POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Matematica: Modelli Matematici e Simulazioni Numeriche

Analisi dei barcodes cellulari e modellazione della dinamica di crescita, selezione e campionamento

Relatori:
Prof. Marco Scianna
Prof. Marcello Edoardo Delitala

Candidata:
Carlotta Sgro

Anno Accademico 2025-2026

Indice

Introduzione **4**

1 Analisi dei dati **7**
1.1 L'esperimento e i dati raccolti 7
1.2 Metodi 12
1.3 Analisi della sovrapposizione clonale 12
1.4 Analisi delle frequenze di lettura 16
1.5 Analisi delle frequenze cumulate in funzione del numero di barcodes considerati 23
1.6 Legge di potenza 26
1.7 Confronto delle abbondanze clonali 31
1.8 Analisi della coerenza del ranking clonale 33
1.9 Possibili interpretazioni dei dati collezionati 35

2 Simulazione dell'esperimento **44**
2.1 Ipotesi di crescita malthusiana 44
2.2 Distribuzione dei tassi intrinseci di crescita 47

3 Conclusioni e possibili sviluppi **53**

Bibliografia **58**

Introduzione

Come indicato nel Report Istat sulle cause di morte in Italia nel 2022, i principali fattori di decesso sono le malattie cardiocircolatorie e i tumori, che insieme causano circa il 55% dei decessi totali [1]. Tuttavia, dal rapporto *"I numeri del cancro 2024"* si evince che il numero di nuovi casi di diagnosi tumorale è stabile rispetto al 2022 e al 2023; inoltre, il tasso di mortalità mostra una diminuzione negli ultimi anni [2]. Questi miglioramenti possono essere imputabili a cure sempre più efficaci, mirate e a una prevenzione a cui aderiscono sempre più persone.

Nonostante i notevoli progressi nelle strategie terapeutiche, l'efficacia a lungo termine dei trattamenti è spesso compromessa dall'insorgenza di meccanismi di resistenza nelle cellule tumorali. Per comprendere se essa sia di origine epigenetica o genetica, i ricercatori dell'Istituto per la Ricerca Contro il Cancro di Candiolo (IRCC) hanno svolto esperimenti su cellule tumorali coltivate in vitro, utilizzando la tecnica del *barcoding cellulare*. Nel primo caso, l'insorgenza sarebbe dovuta all'acquisizione di mutazioni che si presentano durante la terapia, mentre nel secondo caso si ipotizza la presenza di cloni resistenti pre-esistenti nella massa tumorale. Quest'ultima ipotesi riveste implicazioni cliniche significative, in quanto le strategie terapeutiche andrebbero modificate al fine di identificare ed eliminare tali cellule resistenti [3].

Il *barcoding cellulare* è una tecnica che trova applicazione in diversi ambiti di studio, come l'emopoiesi, lo sviluppo cellulare, il cancro e le dinamiche infettive. Essa prevede l'etichettamento di singole cellule tramite una sequenza unica e casuale o semicasuale di acido nucleico chiamata *barcode*. In questo modo, grazie al numero praticamente illimitato di possibili sequenze, è possibile tracciare grandi quantità di cellule attraverso lo spazio, il tempo e la divisione cellulare, e quindi analizzare popolazioni eterogenee di cellule [4],[5].

L'introduzione del barcode avviene in diversi modi: il più semplice è l'assegnazione manuale di una cellula alla volta; questo metodo è molto efficace, ma richiede un elevato onere di lavoro. Il metodo più efficiente e robusto consiste nella produzione di un insieme di barcodes vettoriali in vitro, come plasmidi o virus, ottimizzati per il trasferimento di al più pochi barcodes per cellula e l'etichettamento del solo numero desiderato di cellule. Grazie al maggior numero di barcodes rispetto alle cellule considerate, è molto probabile che ognuna di esse venga classificata in modo univoco. Successivamente, la popolazione di cellule viene lasciata libera di crescere ed evolvere in vitro, infatti, essendo il barcode

integrato nel genoma cellulare, le cellule figlie presenteranno lo stesso codice della cellula madre [4], [5].

Il processo di lettura dei barcodes viene detto *screening* e avviene tramite estrazione di acido nucleico e successiva amplificazione e codifica. Ciò può essere compiuto tramite diversi metodi, tuttavia, ognuno di essi è soggetto a errori di codifica da tenere in considerazione durante l'analisi della popolazione di cellule considerata [4].

Infatti, per poter determinare delle relazioni tra i vari tipi di cellule è necessaria un'elevata confidenza nello stabilire che un dato barcode sia accurato e non frutto di errori di sequenziamento o scarso campionamento. A questo scopo, generalmente, vengono prelevati dalla popolazione alcuni campioni di pari dimensioni e nello stesso momento, detti *replicati tecnici*, ai quali viene successivamente applicato il sequenziamento. Ogni porzione dovrebbe quindi contenere circa lo stesso numero di barcodes differenti e in proporzioni simili. Se si ottiene una buona correlazione i dati sono affidabili e quindi è possibile determinare relazioni con altre cellule [5].

Comunemente, i dati vengono sottoposti anche ad altre procedure di filtraggio, come ad esempio l'esclusione delle cellule aventi un numero di letture al di sotto di una certa frequenza. L'obiettivo è ottenere un insieme di barcodes affidabile, sul quale sia possibile assumere relazioni biologiche con maggiore confidenza [5].

In seguito, è necessario utilizzare strumenti statistici adeguati per analizzare e interpretare correttamente i dati ottenuti e descrivere la distribuzione dei barcodes, la loro persistenza nel tempo e i pattern di selezione emergenti. In questo contesto, l'analisi quantitativa dei dati consente anche di stimare i tassi di crescita associati ai diversi cloni cellulari. A questo scopo, un approccio comunemente utilizzato nello studio di popolazioni cellulari è quello basato sul modello malthusiano, che descrive la dinamica di una popolazione attraverso una crescita esponenziale.

Combinando queste analisi, l'elaborato mira quindi a caratterizzare in che modo i processi di crescita, selezione e campionamento influenzino la sopravvivenza e la diversità clonale osservate e come tali dinamiche possano essere ricondotte a differenze nei tassi di espansione stimati tramite modelli matematici.

Capitolo 1

Analisi dei dati

1.1 L'esperimento e i dati raccolti

I ricercatori dell'Istituto per la Ricerca Contro il Cancro di Candiolo (IRCC) hanno svolto un esperimento con lo scopo di comprendere se l'insorgenza di meccanismi di resistenza alle terapie da parte di cellule cancerogene fosse dovuta a origini di tipo genetico o epigenetico.

Inizialmente, hanno considerato una popolazione di $1M$ di cellule, la quale, in data 26 gennaio, è stata etichettata tramite l'inserimento di un'ampia libreria di barcodes differenti nel genoma cellulare, tramite infezione lentivirale. Il metodo del barcoding utilizza un'infezione causata da un *lentivirus*, cioè un *retrovirus* come l'HIV, il cui RNA viene retrotrascritto in DNA e integrato nel genoma delle cellule bersaglio. Il genoma lentivirale integrato è costituito da una sequenza di DNA casuale che identifica il barcode, da elementi lentivirali che permettono l'infezione e la sua integrazione, e da un *gene reporter* che conferisce alle cellule la resistenza a uno specifico antibiotico o la capacità di esprimere una proteina fluorescente. Quest'ultima proprietà consente di isolare e contare le cellule che esprimono il marcatore fluorescente tramite FACS (*Fluorescence Activated Cell Sorting*), un tipo di citometria a flusso in cui le singole cellule vengono isolate in goccioline elettricamente cariche o meno a seconda dell'espressione del marcatore[6].

Nell'ambito del barcoding cellulare, un parametro regolabile è la *molteplicità di infezione* (MOI), che rappresenta il rapporto tra il numero di particelle virali e il numero di cellule bersaglio. Un MOI elevato, come nel caso in esame ($MOI = 10$), implica che ciascun barcode sia veicolato da molte particelle virali, aumentando la probabilità che la libreria venga rappresentata in modo completo nella popolazione infettata. Tuttavia, un MOI elevato non garantisce l'integrazione di tutti i barcodes, ma riduce la probabilità che alcuni di essi vadano completamente persi durante la fase di trasduzione, ovvero la fase in cui avviene l'integrazione dei vettori virali all'interno delle cellule.[7]. Alcuni giorni dopo, il 26 gennaio, viene effettuata la fase di selezione, in cui, tramite puromicina, ossia un antibiotico, vengono mantenute solamente le cellule che hanno integrato il gene reporter, che in questo caso è il gene PuroR, e con esso il barcode. Utilizzando una libreria di

circa $100K$ barcodes unici, i ricercatori si aspettano che una frazione significativa di essi venga effettivamente recuperata nella popolazione selezionata e quindi di ottenere $100K$ barcodes unici su $100K$ cellule. Successivamente, dopo 26 giorni in cui le cellule vengono lasciate libere di crescere, in data 21 febbraio, la popolazione cellulare raggiunge le $38M$ cellule e da esse vengono campionate $6M$ di cellule per costituire gli insiemi costituenti i replicati tecnici: Bulk-1, Bulk-2 e Bulk-3 di dimensioni $2M$ di cellule ciascuno. Nella stessa data, vengono costituiti ulteriori tre gruppi di cellule, impiegati per il controllo dell'esperimento:

- Linea L di dimensione $2M$ di cellule;
- Linea M di dimensione $200K$ di cellule;
- Linea S di dimensione $20K$ di cellule.

A seguire, la linea L subisce una suddivisione a metà due volte a settimana per un totale di 42 giorni, fino al 4 aprile, mentre le linee M e S vengono coltivate e lasciate crescere fino al raggiungimento delle $2M$ di cellule e poi anch'esse suddivise a metà due volte a settimana fino al 4 aprile. Durante il periodo che intercorre tra una suddivisione e l'altra, le cellule sono lasciate libere di crescere e duplicarsi.

Le cellule rimanenti vengono impiegate per svolgere esperimenti in vivo sui topi, suddividendole in 12 gruppi da $2M$ di cellule ciascuno, al fine di ottenere indicazioni utili sui meccanismi alla base della possibile acquisizione di resistenza terapeutica.

Uno schema riassuntivo dell'esperimento è riportato nella figura 1.1.

Nel corso dell'esperimento, sono stati raccolti alcuni dati tramite *screening/campionamento*, grazie al quale i ricercatori hanno potuto contare approssimativamente il numero di cellule che condividono lo stesso barcode all'interno di un gruppo. Si può considerare che, data una popolazione di N cellule, tale processo seleziona casualmente una cellula, identifica il suo barcode e ripete lo stesso passaggio k volte. Se il rapporto k/N è sufficientemente grande, si ottiene una buona approssimazione della ripartizione dei barcodes presenti nella popolazione cellulare.

I dati acquisiti sono stati raccolti in forma tabellare, in cui la riga *i-esima* riporta nella prima colonna il codice del barcode *i-esimo* e nelle colonne successive il numero di letture corrispondenti nel gruppo indicato. Inoltre, sono disponibili i risultati degli screening avvenuti in due momenti diversi dell'esperimento:

- Per i Bulk-1, Bulk-2 e Bulk-3, appena dopo il loro campionamento successivo alla loro formazione $\rightarrow$ in data 21 febbraio;
- Per le linee L, M e S, appena dopo il loro campionamento successivo al termine del processo di crescita-divisione $\rightarrow$ in data 4 aprile.

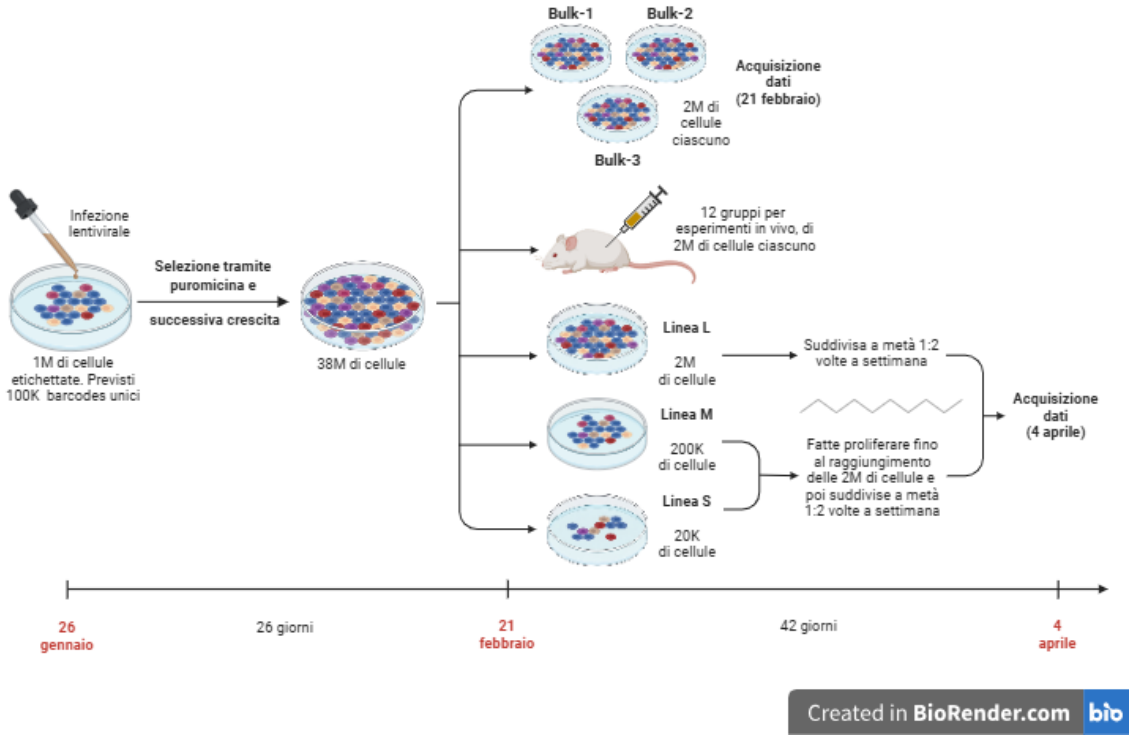


Figura 1.1: Schema temporale dell'esperimento svolto dai ricercatori dell'IRCC. (Created in https://BioRender.com)

Nelle tabelle 1.1 e 1.2 sottostanti sono rappresentati degli estratti delle tabelle per consentirne una visualizzazione più chiara.

Letture barcodes al momento del campionamento dei bulks (21 febbraio)			
Codice barcode	Bulk-1	Bulk-2	Bulk-3
GTGTATA[...]GACTGTG	19548	28333	35595
AGCAGGC[...]GCATGTG	17054	17986	25628
AGGCGCA[...]GTGTGCA	14710	14150	11576
...	...	...	...
TCCTCTT[...]GCAACAC	0	0	0
TCCTCTT[...]ACATGCA	0	0	0
TCCTCTT[...]AACGTGT	0	0	1

Tabella 1.1: Tabella riportante le prime tre e le ultime tre righe dei dati forniti e corrispondenti ai bulks al momento del loro campionamento.

Letture barcodes al momento del campionamento delle linee (4 aprile)			
Codice barcode	Linea L	Linea M	Linea S
GTGTATA[...]GACTGTG	248286	684566	51748
AGCAGGC[...]GCATGTG	10397	52	89
AGGCGCA[...]GTGTGCA	91400	116104	109601
...	...	...	...
TCCTCTT[...]GCAACAC	2	0	0
TCCTCTT[...]ACATGCA	0	1	0
TCCTCTT[...]AACGTGT	0	0	0

Tabella 1.2: Tabella riportante le prime tre e le ultime tre righe dei dati forniti e corrispondenti alle tre linee al momento del loro campionamento.

