni di cellule staminali muscolari (MuSC 1 e MuSC 2), mentre nel presente lavoro queste ricadono entrambe nel *cluster* 4. Mentre nel presente lavoro i *cluster* 0, 1, 2 e 11 ricadono nella classe delle FAPs, nel lavoro di riferimento queste vengono distinte in Fibroblasti 1, Fibroblasti 2, Fibroblasti 3 e Adipociti (**Fig. 1**). Inoltre, nello studio di De Micheli *et al.* viene evidenziata la presenza di popolazioni di mionuclei ed eritroblasti, mentre nella presente analisi viene identificata una popolazione di cellule muscolari di tipo II unica (*cluster* 12).

Le differenze nella specificità d'identificazione delle popolazioni cellulari possono dipendere da diversi fattori. Bisogna tenere conto del diverso approccio bioinformatico utilizzato (“*Seurat + Scanorama*” e “*Seurat puro*”), in quanto i diversi algoritmi di analisi possono andare a favorire la comparsa di alcune

caratteristiche a scapito di altre. Inoltre, anche i parametri utilizzati possono aver influenzato i risultati finali. Ad esempio, il controllo qualità svolto sulle cellule, che da 22.058 ha portato al mantenimento di solo 12.993 di esse, può essere stato utile per diminuire il rumore di fondo dato da possibili errori di procedura o di sequenziamento, ma potrebbe anche aver escluso profili di espressione genica importanti, di fatto andando a diminuire la risoluzione per le popolazioni cellulari meno abbondanti.

5 Bibliografia

De Micheli, A.J., Spector, J.A., Elemento, O. et al. (2020) 'A reference single-cell transcriptomic atlas of human skeletal muscle tissue reveals bifurcated muscle stem cell populations', **Skeletal Muscle**, 10, p. 19. Available at: <https://doi.org/10.1186/s13395-020-00236-3>.

Franzén, O., Gan, L.-M. and Björkegren, J.L.M. (2019) 'PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data', **Database**, 2019, baz046. Available at: <https://doi.org/10.1093/database/baz046>.

Hie, B. L., Kim, S., Rando, T. A., Bryson, B., & Berger, B. (2024). Scanorama: integrating large and diverse single-cell transcriptomic datasets. *Nature protocols*, 19(8), 2283–2297. <https://doi.org/10.1038/s41596-024-00991-3>.

Hu, C., Li, T., Xu, Y., Zhang, X., Li, F., Bai, J., Chen, J., Jiang, W., Yang, K., Ou, Q., Li, X., Wang, P. and Zhang, Y. (2023) 'CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data', **Nucleic Acids Research**, 51(D1), pp. D870-D876. Available at: <https://doi.org/10.1093/nar/gkac947>.

Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F. and Luo, Y. (2022) 'Single-cell RNA sequencing technologies and applications: A brief overview', **Clinical and Translational Medicine**, 12, e694. Available at: <https://doi.org/10.1002/ctm2.694>.

McGinnis, C.S., Murrow, L.M. and Gartner, Z.J. (2019) 'DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors', **Cell Systems**, 8(4), pp. 329-337.e4. Available at: <https://doi.org/10.1016/j.cels.2019.03.003>.

- Molina, T., Fabre, P. and Dumont, N.A. (2021) 'Fibro-adipogenic progenitors in skeletal muscle homeostasis, regeneration and diseases', *Open Biology*, 11(12), 210110. Available at: <https://doi.org/10.1098/rsob.210110>.
- Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. and Regev, A. (2015) 'Spatial reconstruction of single-cell gene expression data', *Nature Biotechnology*, 33(5), pp. 495-502. Available at: <https://doi.org/10.1038/nbt.3192>.
- Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M. III, Hao, Y., Stoeckius, M., Smibert, P. and Satija, R. (2019) 'Comprehensive integration of single-cell data', *Cell*, 177(7), pp. 1888-1902.e21. Available at: <https://doi.org/10.1016/j.cell.2019.05.031>.