Dai dati disponibili, è possibile ottenere il numero totale di barcodes differenti osservati nei due campionamenti, ovvero 43171, e il numero di quelli presenti nei diversi bulks e nelle tre linee. I conteggi sono riportati nella tabella 1.3 ed indicati sia nel caso in cui vengano considerati tutti i barcodes aventi almeno una lettura, sia nel caso in cui vengano scartati quelli aventi un numero di letture inferiore a 10, soglia minima di lettura considerata dai ricercatori.

Di seguito, per semplificare la trattazione che verrà affrontata successivamente, è utile definire una notazione per identificare una specifica riga e/o colonna delle tabelle relative

Linea/bulk	Letture totali	Numero di barcodes distinti	
		Numero letture ≥ 1	Numero letture ≥ 10
Bulk-1	12013595	26443	21858
Bulk-2	11355753	29019	21630
Bulk-3	11521257	29133	21740
Linea L	12982279	7639	2660
Linea M	12694147	8062	2331
Linea S	11716979	9977	2589

Tabella 1.3: Tabella riportante il numero di barcodes distinti all'interno dei campioni al momento del loro campionamento, ovvero in data 21 febbraio per i tre bulks e in data 4 aprile per le tre linee.

ai dati raccolti. Si indica con:

$$S_i^C(T)$$

il numero di letture corrispondenti al barcode *i-esimo*, compiute in un bulk o in una linea C, al giorno T. L'apice C può quindi essere pari a B_j con $j = 1, 2, 3$ per indicare rispettivamente Bulk-1, Bulk-2 o Bulk-3; altrimenti sarà pari a L, M o S nel caso in cui venga considerata una linea. Ad esempio,

$$S^{B_1}_{GTGTATA[...]GACTGTG}(26) = 19548$$

indica il numero di letture di cellule etichettate dal barcode GTGTATA[...]GACTGTG, avvenute nel Bulk-1, 26 giorni dopo l'inizio dell'esperimento, ovvero in data 21 febbraio. Scrivendo invece

$$K_l^C(T)$$

si denota il numero di barcodes distinti presenti nel bulk o nella linea C al giorno T e aventi almeno l letture. Perciò,

$$K_{10}^L(68) = 2660$$

denota il numero totale di barcodes distinti presenti nella linea L e aventi almeno 10 letture, 68 giorni dopo l'inizio dell'esperimento, cioè in data 4 aprile. Infine, si può esprimere il numero totale di letture effettuate nel bulk o nella linea C, al giorno T con:

$$S^C(T)$$

per cui

$$S^{B_2}(26) = 11355753$$

indica il numero totale di letture effettuate nel Bulk-2 26 giorni dopo l'inizio dell'esperimento, cioè in data 21 febbraio.

1.2 Metodi

Nelle sezioni successive, ai fini dell'analisi dei dati osservati, verranno utilizzati diversi strumenti statistici, quali:

- *Diagramma di Venn* → permette di organizzare visivamente le relazioni tra gli insiemi in modo efficace, per identificarne somiglianze e differenze. Pertanto, evidenzia sia la frazione condivisa sia i cloni esclusivi di ciascun campione;
- *Grafico di frequenza* → permette di analizzare le distribuzioni delle frequenze clonali e avere una visione globale della struttura della popolazione in ogni campione;
- *Grafico delle frequenze cumulate* → permette di visualizzare come le letture totali si distribuiscono al variare del numero di barcodes considerati. Essa permette di valutare quanto il contributo alla popolazione di un campione sia concentrato nei cloni più abbondanti e quanto, invece, sia distribuito nella sua frazione rara;
- *Scatter plot a scala logaritmica* → permette di rappresentare ogni barcode come un punto, con le sue abbondanze nei due campioni considerati, poste sugli assi x e y, entrambi trasformati in scala logaritmica per visualizzare bene sia i barcodes rari che quelli abbondanti;
- *Heatmap di intersezione dei quartili* → permette di evidenziare quanti cloni, nelle coppie di campioni, mantengono una posizione simile in termini di abbondanza e quanti, invece, cambiano quartile, riflettendo possibili effetti di selezione dovuti al processo di crescita-divisione.

1.3 Analisi della sovrapposizione clonale

Per studiare la sovrapposizione tra i barcodes presenti nelle diverse linee e nei tre bulks, vengono utilizzati i diagrammi di Venn qualitativi. Il diagramma viene definito *qualitativo* quando i cerchi e gli spazi di intersezione non sono di dimensioni proporzionali ai valori al loro interno.

In prima analisi, è possibile osservare i diagrammi di Venn qualitativi, relativi al numero di barcodes distinti, aventi almeno una lettura e presenti nei tre bulks e nelle tre linee al momento del loro campionamento. Essi sono riportati nelle figure 1.2 e 1.3, per poterne formulare un'interpretazione accurata.

Innanzitutto, il grafico non permette di quantificare il numero di cellule presenti o di letture compiute nelle tre linee, ma solo quanti barcodes distinti vi siano. Da esso si può osservare che, nel caso dei tre bulks, i valori corrispondenti al numero totale di barcodes distinti risultano simili tra loro, anche se il Bulk-1 ne presenta un numero minore, e le intersezioni indicano quasi tutte quantità abbastanza vicine. Nel caso delle linee,

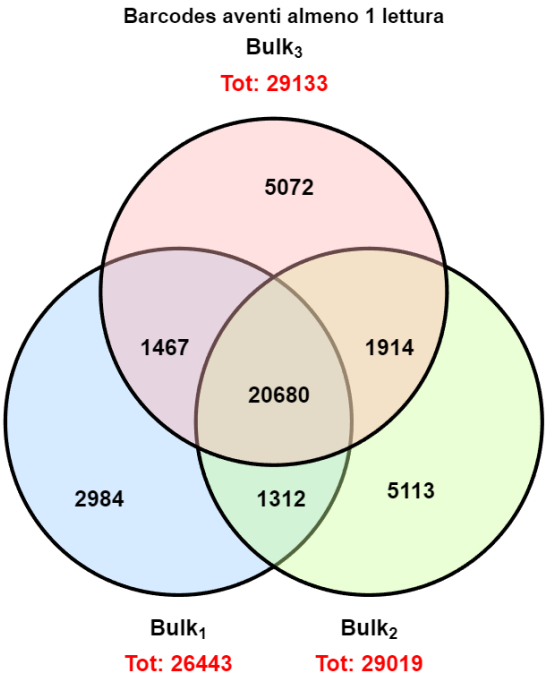


Figura 1.2: Diagramma di Venn qualitativo per i barcodes distinti aventi almeno una lettura, e costituenti i tre bulks al momento della loro formazione.

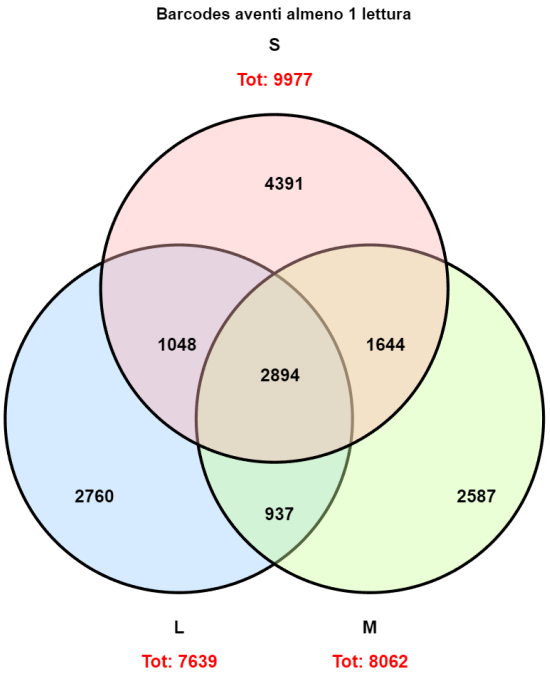


Figura 1.3: Diagramma di Venn qualitativo per i barcodes distinti aventi almeno una lettura, e costituenti le tre linee alla fine del processo di crescita-divisione.

invece, il numero totale di barcodes distinti presenti in S è maggiore e il suo numero di barcodes esclusivi è molto distante dagli altri; infatti, sembra essercene molti di più di quelli contenuti nelle linee L e M, nonostante le dimensioni iniziali inferiori:

$$K_1^S(68) >> K_1^L(68), K_1^M(68).$$

Le linee L e M, invece, hanno valori simili e questo può essere giustificato dalla loro dimensione iniziale maggiore.

Tuttavia, questi dati possono essere influenzati da errori di campionamento. Di conseguenza, è necessario apprendere che, generalmente, la procedura di *screening* prevede che k, ovvero il numero di letture, sia circa $12M$. Pertanto, considerando una popolazione di $N = 2M$ di cellule, come nel caso dei bulks, si ottiene $k/N \geq 6$. Per tale coltura cellulare, si può quindi supporre che ogni cellula venga selezionata circa 6 volte e, di conseguenza, per rimuovere i falsi positivi, non devono essere considerati i barcodes che sono stati letti meno di 5 volte.

Quindi è importante analizzare ulteriori diagrammi di Venn: uno relativo ai bulks e uno alle linee, entrambi definiti per i barcodes distinti aventi almeno 5 letture e riportati nelle figure 1.4 e 1.5.

Confrontando le diverse quantità tra i tre bulks esse risultano più simili rispetto al diagramma precedente e lo stesso avviene per le quantità totali delle tre linee. Inoltre, a

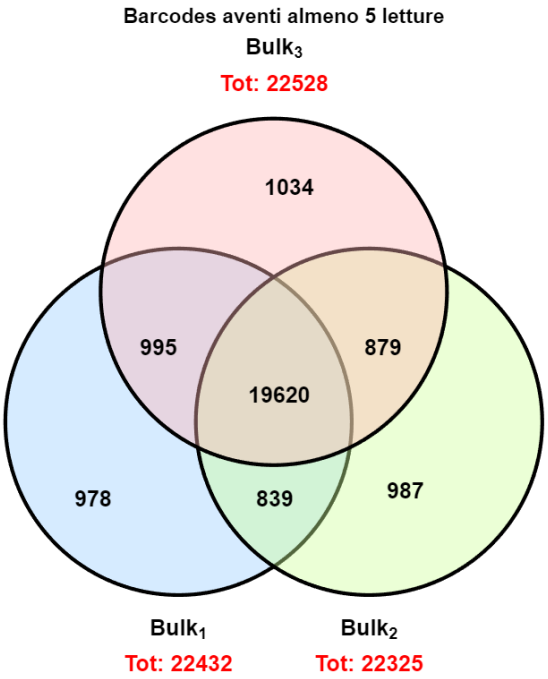


Figura 1.4: Diagramma di Venn qualitativo per i barcodes distinti aventi almeno 5 letture, e costituenti i tre bulks al momento della loro formazione.

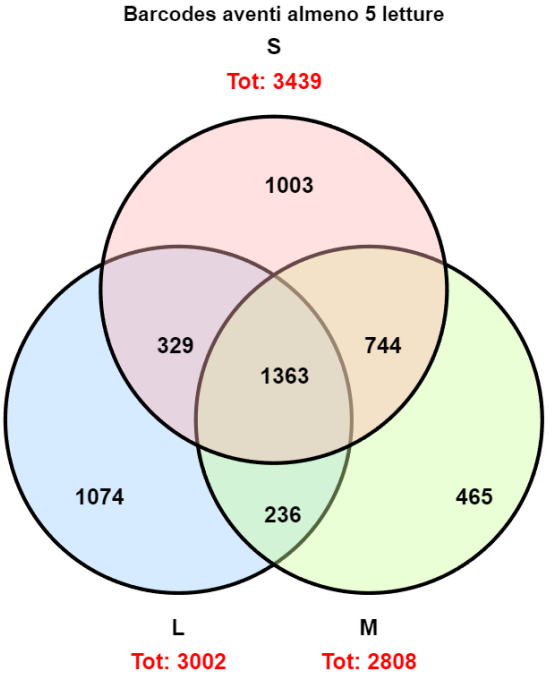


Figura 1.5: Diagramma di Venn qualitativo per i barcodes distinti aventi almeno 5 letture, e costituenti le tre linee alla fine del processo di crescita-divisione.

partire da questa soglia di lettura, risulta di particolare interesse determinare le percentuali corrispondenti sia ai barcodes distinti comuni ai tre campioni che ai barcodes esclusivi nei tre bulks, sul totale di ciascuno di essi. Le percentuali per i bulks sono riportate nella tabella 1.4.

Percentuali sul totale dei barcodes distinti aventi almeno 5 letture			
Bulk	% comuni	% esclusivi	Totale
Bulk-1	87.5%	4.6%	22528
Bulk-2	87.9%	4.4%	22325
Bulk-3	87.1%	4.4%	22432

Tabella 1.4: Tabella riportante le percentuali corrispondenti al numero di barcodes distinti comuni ed esclusivi in ciascun bulks, rispetto al corrispondente totale di barcodes distinti aventi almeno 5 letture.

Le prime appaiono elevate mentre i barcodes esclusivi costituiscono meno del 5% dei campioni. Pertanto, i bulks possono rappresentare una buona approssimazione della distribuzione dei barcodes per la popolazione di $38M$ cellule. Effettuando lo stesso calcolo

per le tre linee si ottengono le percentuali riportate in tabella 1.5.

Percentuali sul totale dei barcodes distinti aventi almeno 5 letture			
Linea	% comuni	% esclusivi	Totale
Linea L	45.4%	35.8%	3002
Linea M	48.5%	16.6%	2808
Linea S	39.6%	29.2%	3439

Tabella 1.5: Tabella riportante le percentuali corrispondenti al numero di barcodes distinti comuni ed esclusivi in ciascuna linea, rispetto al corrispondente totale di barcodes distinti aventi almeno 5 letture.

Da esse si evince che quasi la metà della popolazione di barcodes presente è comune alle tre linee e l'altra metà è in gran parte costituita da barcodes esclusivi. In aggiunta, dal diagramma di Venn con soglia minima di 5 letture, si può osservare che tale soglia di lettura porta le tre linee ad avere un numero simile di barcodes distinti totali, anche se si ha una differenza più accentuata tra le linee S e M. Inoltre, la dissomiglianza più evidente è quella tra i numeri di barcodes distinti presenti solo nella linea M.

I dati forniti suggeriscono la possibilità di analizzare un secondo diagramma di Venn in cui vengono considerati solo i barcodes aventi almeno 10 letture, per svolgere analisi più affidabili, in quanto, barcodes con scarse letture potrebbero essere maggiormente influenzati dalla componente stocastica del sequenziamento. Essi sono riportati nelle figure 1.6 e 1.7.

Per i bulks si può osservare che, il numero totale di barcodes distinti rimane sempre simile tre i tre bulks, e, come ci si aspetta, leggermente minore. Inoltre, le rispettive percentuali calcolate sui barcodes comuni e non sono riportate nella tabella 1.6.

Percentuali sul totale dei barcodes distinti aventi almeno 10 letture			
Bulk	% comuni	% esclusivi	Totale
Bulk-1	86.7%	4.6%	21858
Bulk-2	87.7%	4.4%	21630
Bulk-3	87.2%	4.4%	21740

Tabella 1.6: Tabella riportante le percentuali corrispondenti al numero di barcodes distinti comuni ed esclusivi in ciascun bulks, rispetto al corrispondente totale di barcodes distinti aventi almeno 10 letture.

Esse risultano pressoché invariate, confermando la riproducibilità del sequenziamento

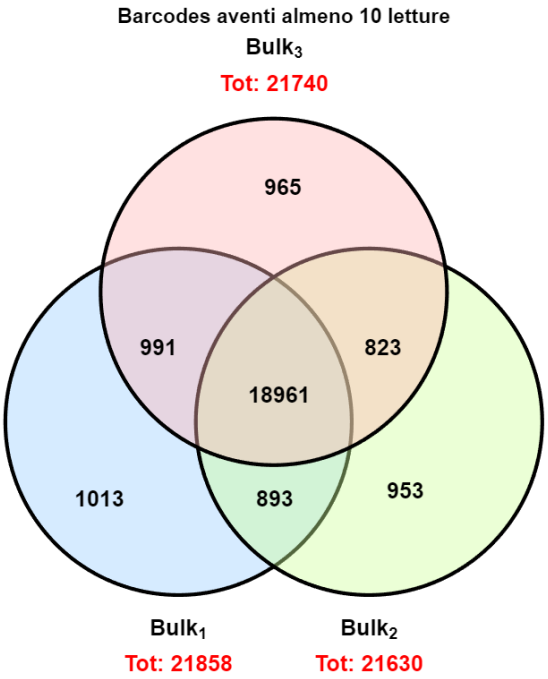


Figura 1.6: Diagramma di Venn qualitativo per i barcodes distinti aventi almeno 10 letture, e costituenti i tre bulks al momento della loro formazione.

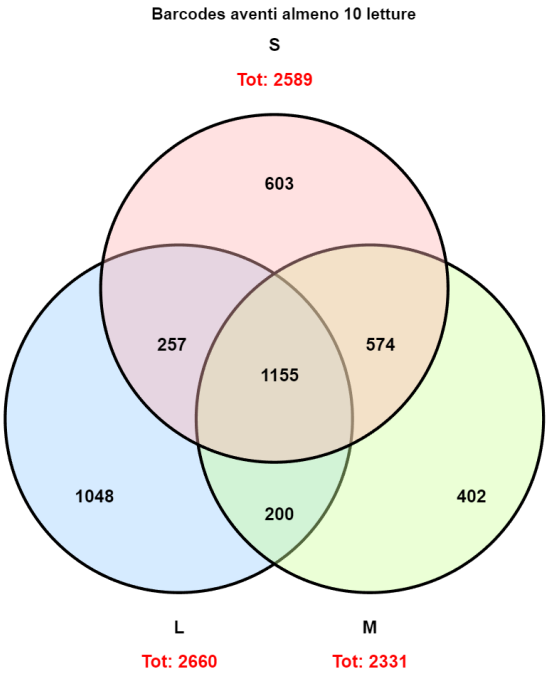


Figura 1.7: Diagramma di Venn qualitativo per i barcodes distinti aventi almeno 10 letture, e costituenti le tre linee alla fine del processo di crescita-divisione.

e suggeriscono che i barcodes sotto soglia rientrano sia nei barcodes esclusivi che in quelli comuni.

Inoltre, si osserva che, anche nel caso delle linee, i valori totali ora sono molto più simili tra loro :

$$K_{10}^{S}(68) \sim K_{10}^{L}(68) \sim K_{10}^{M}(68) \sim 2.5 \cdot 10^{3}$$

e il loro valore di intersezione è stato più che dimezzato rispetto al primo diagramma, avendo diminuito il numero di barcodes esaminabili. In aggiunta, non è più la linea S ad avere un numero maggiore di barcodes distinti e presenti solo in essa, ma è la linea L, e questo è coerente con la maggiore dimensione iniziale di L. Calcolando le rispettive percentuali come indicato prima, si ottengono i valori riportati nella tabella 1.7.

Pertanto, per le linee M e S, le percentuali sui barcodes comuni sono leggermente aumentate, portandoli a ricoprire quasi la metà dei barcodes distinti. Tuttavia, i barcodes esclusivi sono ancora numerosi per tutte e tre le linee.

1.4 Analisi delle frequenze di lettura

Per studiare la frequenza con cui i diversi barcodes sono presenti nei tre bulks e nelle tre linee, è possibile utilizzare grafici di frequenza come i *barplot*, le cui barre hanno un'altezza

Percentuali sul totale dei barcodes distinti aventi almeno 10 letture			
Linea	% comuni	% esclusivi	Totale
Linea L	43.4%	39.4%	2660
Linea M	49.6%	17.2%	2331
Linea S	44.6%	23.3%	2589

Tabella 1.7: Tabella riportante le percentuali corrispondenti al numero di barcodes distinti comuni ed esclusivi in ciascuna linea, rispetto al corrispondente totale di barcodes distinti aventi almeno 10 letture.

proporzionale al numero di letture di ciascun barcode.

In figura 1.8 sono illustrati i grafici ottenuti riportando sull'asse delle ascisse i barcodes utilizzati per l'esperimento, ordinati per frequenza di letture decrescente in ciascun bulk e aventi almeno 10 letture, e sull'asse delle ordinate il corrispondente numero di letture.

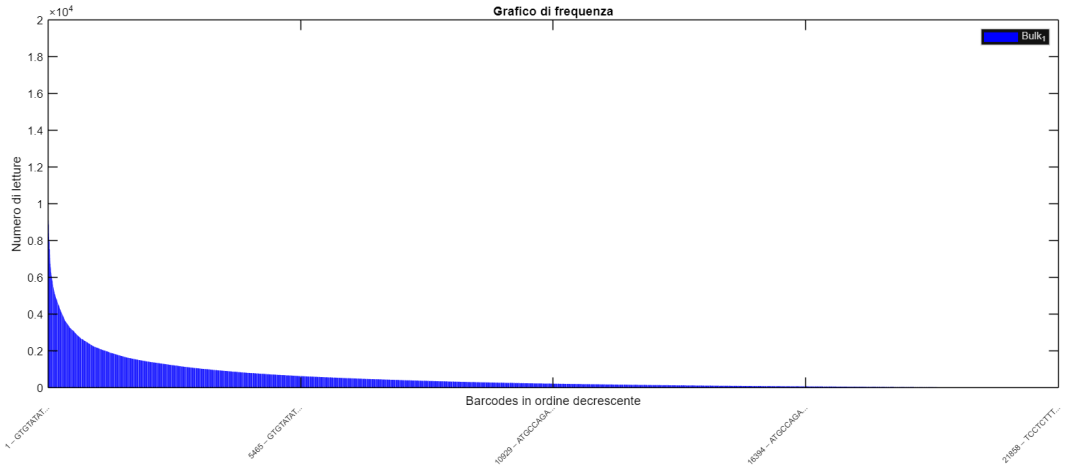
Si può notare che le tre curve di frequenza mostrano forme molto simili, suggerendo che i tre bulks rappresentano repliche coerenti e affidabili della popolazione di partenza. Le differenze in altezza possono riflettere variazioni nella profondità di sequenziamento o di natura biologica.

É possibile generare gli stessi grafici anche nel caso delle tre linee, in questo modo si ottengono gli andamenti riportati nella figura 1.9.

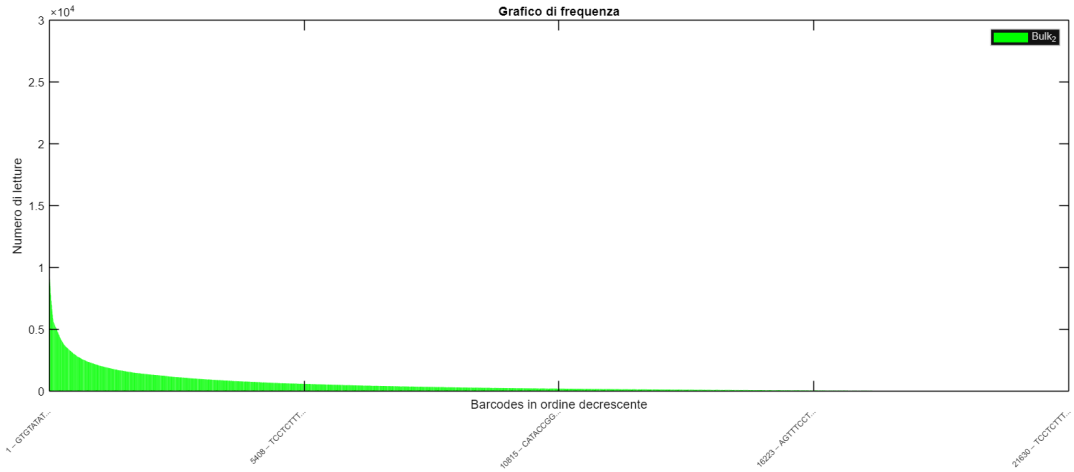
Esaminando i grafici, si osserva che le tre curve di frequenza mostrano andamenti complessivamente simili ma altezze e pendenze differenti, indicando che la struttura generale della distribuzione clonale è comparabile tra i campioni. Tuttavia, la linea M mostra una pendenza iniziale più marcata e, come la linea S, una coda più lunga, suggerendo che i primi barcodes più abbondanti si distacchino maggiormente rispetto a quelli intermedi in termini di letture, e la presenza di una frazione maggiore di barcodes rari. Si osserva anche che nel caso delle linee il numero di letture relativo ai primi barcodes più abbondanti è di un ordine di grandezza superiore se confrontato con i bulks e questo potrebbe essere dovuto nuovamente a effetti biologici o a una diversa profondità di campionamento. Infatti, analizzando le letture corrispondenti ai primi tre barcodes di ciascun campione, mostrate in tabella 1.8, si osserva un distacco significativo tra i primi tre barcodes della linea M.

Tuttavia, tramite questi grafici non è possibile affermare se i barcodes maggiormente presenti in un campione lo siano anche negli altri, ovvero non è possibile determinare se gli stessi barcodes occupino le stesse posizioni per abbondanza.

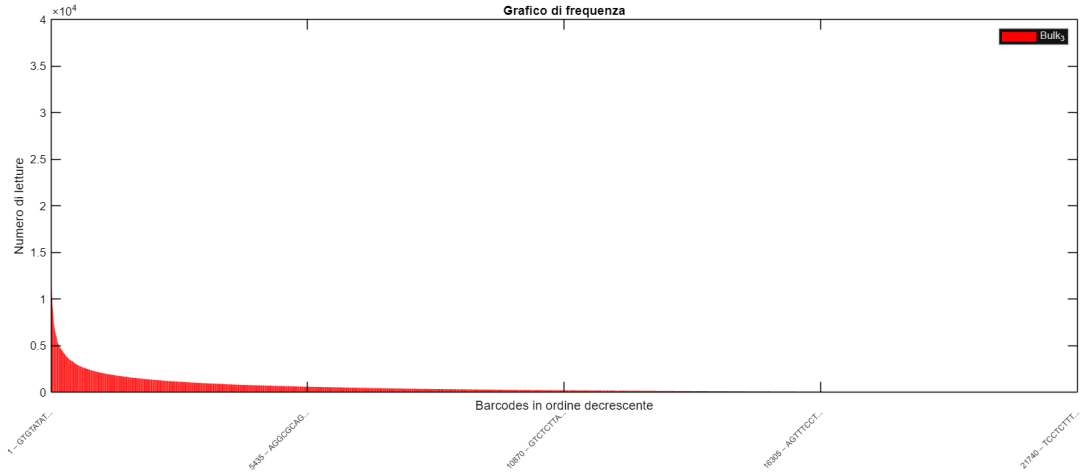
Si può però approfondire questa analisi isolando i barcodes esclusivi di ciascun campione per osservare in che zona della distribuzione essi si trovino. In figura 1.10 sono visibili i grafici di frequenza relativi ai bulks, in cui sono evidenziati i barcodes non comuni.



(a) Bulk-1

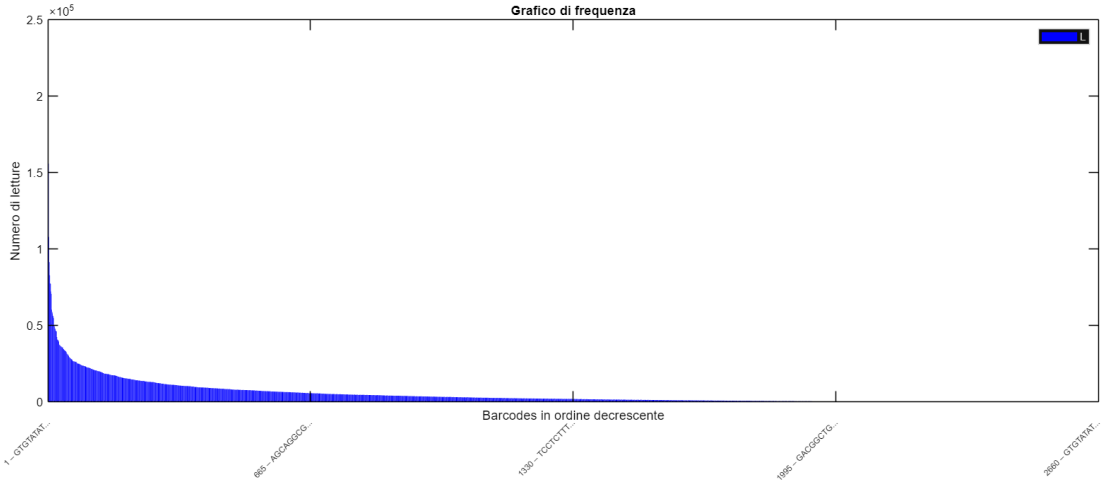


(b) Bulk-2

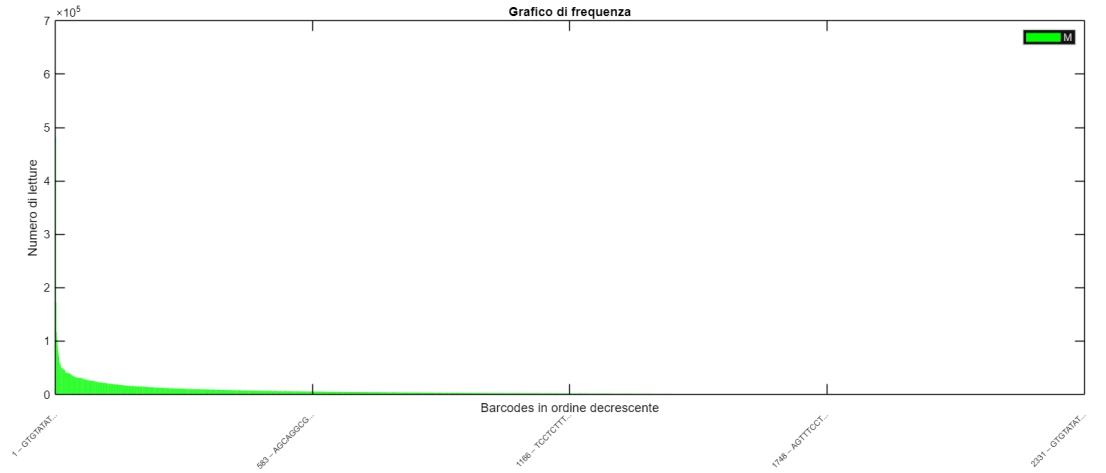


(c) Bulk-3

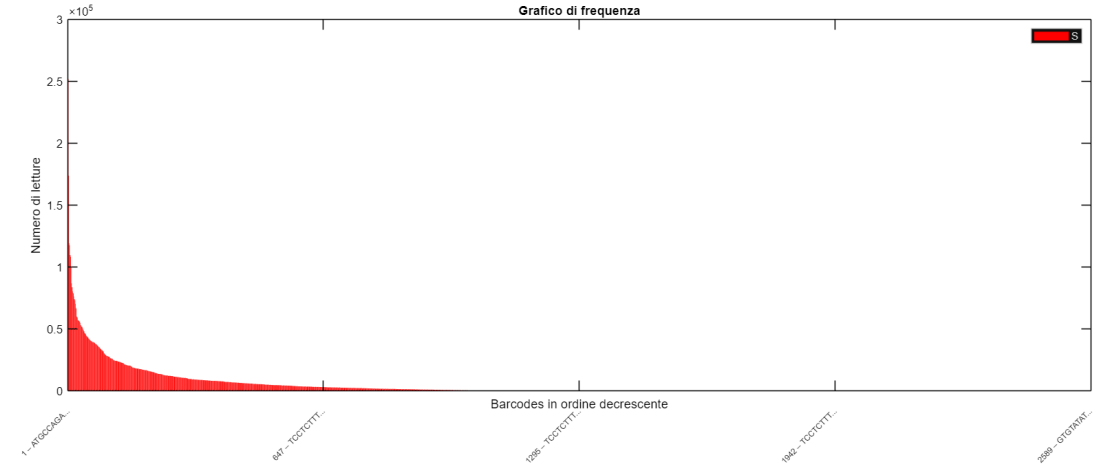
Figura 1.8: Grafici di frequenza per i barcodes ordinati per frequenza di lettura decrescenti e aventi almeno 10 letture in ciascun bulk.



(a) Linea L

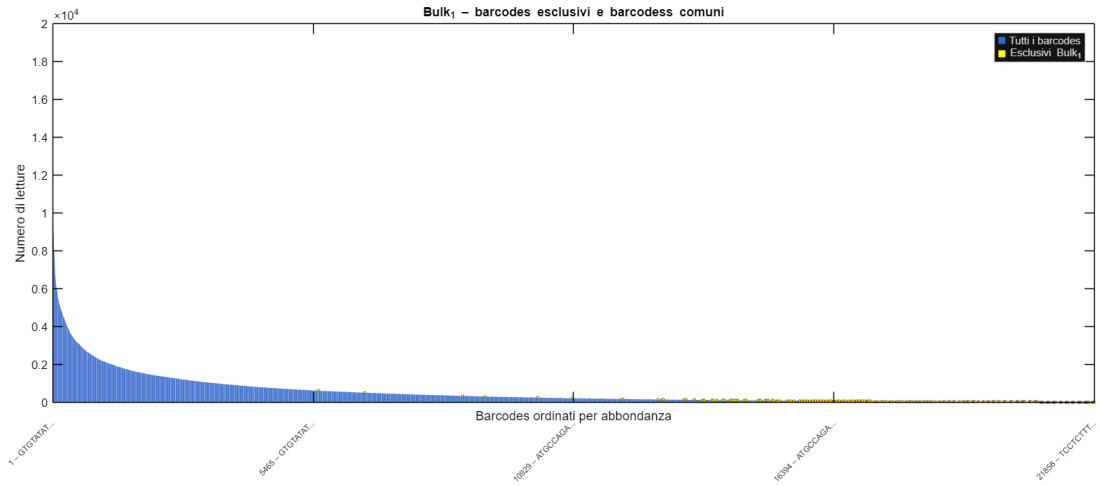


(b) Linea M

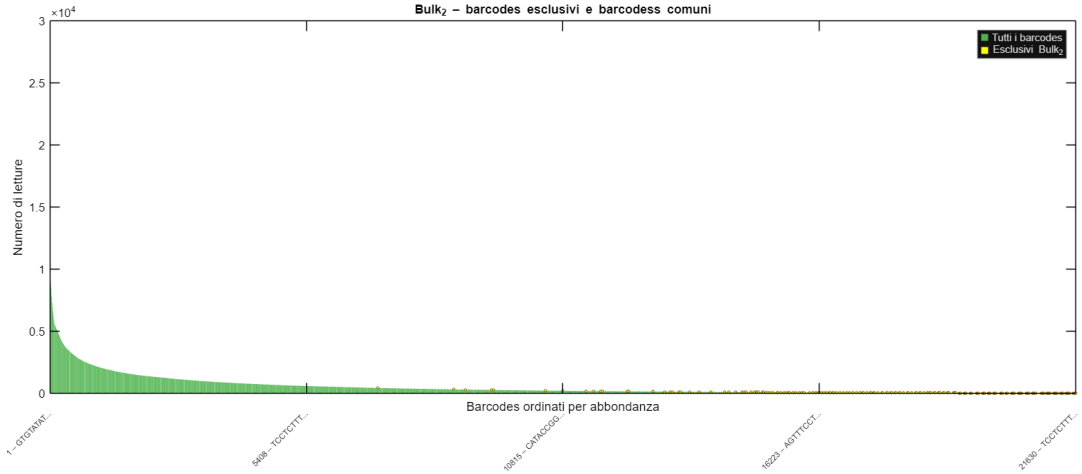


(c) Linea S

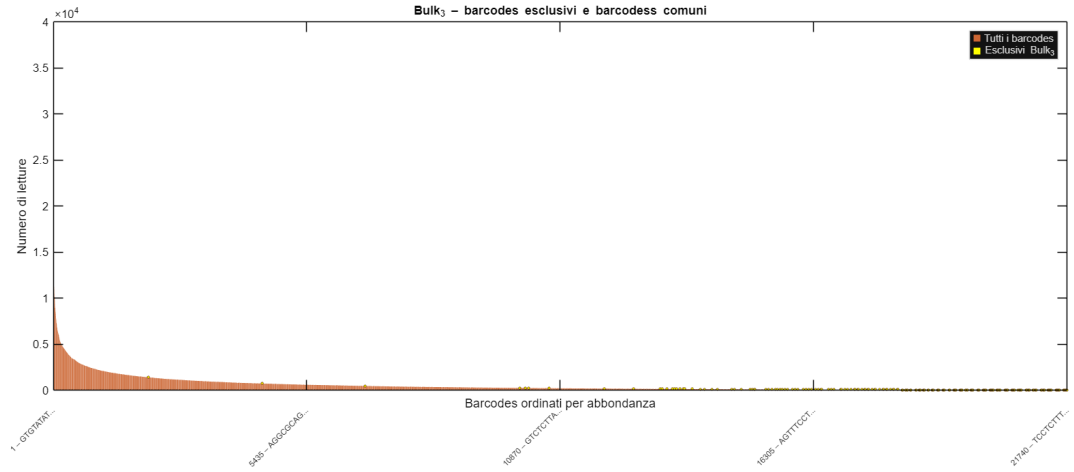
Figura 1.9: Grafici di frequenza per i barcodes ordinati per frequenza di lettura decrescenti e aventi almeno 10 letture in ciascuna linea.



(a) Bulk-1



(b) Bulk-2



(c) Bulk-3

Figura 1.10: Grafici di frequenza per i barcodes ordinati per frequenza di lettura decrescenti e aventi almeno 10 letture in ciascun bulk, e in cui sono evidenziati i barcodes esclusivi di ciascun bulk.

Letture dei primi tre barcodes più abbondanti			
Linea/Bulk	Primo	Secondo	Terzo
Bulk-1	$2 \cdot 10^4$	$1.7 \cdot 10^4$	$1.5 \cdot 10^4$
Bulk-2	$2.8 \cdot 10^4$	$1.8 \cdot 10^4$	$1.8 \cdot 10^4$
Bulk-3	$3.6 \cdot 10^4$	$2.6 \cdot 10^4$	$2.5 \cdot 10^4$
Linea L	$2.5 \cdot 10^5$	$1.6 \cdot 10^5$	$1.1 \cdot 10^5$
Linea M	$6.8 \cdot 10^5$	$4.8 \cdot 10^5$	$1.7 \cdot 10^5$
Linea S	$2.8 \cdot 10^5$	$2.5 \cdot 10^5$	$1.7 \cdot 10^5$

Tabella 1.8: Tabella riportante il numero di letture effettuato per i primi tre barcodes più abbondanti in ciascun campione.

Si osserva che essi sono concentrati nella coda delle tre distribuzioni, pertanto, corrispondono ai barcodes meno abbondanti. Infatti, calcolando la percentuale di letture che tali barcodes, quelli comuni a tutti e tre e quelli comuni a due campioni ricoprono sul totale, si ottengono le quantità in tabella 1.9.

Percentuale delle letture				
Bulk	%comuni a tre campioni	%comuni a due campioni	% esclusivi	Totali
Bulk-1	99%	0.7%	0.3%	12004490
Bulk-2	99%	0.7%	0.3%	11342217
Bulk-3	98.8%	0.9%	0.3%	11507035

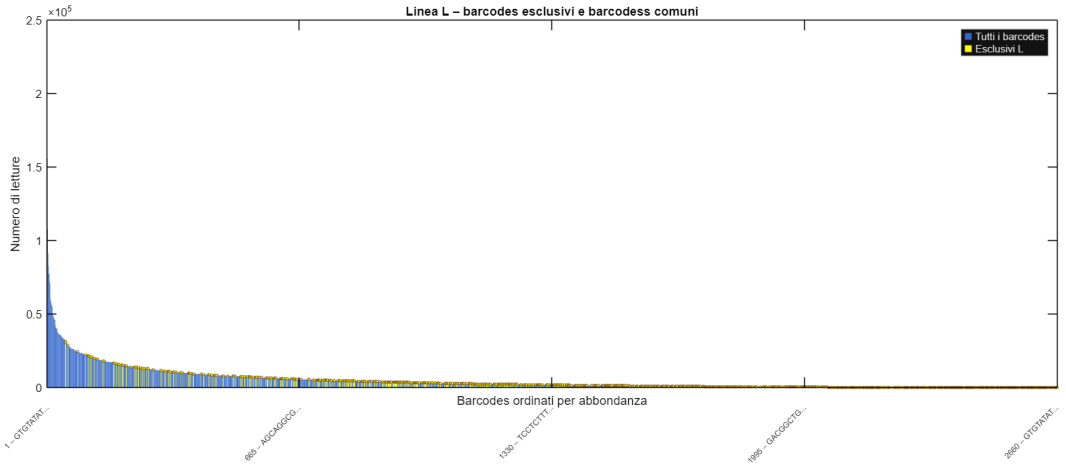
Tabella 1.9: Tabella riportante la percentuale di letture ricoperte dai barcodes comuni a tutti e tre i campioni, comuni a due campioni e dai barcodes esclusivi di ciascuno; ottenute considerando 10 come soglia minima di lettura.

Pertanto, le percentuali per i barcodes esclusivi e i barcodes comuni a soli due campioni risultano vicine allo zero, indicando che il loro contributo nelle popolazioni dei tre bulks è quasi trascurabile. Infatti, risulta che le letture totali siano quasi completamente dovute ai barcodes comuni a tutti e tre i campioni.

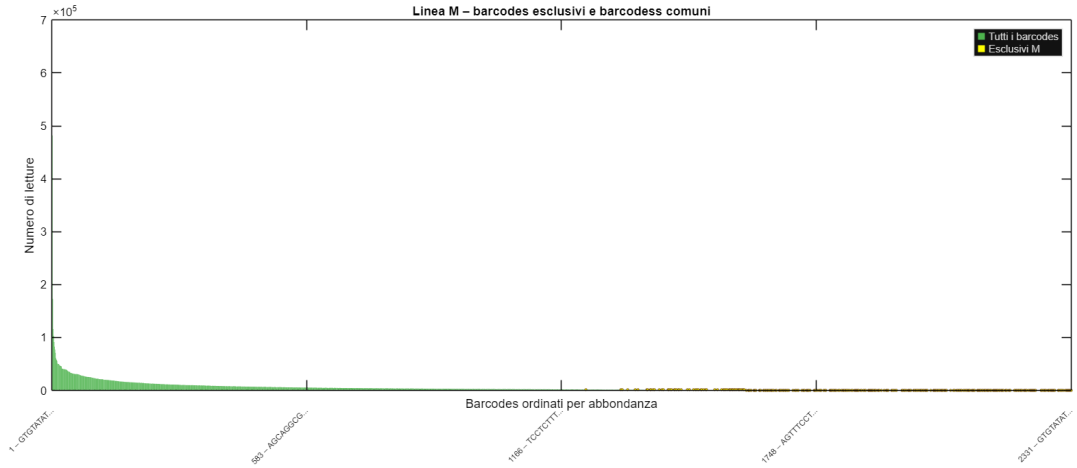
Svolgendo le stesse analisi per le tre linee, si ottengono i grafici di frequenza illustrati nella figura 1.11.

Il comportamento osservato nei bulks viene mantenuto solo dalla linea M, la quale presenta barcodes esclusivi solo tra i barcodes più rari. Le linee L e S, invece, presentano numerosi barcodes esclusivi, anche di media e grande abbondanza. Le percentuali calcolate sulle letture totali, infatti, sono pari alle quantità riportate nella tabella 1.10.

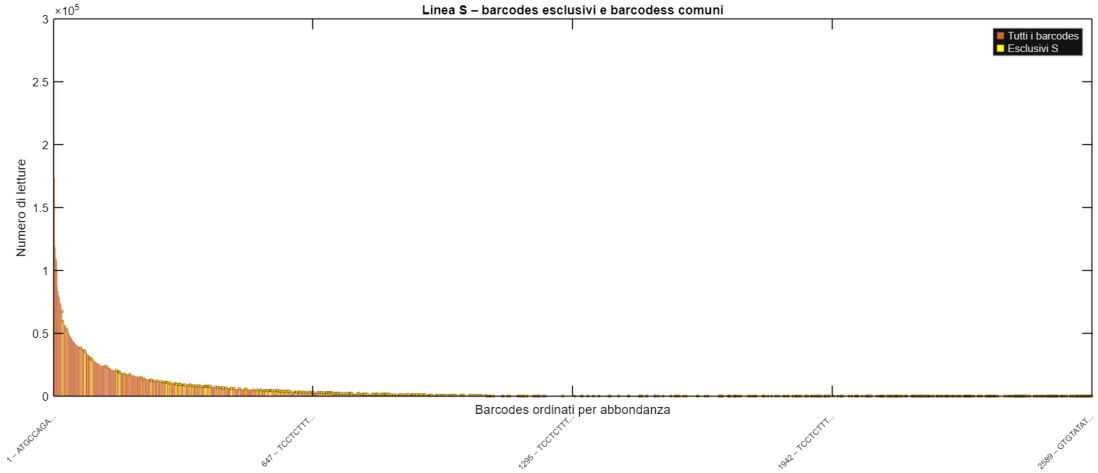
Da esse si deduce che le linee L e S presentano una maggiore diversità di barcodes, mostrando percentuali elevate per i barcodes esclusivi, e quindi non trascurabili.



(a) Linea L



(b) Linea M



(c) Linea S

Figura 1.11: Grafici di frequenza per i barcodes ordinati per frequenza di lettura decrescenti e aventi almeno 10 letture in ciascuna linea, e in cui sono evidenziati i barcodes esclusivi di ciascuna linea.

Percentuale delle letture				
Linea	%comuni a tre campioni	%comuni a due campioni	% esclusivi	Totali
Linea L	70.9%	11.3%	17.8%	12973268
Linea M	82.1%	17.0%	0.9%	12682910
Linea S	66.4%	23.4%	11.2%	11700846

Tabella 1.10: Tabella riportante la percentuale di letture ricoperte dai barcodes comuni a tutti e tre i campioni, a due campioni e dai barcodes esclusivi di ciascuno; ottenute considerando 10 come soglia minima di lettura.

1.5 Analisi delle frequenze cumulate in funzione del numero di barcodes considerati

Il grafico delle frequenze cumulative rappresenta la somma progressiva delle letture ordinate in senso decrescente, mostrando come la distribuzione dei valori si accumuli al crescere del numero di barcodes considerati. Questo tipo di analisi è utile per confrontare la diversità clonale tra i bulks e le linee al momento del loro campionamento.

Questi grafici si ottengono riportando sull'asse delle x il numero di barcodes considerati, in ordine di frequenza decrescente, e sull'asse delle y le corrispondenti frequenze cumulative delle letture di tali barcodes. Le curve così ottenute sono visibili nella figura 1.12.

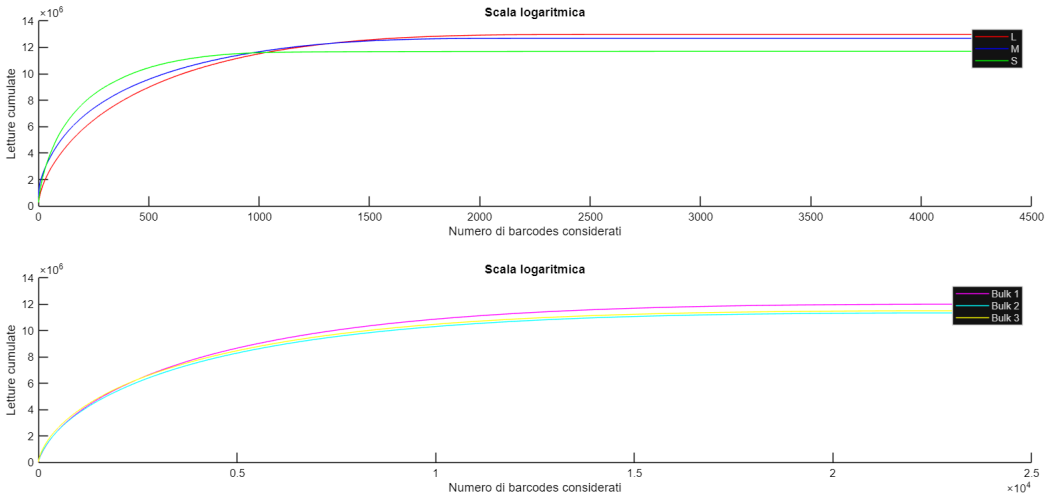


Figura 1.12: Grafici delle frequenze cumulate in funzione dei barcodes considerati aventi almeno 10 letture in una linea o in un bulk.

In questa analisi, sono stati considerati solamente i barcodes aventi almeno 10 letture in uno dei tre campioni confrontati, prima di calcolarne la frequenza cumulata, al fine di valutare il contributo dei cloni più abbondanti in ciascuna popolazione.

Si può osservare che le curve relative alle linee L, M e S presentano un aumento iniziale più ripido rispetto ai bulks, con la linea S che mostra una ripidità maggiore e la linea M intermedia. Questo è coerente con le loro dimensioni iniziali differenti ed è riconducibile al fatto che i primi barcodes più dominanti hanno frequenze molto maggiori dei bracodes intermedi.

Analizzando invece il grafico relativo ai bulks, si nota una crescita delle curve più graduale, e questo è coerente con la constatazione evidenziata dai grafici di frequenza, secondo cui la distribuzione delle dominanze è più uniforme. Inoltre, le tre curve sembrano quasi sovrapposte, e questo conferma che i bulks possono essere considerati come rappresentativi della popolazione in data 21 febbraio. Infatti, calcolando il numero minimo di barcodes necessari per ricoprire il 70% delle letture totali di ciascun campione, si ottengono:

- Bulk-1 $\rightarrow$ 4633;
- Bulk-2 $\rightarrow$ 4485;
- Bulk-3 $\rightarrow$ 4398;
- Linea L $\rightarrow$ 514;
- Linea M $\rightarrow$ 403;
- Linea S $\rightarrow$ 233;

Pertanto, calcolando la percentuale sul numero totale di barcodes distinti presenti in ogni campione, si ottiene:

- Bulk-1 $\rightarrow$ 21.3%;
- Bulk-2 $\rightarrow$ 20.7%;
- Bulk-3 $\rightarrow$ 20.1%;
- Linea L $\rightarrow$ 19.3%;
- Linea M $\rightarrow$ 17.3%;
- Linea S $\rightarrow$ 9%;

Queste percentuali quantificano quanto sia concentrata la distribuzione delle frequenze in ciascun campione e confermano la maggiore ripidità delle linee. Esse presentano una distribuzione più sbilanciata e una forte concentrazione delle frequenze in pochi cloni dominanti, contrariamente ai tre bulks. Tuttavia, la percentuale per la linea L si avvicina maggiormente a quella dei bulk, e quella della linea M è intermedia tra L e S. Questi valori

sono coerenti con le dimensioni iniziali delle tre linee.

Inoltre, ognuna di esse non raggiunge la stessa altezza per la regione di plateau; ciò è legato al numero totale di letture, che può essere diverso per ogni bulk e per ogni linea, probabilmente a causa delle diverse profondità di campionamento, come precedentemente evidenziato tramite i grafici di frequenza.

Applicando la scala logaritmica a entrambi gli assi, si ottengono i grafici riportati nella figura 1.13.

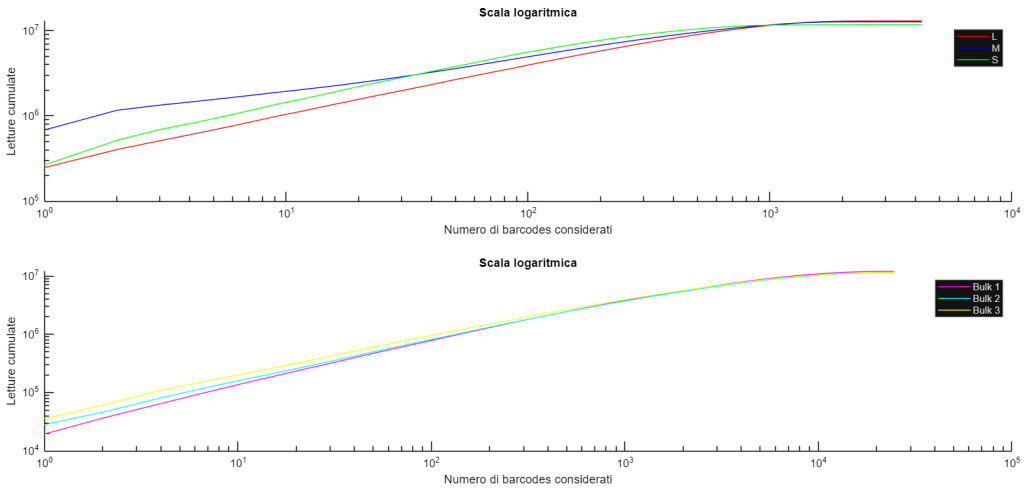


Figura 1.13: Grafici delle frequenze cumulate in funzione dei barcodes considerati, in scala logaritmica.

L'obiettivo dell'utilizzo della scala logaritmica è determinare se esiste una relazione del tipo:

$$frequenze\ cumulate(numero\ di\ barcodes\ considerati) = (numero\ di\ barcodes\ considerati)^{-\beta}$$

ovvero una *legge di potenza*, con $\beta > 0$ parametro costante detto *esponente* o *parametro di scalatura* [9].

Per ogni bulk e per ogni linea si osserva un andamento lineare a tratti, con pendenze differenti. In aggiunta, ciascuno di essi ha mantenuto anche un tratto di plateau. Tale regione indica la saturazione delle letture in quanto i dati in x e in y sono finiti. Pertanto, una legge di potenza non è applicabile globalmente, ma solo localmente, come nella regione iniziale. Inoltre, la linea M presenta un tratto iniziale più ripido che forma con il tratto successivo una zona a gomito; ciò suggerisce che in essa i barcodes maggiormente abbondanti presentano molte più letture dei barcodes intermedi. Infatti, come analizzato nella sezione relativa ai grafici di frequenza, i primi tre barcodes più abbondanti presen-

tano ampi salti di frequenza.

Il grafico in scala logaritmica suggerisce di limitare l'applicazione di una legge di potenza ai primi 10^3 barcodes per ogni campione.

1.6 Legge di potenza

Come suggerito nella sezione precedente, relativa alle curve cumulative, in cui viene illustrata in scala logaritmica la relazione tra le letture cumulate e il numero di barcodes considerati (da qui identificati come: $(R, Y(R))$), è possibile applicare una legge di potenza ai primi 10^3 barcodes, del tipo:

$$Y(R) = C'R^{-\beta}$$

Il grafico citato, però, ha solo funzione diagnostica volta a valutare l'esistenza di un tratto rettilineo a cui applicare la legge di potenza. In letteratura infatti, il grafico generalmente utilizzato per poter determinare i parametri della legge è definito come *Zipf plot* o *rank-abundance plot* (r, y_r) e riporta sull'asse delle y le letture corrispondenti al barcode r-esimo, la cui posizione in ordine decrescente di abbondanza è riportata sull'asse x e denotata con la lettera r [10]. Sapendo che le variabili considerate sono discrete e intere, è possibile scrivere la legge di potenza come:

$$y_r = Cr^{-\alpha}$$

dove C è una costante e $\alpha > 0$ è un esponente positivo che fornisce una misura della dominanza clonale nelle tre linee e nei tre bulks ed è legato a β dalla relazione $\beta = 1 - \alpha$.

Per determinare i valori di tali parametri, si applica il logaritmo a entrambi i lati:

$$\log y_r = -\alpha \log r + b, \quad b = \log C$$

e pertanto è necessario utilizzare la funzione nativa *polyfit* del software *Matlab*. Essa permette di determinare i coefficienti di un polinomio di grado n che meglio si adatta ai dati, utilizzando il metodo dei minimi quadrati [10]. Nel caso in esame, il polinomio ricercato è di grado 1 e i valori ottenuti sono riportati nella tabella 1.11.

Parametri α e C delle leggi di potenza			
α_{Bulk_1}	0.3637	C_{Bulk_1}	$3.0182 \cdot 10^4$
α_{Bulk_2}	0.3867	C_{Bulk_2}	$3.3525 \cdot 10^4$
α_{Bulk_3}	0.4300	C_{Bulk_3}	$4.4552 \cdot 10^4$
α_L	0.6883	C_L	$4.9962 \cdot 10^5$
α_M	0.7796	C_M	$7.5276 \cdot 10^5$
α_S	1.1641	C_S	$5.1472 \cdot 10^6$

Tabella 1.11: Tabella riportante i valori dei parametri α e C ottenuti tramite la funzione *polyfit* di *Matlab*. I risultati sono ordinati per valori crescenti di α e di C.

Utilizzando questi valori dei parametri, si ottengono le rette riportate nei grafici in figura 1.14, dalle quali si può notare che esse approssimano bene i dati, ad eccezione della linea S.

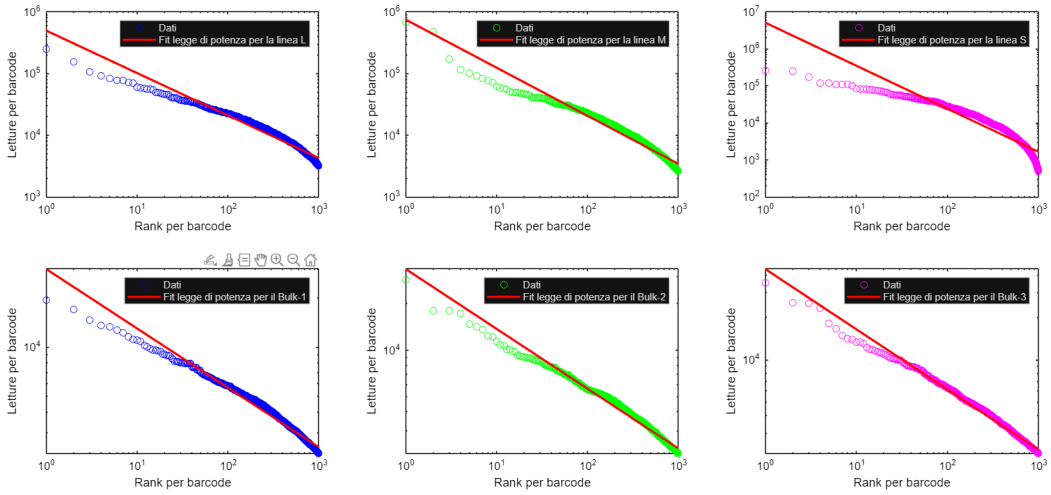


Figura 1.14: Fit dei dati tramite legge di potenza, corrispondenti ai primi 10^3 barcodes più abbondanti nelle tre linee e nei tre bulks.

Valori minori del parametro α indicano che la curva decresce lentamente e, quindi, si ha una distribuzione di barcodes più uniforme; al contrario, valori più elevati indicano che la curva decresce rapidamente; pertanto, si ha la dominanza di pochi barcodes più abbondanti. Il valore assunto dal parametro C identifica il valore di y_r quando $r = 1$, ovvero il numero di letture del barcode più abbondante. I primi barcodes, infatti, corrispondono a quelli più presenti, mentre gli ultimi sono i più rari.

Analizzando i risultati ottenuti, si può osservare che:

$$\alpha_{B_1,B_2,B_3} < \alpha_{L,M,S}$$

$$C_{B_1,B_2,B_3} < C_{L,M,S} \implies S^{B_1,B_2,B_3}_{Barcode_1}(26) < S^{L,M,S}_{Barcode_1}(68)$$

Coerentemente con ciò che si è osservato tramite i grafici di frequenza. Inoltre, i risultati sono coerenti con l'assunzione secondo cui i bulks, presentando una distribuzione più omogenea delle frequenze, vengono considerati rappresentativi della distribuzione dei barcodes presenti nella popolazione di $38M$ di cellule.

Il parametro α relativo alla linea S, però, risulta eccessivamente elevato e indicherebbe frequenze estreme per i primi barcodes; infatti, il parametro C presenta un ordine di grandezza in più rispetto al valore corrispondente nelle altre linee. In aggiunta, analizzando il

grafico relativo, emerge che il valore trovato è significativamente sovrastimato.

Tuttavia, osservando le curve approssimate dalle rette trovate e relative alle tre linee, si nota che, considerando i primi 10^3 barcodes, le linee stimate comprendono anche un marcato principio di flessione e la parte iniziale della distribuzione si distacca più o meno significativamente dalla retta.

Per poter valutare la qualità dei fitting ottenuti, è necessario calcolare due diversi errori tramite l'ausilio della funzione *polyval*, anch'essa nativa di *Matlab*. Avendo applicato la regressione lineare, ovvero una tecnica statistica che modella la relazione tra una variabile dipendente (y_r) e una variabile indipendente (r), ricercando la retta che minimizza la distanza tra i dati, si possono analizzare gli errori definiti da:

- *Coefficiente di determinazione* R^2: è un indice che misura il legame tra la variabilità dei dati e la correttezza del modello statistico utilizzato. Indica quanto bene i dati osservati vengono approssimati dal modello. Tale coefficiente viene definito come:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

dove SS_{res} è la *devianza residua* e SS_{tot} è la *devianza totale*, le quali vengono calcolate nel modo seguente:

$$SS_{res} = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \widehat{y}_i)^2$$

$$SS_{tot} = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

dove e_i sono i *residui*, y_i sono i dati osservati, $\widehat{y}_i$ sono i dati stimati dal modello e $\bar{y}$ è la media degli N dati osservati. Tale stimatore può assumere valori compresi nell'intervallo $(-\infty, 1]$: quantità prossime a 1 mostrano che il modello approssima bene i dati, al contrario, valori vicini a 0 o negativi denotano una cattiva approssimazione;

- *Scarto quadratico medio* $RMSE$: è una misura della deviazione standard dei valori residui, i quali indicano la distanza dei dati osservati dalla retta determinata. Tale stimatore definisce quindi la concentrazione dei dati attorno alla retta di migliore approssimazione e fornisce un'indicazione sulle prestazioni del modello in termini di accuratezza. Esso viene calcolato come:

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(\widehat{y}_i - y_i)^2}{N}}$$

in cui y_i sono gli N dati osservati e $\widehat{y}_i$ sono i dati stimati dal modello. La presenza della radice quadrata garantisce che gli errori negativi e positivi non si annullino tra loro. Valori di $RMSE$ piccoli, e quindi prossimi a 0, indicano una buona descrizione dei dati da parte della retta definita.

I due stimatori dell'errore di fitting devono essere considerati insieme, in quanto misurano due aspetti complementari, ovvero quanto bene la retta descrive i dati e quanto essi se ne discostano [10]. I risultati ottenuti dal loro calcolo, nei casi in esame delle tre linee e dei tre bulks, e considerando i primi 10^3 barcodes più abbondanti, sono riportati nella tabella 1.12.

Stime degli errori di fitting			
$R^2_{B_1}$	0.9745	$RMSE_{B_1}$	0.0251
$R^2_{B_2}$	0.9812	$RMSE_{B_2}$	0.0229
$R^2_{B_3}$	0.9897	$RMSE_{B_3}$	0.0187
R^2_L	0.9495	$RMSE_L$	0.0679
R^2_M	0.9564	$RMSE_M$	0.0712
R^2_S	0.8554	$RMSE_S$	0.2047

Tabella 1.12: Tabella riportante i valori degli stimatori degli errori di fitting R^2 e $RMSE$, ottenuti tramite la funzione *polyval* di *Matlab* e considerando solo i primi 10^3 barcodes più abbondanti, nelle tre linee e nei tre bulks.

Si può osservare che, nel caso dei bulks, si ha $R^2 > 0.97$ e $RMSE < 0.03$, da cui si deduce che essi presentano una distribuzione dei primi 10^3 barcodes più abbondanti, regolare e ben descrivibile da una legge di potenza. Anche le linee L ed M presentano ottimi valori per lo stimatore R^2, tuttavia, il valore per $RMSE$ è leggermente elevato. Analizzando invece i valori ottenuti per la linea S, si denota un R^2 inferiore e un $RMSE$ più che doppio rispetto agli altri. Queste quantità indicano che la curva relativa alla linea S per i primi 10^3 barcodes più abbondanti non può essere ben approssimata da una legge di potenza.

Pertanto, si evince la necessità di svolgere una seconda analisi restringendo l'intervallo su cui definire la legge di potenza. Avendo precedentemente osservato che le curve relative alle frequenze cumulative delle tre linee crescono rapidamente, si può identificare l'intervallo da considerare come costituito dal numero minimo di barcodes necessario per ricoprire circa il 70% delle letture totali. Pertanto, scegliendo intervalli diversi per ciascun campione e di ampiezze pari ai numeri ottenuti nella sezione precedente si ottengono le quantità indicate nella tabella 1.13 per le quali vengono generati i grafici in figura 1.15.

L'utilizzo di questi valori permette di ottenere rette che seguono meglio la distribuzione dei dati nel caso delle linee, mentre le curve stimate per i bulks comprendono ora anch'esse una parte di flessione; pertanto, si discostano significativamente dalla parte iniziale del-

Parametri α e C della nuova legge di potenza.			
α_{B_1}	0.5311	C_{B_1}	$7.9484 \cdot 10^4$
α_{B_2}	0.5454	C_{B_2}	$8.3803 \cdot 10^4$
α_{B_3}	0.5667	C_{B_3}	$9.8414 \cdot 10^4$
α_L	0.5533	C_L	$2.6644 \cdot 10^5$
α_M	0.6116	C_M	$3.4983 \cdot 10^5$
α_S	0.5874	C_S	$3.9788 \cdot 10^5$

Tabella 1.13: Tabella riportante i valori dei parametri α e C ottenuti tramite la funzione *polyfit* di *Matlab*, considerando solo i barcodes più abbondanti necessari a ricoprire il 70% delle letture totali in ciascun campione.

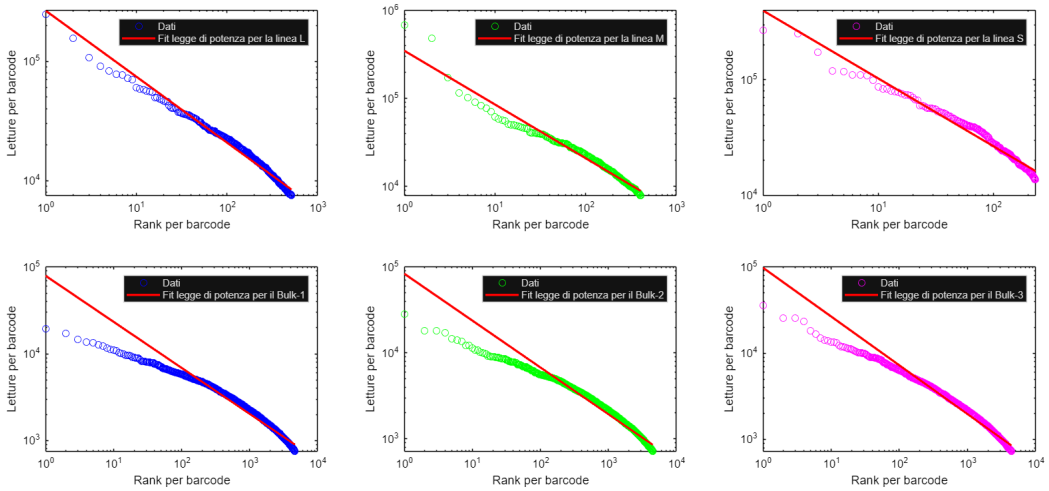


Figura 1.15: Fit dei dati tramite legge di potenza, corrispondenti ai barcodes necessari per ricoprire il 70% delle letture totali in ciascun campione.

la curva. Di conseguenza, i valori dei due parametri si sono ridotti nel caso delle linee, specialmente per la linea S, i cui valori si sono dimezzati, indicando una pendenza meno estrema. Nel caso dei bulks invece, i valori sono drasticamente aumentati. Analizzando il parametro C e confrontandolo con il numero di letture relative al primo barcodes, riportato nella tabella 1.8, si deduce che esso viene significativamente sovrastimato per i bulks; contrariamente, per la linea M esso viene sottostimato, coerentemente con la presenza della zona a gomito nella curva cumulativa. Tuttavia, quest'ultimo, come per le linee L e S, si avvicina ora maggiormente al suo valore reale.

Ricalcolando gli stimatori per le tre linee e i tre bulks, considerando per ciascuno gli intervalli precedentemente indicati, si ottengono i valori riportati nella tabella 1.14.

Si nota che R^2 è ora superiore a 0.97 per tutte e tre le linee, mentre, contrariamente

Stime degli errori di fitting per le nuove leggi di potenza			
$R^2_{B_1}$	0.9577	$RMSE_{B_1}$	0.0483
$R^2_{B_2}$	0.9623	$RMSE_{B_2}$	0.0467
$R^2_{B_3}$	0.9714	$RMSE_{B_3}$	0.0420
R^2_L	0.9816	$RMSE_L$	0.0321
R^2_M	0.9736	$RMSE_M$	0.0425
R^2_S	0.9710	$RMSE_S$	0.0422

Tabella 1.14: Tabella riportante i valori degli stimatori degli errori di fitting R^2 e $RMSE$, ottenuti tramite la funzione *polyval* di *Matlab* e considerando solo i barcodes più abbondanti necessari per ricoprire il 70% delle letture totali in ciascun campione.

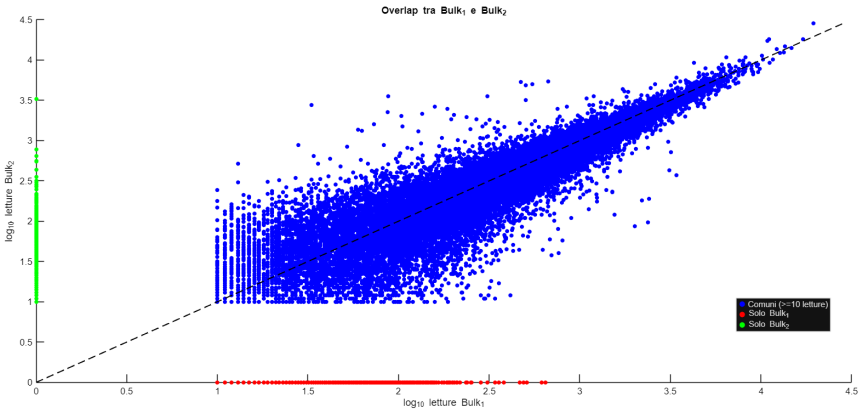
ai valori ottenuti utilizzando l'intervallo 10^3, Bulk-1 e Bulk-2 presentano valori inferiori a tale soglia. In aggiunta, il valore di $RMSE$ risulta diminuito per le tre linee e aumentato per i tre bulks. Da ciò si deduce che, riducendo l'intervallo, si ottengono approssimazioni migliori per le distribuzioni delle tre linee, soprattutto per la linea S, mentre si osserva l'effetto opposto nei bulks. Questo suggerisce che i bulks presentino una zona intermedia in cui i barcodes mostrano variazioni più marcate nella frequenza relativa e una coda che segue maggiormente una legge di potenza.

1.7 Confronto delle abbondanze clonali

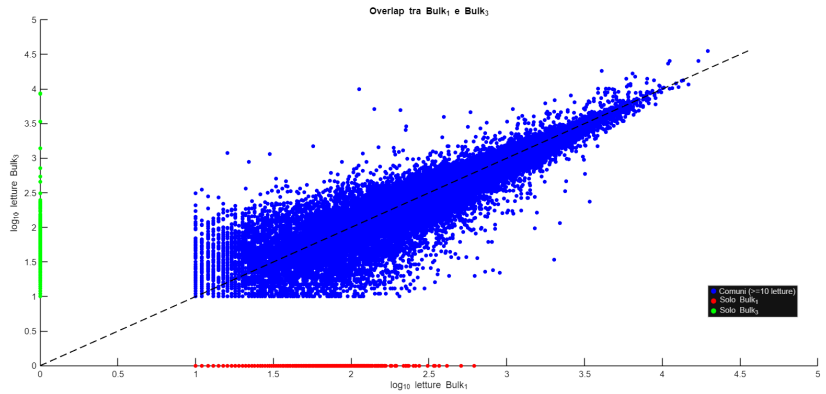
Per analizzare le relazioni tra le abbondanze clonali nei campioni, si utilizza uno scatter plot in scala logaritmica, in cui si considerano i barcodes comuni e non tra due campioni. Questo grafico fornisce una visione globale della correlazione tra le abbondanze nei due campioni, evidenziando la presenza di cloni altamente proliferanti, localizzati in alto a destra, e di cloni che mostrano variazioni marcate. Inoltre, distinguere tra barcodes comuni e non permette di evidenziare se la loro assenza in uno o più campioni possa essere legata a una bassa frequenza.

Nelle figure 1.16 sono riportati gli scatter plot relativi ai tre bulks ottenuti considerando solo i barcodes aventi almeno 10 letture.

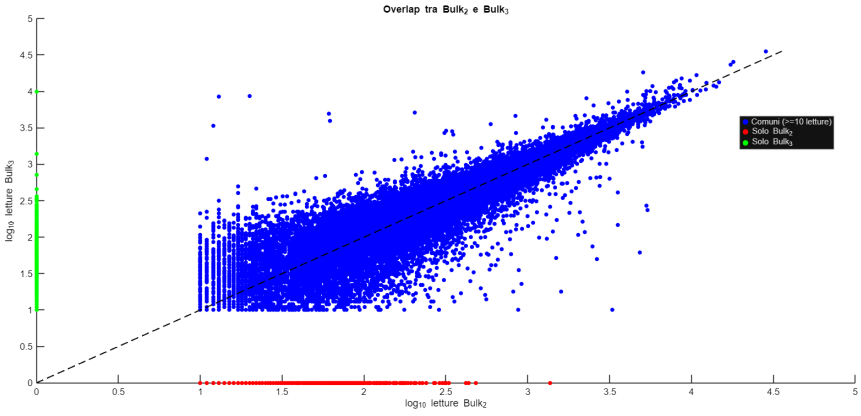
Le distribuzioni dei punti comuni presentano differenze minime nei tre casi e appaiono concentrate attorno alla bisettrice; ciò significa che la maggioranza dei barcodes mantiene le abbondanze relative in ognuno dei tre campioni. Questo suggerisce una buona rappresentazione della diversità clonale dell'intera popolazione. Tuttavia, considerando via via i punti relativi a frequenze sempre minori, si osserva un aumento graduale della presenza di punti distanti dalla bisettrice, suggerendo che la diversità delle abbondanze nei tre campioni aumenta al diminuire della frequenza di ciascun barcode. Inoltre, si può osservare che



(a) Bulk-1 vs Bulk-2



(b) Bulk-1 vs Bulk-3



(c) Bulk-2 vs Bulk-3

Figura 1.16: Scatter plot in scala logaritmica delle letture tra coppie di bulks e con soglia di lettura minima pari a 10.

i punti più a sinistra sembrano essere allineati lungo lo stesso valore dell'asse orizzontale. Questo può essere imputabile alla presenza di numerosi barcodes rari aventi abbondanze che differiscono di poche letture e per cui i valori dei corrispondenti logaritmi sono tra loro approssimabili. In aggiunta, i punti relativi ai barcodes non comuni alle coppie esaminate risultano concentrati nelle frequenze più basse, suggerendo, come indicato nelle analisi precedenti, che essi non siano osservabili in uno o due campioni perché rari.

Svolgendo la stessa analisi anche per le tre linee, si ottengono gli scatter plot illustrati nella figura 1.17.

Le distribuzioni dei punti comuni risultano molto disperse e lontane dalla bisettrice, fornendo un quadro di difficile interpretazione. Questo indica una bassa coerenza tra le abbondanze clonali. Tuttavia, la linea M e la linea S mostrano una fascia di concentrazione lontana dalla bisettrice, corrispondente ai cloni di abbondanza intermedia per la linea M e di bassa abbondanza per la linea S, suggerendo che le uguaglianze tra le due linee siano principalmente sostenute da questi cloni ed evidenziando una relazione monotona tra di esse. Tale comportamento non riguarda l'intero insieme dei barcodes comuni tra le due linee, ma rappresenta una tendenza prevalente.

Inoltre, i barcodes non comuni alle intersezioni presentano frequenze miste.

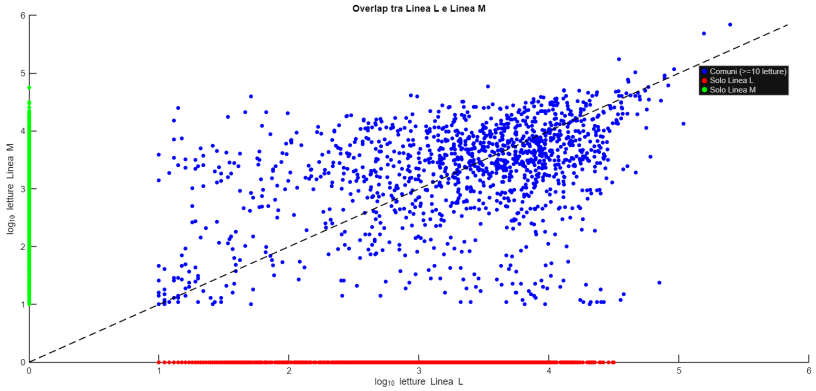
1.8 Analisi della coerenza del ranking clonale

Per analizzare la coerenza del ranking dei barcode tra due campioni e quindi quantificare quanti barcodes mantengono la stessa frequenza nei diversi campioni, valutando in modo aggregato le transizioni tra classi di frequenza nei due campioni, si utilizza una heatmap di intersezione dei quartili, ottenuta dividendo i barcodes di ciascun campione in quattro gruppi, detti quartili (Q), in base alla loro frequenza e corrispondenti a:

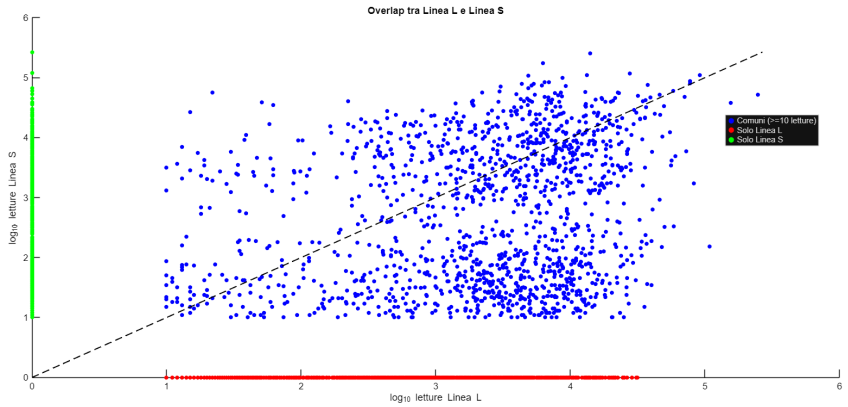
- $Q_1 \rightarrow$ top 25%;
- $Q_2 \rightarrow 25\% - 50\%$;
- $Q_3 \rightarrow 50\% - 75\%$;
- $Q_4 \rightarrow$ bottom 25%.

Successivamente, tramite intersezione si calcola quanti, in termini di percentuale, dei barcodes che ricadono in un determinato quartile sono comuni ai due campioni.

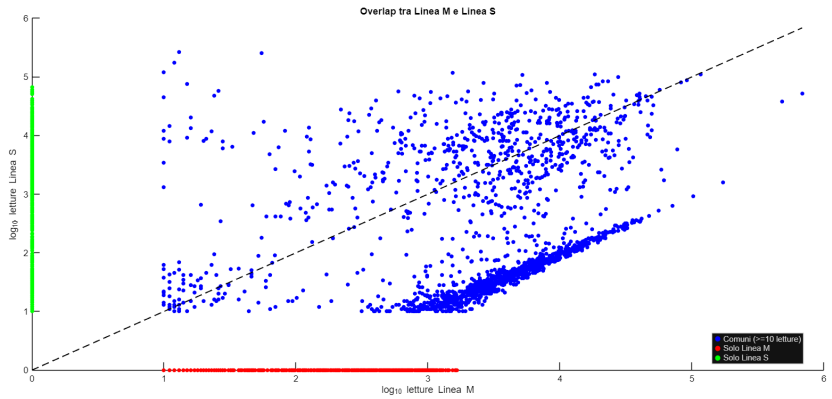
A differenza dello scatter plot, che rappresenta le abbondanze clonali utilizzando i valori continui delle letture, le heatmap quantificano in modo aggregato la proporzione di barcodes che rimane nello stesso quartile o che transita da quartili elevati a quartili inferiori e viceversa, permettendo di descrivere la riorganizzazione delle classi di abbondanza tra i campioni.



(a) Linea L vs Linea M



(b) Linea L vs Linea S



(c) Linea M vs Linea S

Figura 1.17: Scatter plot in scala logaritmica delle letture tra linee e con soglia di lettura minima pari a 10.

Le heatmaps così definite e relative alle intersezioni tra i tre bulks sono riportate nelle figure presenti in 1.18 e sono ottenute considerando solo i barcodes comuni e aventi almeno 10 letture.

Osservando le sfumature dei colori corrispondenti ai diversi quartili, i tre grafici sono sovrapponibili; in particolare, le percentuali relative ai quartili di intersezione tra il Bulk-3 e gli altri bulks differiscono solamente per pochi decimali. Particolarmente rilevanti per l'analisi sono i valori sulla diagonale, i quali rappresentano la percentuale di barcodes che occupano lo stesso quartile in entrambi i campioni. Sommando le percentuali relative, si ottiene:

- Bulk-1 VS Bulk-2 : 69%;
- Bulk-1 VS Bulk-3 : 69.7%;
- Bulk-2 VS Bulk-3 : 70.1%;

Pertanto, i bulks mostrano una buona coerenza del ranking clonale e le basse percentuali negli altri quartili ne confermano la stabilità.

Infine, è necessario definire le stesse heatmaps anche per le intersezioni tra le tre linee; esse sono riportate nella figura 1.19.

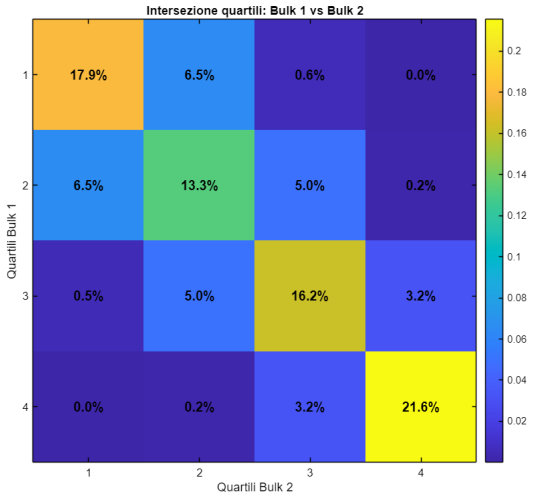
Questi confronti portano a ottenere percentuali elevate al di fuori della diagonale, soprattutto nel caso dell'intersezione tra le linee M e S, in cui si osserva una diagonale inferiore significativa. Questo conferma una relazione monotona tra le due linee, per cui la linea M tende a ridistribuire i barcodes abbondanti della linea S verso quartili inferiori in modo graduale, ovvero un quartile alla volta, sebbene tale comportamento non riguardi l'intero insieme dei barcodes ma il 39.7% dei loro barcodes comuni. Sommando nuovamente le percentuali sulle diagonali, si ottiene:

- Linea L VS Linea M : 36.6%;
- Linea L VS Linea S : 29.5%;
- Linea M VS Linea S : 30.8%;

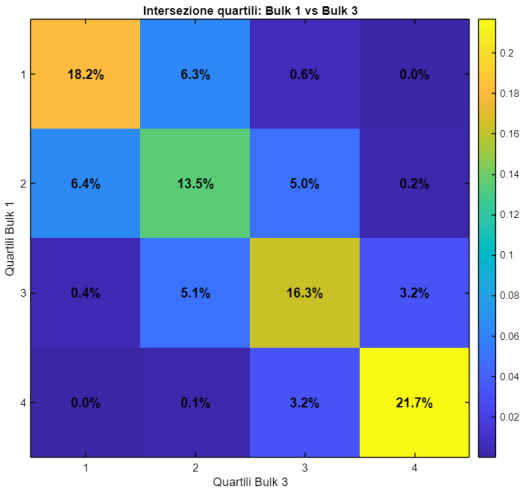
Pertanto, conducono a identificare una forte divergenza clonale tra le linee, in quanto la distribuzione delle abbondanze non è conservata tra i tre campioni. Infatti, si evince che i barcodes più abbondanti in una linea non risultano necessariamente abbondanti nelle altre, e lo stesso vale per i barcodes più rari.

1.9 Possibili interpretazioni dei dati collezionati

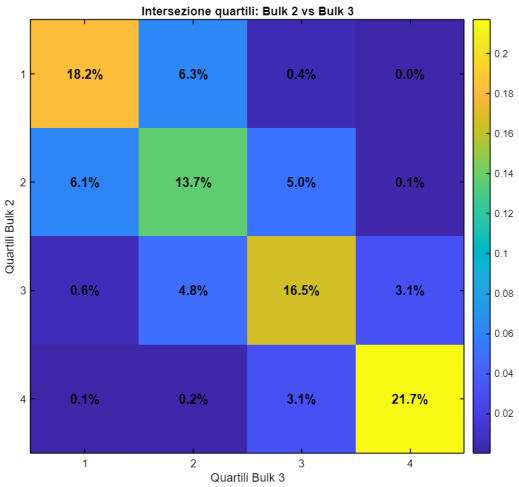
Dopo l'analisi descrittiva dei dati collezionati svolta nei paragrafi precedenti, assumendo che il numero di letture sia proporzionale al numero di cellule, è possibile compiere un'analisi interpretativa.



(a) Bulk-1 vs Bulk-2

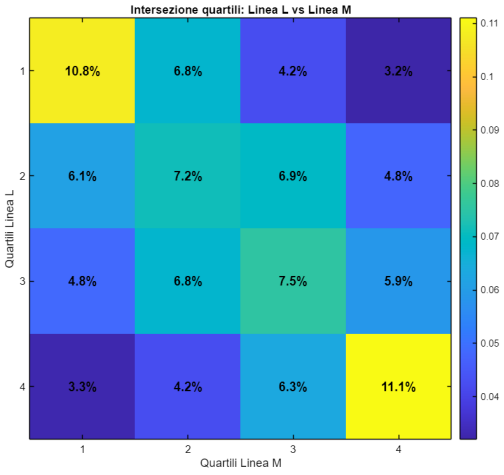


(b) Bulk-1 vs Bulk-3

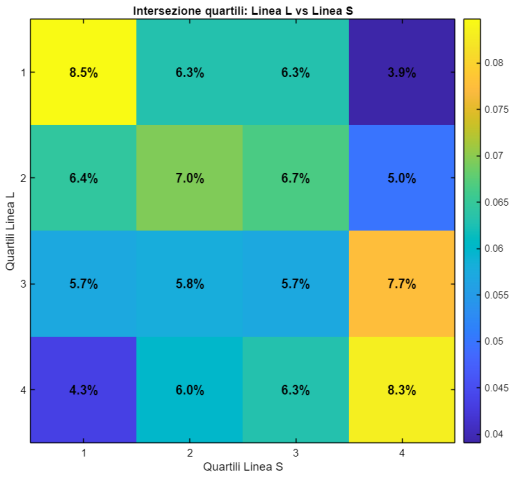


(c) Bulk-2 vs Bulk-3

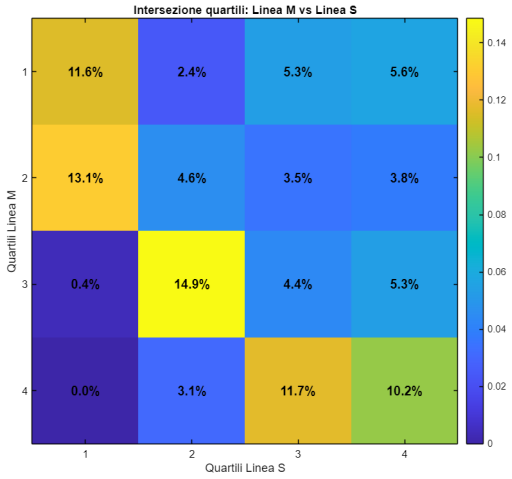
Figura 1.18: Heatmap di intersezione tra i quartili dei tre bulks, in cui le percentuali indicano la distribuzione dei barcode comuni.



(a) Linea L vs Linea M



(b) Linea L vs Linea S



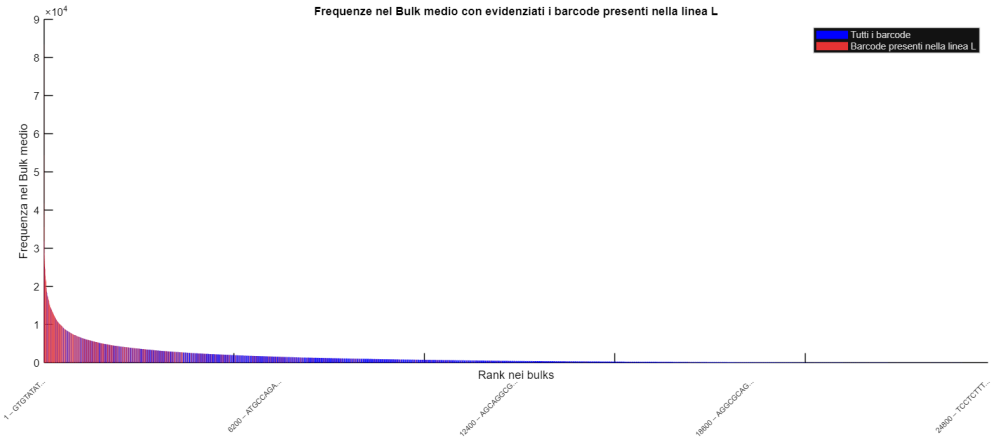
(c) Linea M vs Linea S

Figura 1.19: Heatmap di intersezione tra i quartili delle tre linee, in cui le percentuali indicano la distribuzione dei barcodes comuni.

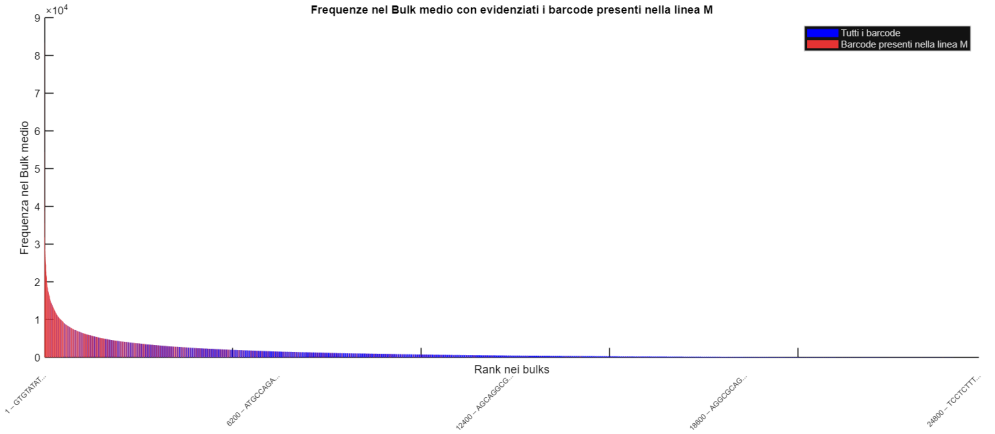
Considerando il processo di crescita-divisione che affrontano le tre linee durante l'esperimento e i diagrammi di Venn nelle figure 1.3 e 1.5, relativi ai barcodes distinti aventi almeno una e 5 letture, si può ipotizzare di giustificare il fatto che S presenta un numero maggiore di barcodes rispetto alle altre linee, supponendo che, durante il processo di crescita iniziale, fino al raggiungimento delle $2M$ di cellule, partendo da dimensioni ridotte, molti cloni, anche quelli inizialmente sotto soglia, crescano notevolmente, mentre vengano persi pochi barcodes rari per effetti stocastici, grazie alla minore competizione. Contrariamente, la linea L non affronta una fase di espansione iniziale e quindi probabilmente mantiene la gerarchia iniziale, portando a una maggiore sopravvivenza dei più presenti e all'estinzione dei meno proliferanti attraverso il maggior numero di suddivisioni che affronta. La linea M, invece, partendo con dimensioni intermedie, potrebbe presentare una popolazione più eterogenea e quindi, al termine del processo di crescita-divisione, potrebbe comprendere meno barcodes che superano la soglia di lettura a causa della maggiore competizione per la sopravvivenza rispetto a quella presente nella linea S. Esaminando invece il diagramma riportato in figura 1.7 e ottenuto escludendo i barcodes aventi meno di 10 letture, si osserva che i barcodes comuni alle tre linee sono in numero minore rispetto a prima, ma sono più significativi in quanto costituiscono i cloni che sono sopravvissuti in abbondanza alla selezione dovuta all'espansione iniziale nelle linee M e S e alle suddivisioni ripetute. In aggiunta, si può presumere che i barcodes comuni a tutte le linee identifichino le cellule con maggiore probabilità di sopravvivenza e quindi con maggiore probabilità di essere campionate. Infine, il fatto che, considerando almeno 10 letture, i valori totali nelle tre linee appaiano simili tra loro può significare che il numero di cloni che riescono a crescere oltre una certa soglia è limitato. Inoltre, analizzando le percentuali dei barcodes comuni e unici sul totale dei barcodes distinti, si osserva che il processo di crescita-divisione ha aumentato la diversità tra i barcodes presenti nelle linee rispetto a quella relativa ai bulks, nonostante le loro dimensioni iniziali. Per di più, la percentuale di barcodes presenti nelle linee ma assenti nei bulks è pari al 28.5% e hanno una media di letture di 2.65 ciascuno. Pertanto, molti potrebbero essere frutto di errori di campionamento, e altri potrebbero essere barcodes presenti nella popolazione totale di $38M$ ma che non sono stati campionati nei bulks.

Per consolidare ulteriormente queste ipotesi, è necessario analizzare quali frequenze di lettura presentino i barcodes che compaiono sia nei tre bulks che nelle tre linee. A tal fine, è necessario definire il *Bulk medio* come somma dei tre bulks e pertanto costituito da $6M$ di cellule, in cui ogni barcodes ha letture pari alla somma delle letture effettuate nei singoli bulks. Si analizzano quindi i grafici di frequenza riportati nella figura 1.20 e ottenuti considerando solo i barcodes aventi almeno 10 letture, in cui è possibile distinguere i barcodes presenti solo nei bulks da quelli presenti anche in ciascuna linea.

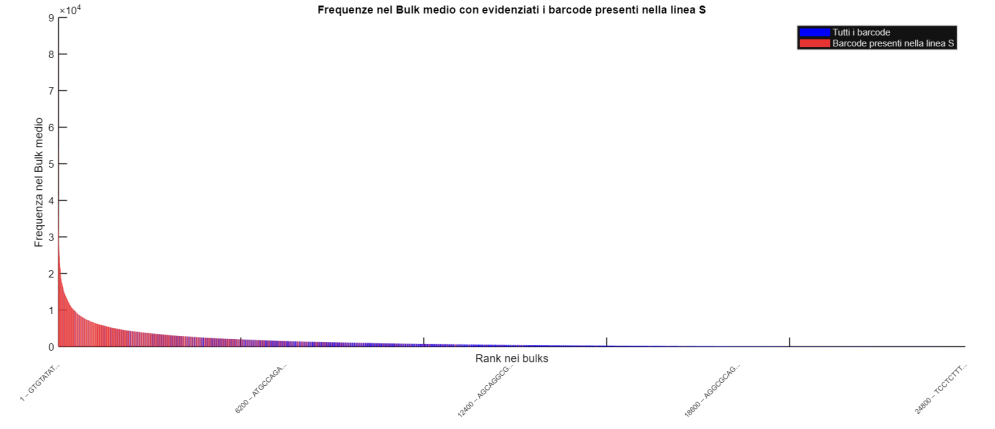
Da essi si osserva che la linea che mantiene il maggior numero di barcodes presenti nei bulks è la Linea S, in contrasto con le sue dimensioni iniziali. Le linee L e M, invece, mostrano quantità circa uguali. Tuttavia, si evidenzia che i barcodes presenti nei bulks, ma



(a) Bulk medio e Linea L



(b) Bulk medio e Linea M



(c) Bulk medio e Linea S

Figura 1.20: Grafici di frequenza per i barcodes ordinati per frequenza di lettura decrescenti nel bulk medio e aventi almeno 10 letture. Sono evidenziati i barcodes presenti anche in ciascuna linea.

assenti nelle tre linee, sono principalmente collocati nella coda del Bulk medio e, quindi, sono i meno proliferanti. Di conseguenza, anche questi diagrammi di frequenza sostengono l'assunzione di considerare i tre bulks come rappresentativi della popolazione di $38M$ di cellule.

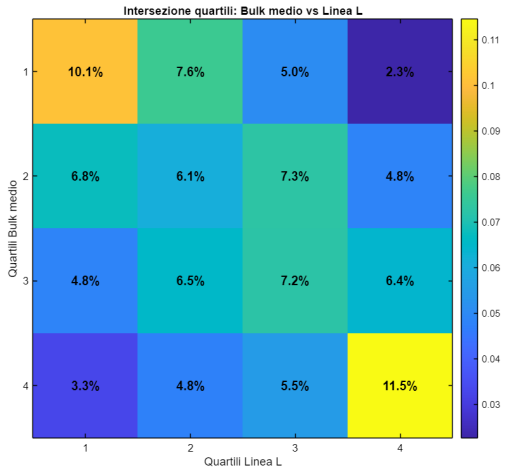
Questa assunzione è confermata anche dai grafici delle frequenze cumulate, i quali suggeriscono anche che le frequenze maggiori nelle tre linee, per i primi barcodes più abbondanti, potrebbero essere legate al processo di crescita-divisione che coinvolge le tre linee, il quale potrebbe comportare una selezione clonale più stringente e, quindi, una minore diversità clonale rispetto i bulks, ma maggiore tra di esse e una forte espansione, con conseguenti maggiori abbondanze.

L'ipotesi di considerare i bulks come rappresentativi è convalidata anche dai grafici di heatmap e scatter plot che, in aggiunta, forniscono ulteriori indicazioni sulle possibili ipotesi di dinamica clonale che caratterizzano le tre linee. Le heatmaps mostrano diagonali deboli e percentuali elevate fuori dalla diagonale, suggerendo che la distribuzione dei barcodes tra i quartili non viene mantenuta durante il processo di crescita-divisione. Tale osservazione è coerente con quanto emerge dagli scatter plot, nei quali i barcodes condivisi risultano ampiamente dispersi rispetto alla bisettrice, indicando variazioni sostanziali nelle abbondanze tra le linee. Nel complesso, questi risultati suggeriscono che il processo di crescita-divisione potrebbe non limitarsi ad aumentare la diversità clonale tra le linee, ma potrebbe comportare anche una riorganizzazione della gerarchia dei barcodes, con perdita di alcuni cloni ed espansione di altri. Inoltre, gli scatter plot relativi ai bulks mostrano che i barcodes non comuni a tutti e tre i campioni presentano frequenze medie e basse, suggerendo che le letture ad essi corrispondenti nei bulks e nelle linee siano proporzionali alla loro abbondanza. Ovvero, è possibile pensare che cloni più numerosi generino più letture perché le stesse cellule hanno una maggiore probabilità di essere campionate e sequenziate rispetto ai cloni rari.

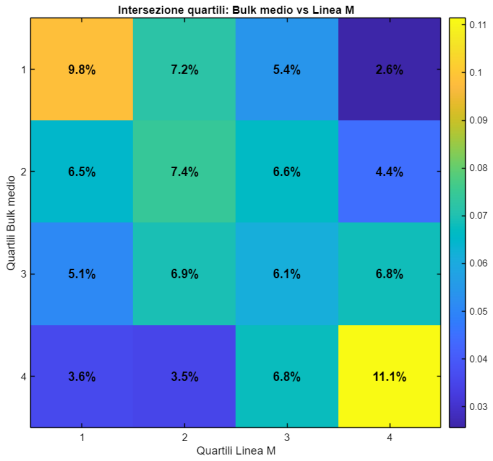
Tuttavia, considerando i bulks come rappresentativi della popolazione totale, è possibile valutare se e in quale misura i barcodes abbondanti o rari nei bulks mantengano il proprio rango anche al termine del processo di crescita-divisione affrontato dalle tre linee. A questo scopo si analizzano le heatmaps corrispondenti alle intersezioni tra il Bulk medio e le tre linee, pur riferendosi a due istanti temporali differenti. Esse sono riportate nella figura 1.21.

Come ci si aspetta, si notano percentuali significative anche al di fuori della diagonale e i colori dei relativi quartili differiscono tra le tre intersezioni. Infatti, sommando le percentuali sulla diagonale si ottiene:

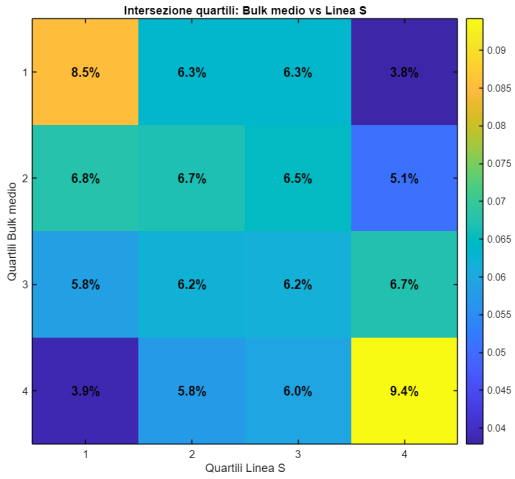
- Bulk medio VS Linea L : 34.9%;
- Bulk medio VS Linea M : 34.4%;



(a) Bulk medio vs Linea L



(b) Bulk medio vs Linea M



(c) Bulk medio vs Linea S

Figura 1.21: Heatmap di intersezione tra i quartili delle tre linee e il Bulk medio, in cui le percentuali indicano la distribuzione dei barcodes comuni.

- Bulk medio VS Linea S : 30.8%;

Pertanto, i valori totali bassi riflettono un'instabilità del ranking e cambiamenti di quartile da parte dei barcodes durante il processo di crescita-divisione che affrontano le tre linee. Tuttavia, tra le percentuali più alte, si notano quelle corrispondenti ai barcodes più rari e ai barcodes più abbondanti, molti dei quali rimangono nello stesso quartile nonostante il processo da loro affrontato. In aggiunta, la percentuale più elevata è relativa alla linea L, ciò sostiene l'ipotesi precedente secondo cui essa è maggiormente propensa al mantenimento delle gerarchie non dovendo affrontare la fase di estrema crescita all'inizio del processo. Tuttavia, la differenza con la percentuale relativa alla linea M è quasi trascurabile e, quindi, si può ipotizzare che la riorganizzazione del rango dei barcodes aumenti con il diminuire delle dimensioni iniziali del campione.

Capitolo 2

Simulazione dell'esperimento

2.1 Ipotesi di crescita malthusiana

La corretta interpretazione dei dati di barcoding richiede modelli matematici adeguati che riflettano la natura intrinsecamente stocastica dei processi biologici coinvolti. Pertanto, al fine di comprendere come i processi di crescita, selezione e campionamento conducano, al termine del processo di crescita, all'ottenimento degli insiemi definiti dai tre bulks, è utile assumere alcune ipotesi per poter simulare matematicamente l'esperimento.

Una prima ipotesi consiste nell'assumere che ogni giorno una cellula possa dare origine a un'altra cellula identica, non fare nulla o andare incontro alla morte. Il modello di crescita più semplice e comunemente utilizzato nel contesto delle popolazioni cellulari è il *modello di crescita malthusiana*.

Questa assunzione mira a sviluppare un modello di dinamica cellulare basato su un tasso di crescita netto, utile per simulare l'evoluzione della popolazione e analizzare come differenti distribuzioni dei tassi determinino l'abbondante proliferazione di alcuni barcodes e quella ridotta di altri, ottenendo così informazioni utili all'interpretazione dei meccanismi di dinamica cellulare.

Tuttavia, il modello di crescita malthusiano presenta diversi limiti quando viene applicato ai dati di barcoding. In primo luogo, esso nasce come modello *deterministico*: un barcode non può mai estinguersi e mantiene inalterate le gerarchie relative nel tempo. Queste caratteristiche impediscono di riprodurre fenomeni reali come l'estinzione di barcodes rari e i salti di soglia del numero di letture osservabili nei dati sperimentali. Inoltre, non contempla la variabilità stocastica associata ai processi di crescita e divisione cellulare. Di conseguenza, non può riprodurre le dinamiche di selezione che emergono nelle tre linee durante il processo che affrontano, pur rimanendo utile allo studio della distribuzione dei tassi nei tre bulks al momento del loro campionamento.

Per coerenza con la notazione definita nel capitolo precedente e avendo supposto che il numero di letture sia proporzionale al numero di barcodes, si indica con $S(T)$ il numero

di cellule al giorno T. Siano inoltre b il *tasso medio di riproduzione* e d il *tasso medio di mortalità*, due costanti positive. Allora, trascorso un intervallo di tempo ΔT, il numero di nuove cellule generate sarà $b\Delta TS$, e il numero di cellule che sono morte sarà $d\Delta TS$, da cui si ottiene un'equazione per S al giorno $T + \Delta T$:

$$S(T + \Delta T) = S(T) + b\Delta TS(T) - d\Delta TS(T) = S(T) + S(T)(b - d)\Delta T$$

cioè

$$\frac{S(T + \Delta T) - S(T)}{\Delta T} = (b - d)S(T)$$

passando quindi al limite per $\Delta T \to 0$, si ottiene:

$$\frac{dS}{dT}(T) = (b - d)S(T)$$

indicando con S_0 la popolazione iniziale e con $r = b - d$ costante il *tasso intrinseco di crescita*, la soluzione $S = S(T)$ evolve esponenzialmente come:

$$S(T) = S_0 e^{rT}$$

a seconda del segno di r, il modello predice che:

- r>0 $\to$ la popolazione cresce esponenzialmente senza limiti;
- r=0 $\to$ la popolazione mantiene dimensioni costanti;
- r<0 $\to$ la popolazione decresce esponenzialmente e si avvicina a 0.

Inoltre, grazie alla forma che assume la soluzione del modello, esso è anche denominato modello di *crescita esponenziale* [11], [12].

Per poter applicare il modello, è necessario ipotizzare anche che i tassi r associati a ciascun barcode rimangano stabili durante tutta la durata dell'esperimento e che ogni cellula generata abbia le stesse probabilità della cellula generatrice.

Al fine di stimare i tassi intrinseci di crescita associati a ogni barcode, è necessario procedere riscrivendo la soluzione del modello per ognuno di essi:

$$S_i^C(T) = S_0 e^{r_i T}$$

successivamente, considerando i due istanti di tempo:

- T_0 $\to$ giorno di inizio dell'esperimento in cui $S_0 = 1$ per ogni barcode;
- T_1 $\to$ giorno del campionamento dei bulks e della creazione delle linee, ovvero il 26-esimo giorno;

si ottiene

$$S_i^C(T_1) = S_0 e^{r_i(T_1 - T_0)}$$

applicando il logaritmo naturale a entrambi i lati, si ricava:

$$r_i = \frac{1}{T_1 - T_0} \ln\left(\frac{S_i^C(T_1)}{S_0}\right) = \frac{1}{T_1 - T_0}[\ln(S_i^C(T_1)) - \ln S_0]$$

In questo modo si determinano i tassi associati ai diversi barcodes nei tre bulks ed è quindi possibile visualizzarne una distribuzione in funzione del numero di barcodes aventi lo stesso tasso. Inoltre, per studiare quali dei tassi determinati comportano la sopravvivenza delle cellule successivamente al processo di crescita-divisione che esse affrontano nelle tre linee, è necessario identificare quali barcodes e quindi quali tassi presentano in comune le linee e i bulks. A tal fine, è essenziale considerare il *Bulk medio* come precedentemente definito. Pertanto, i tassi relativi ai barcodes presenti nelle tre linee vengono calcolati attraverso i seguenti passaggi:

$$S_i^{B_{medio}}(T_1) = S_i^{B_1}(T_1) + S_i^{B_2}(T_1) + S_i^{B_3}(T_1)$$

$$S_i^{B_{medio}}(T_1) = S_0 e^{r_{i,B_1}(T_1 - T_0)} + S_0 e^{r_{i,B_2}(T_1 - T_0)} + S_0 e^{r_{i,B_3}(T_1 - T_0)}$$

ma essendo $S_0 = 1$ per ogni barcode si ha:

$$S_i^{B_{medio}}(T_1) = e^{r_{i,medio}(T_1 - T_0)}$$

e, pertanto:

$$r_{i,medio} = \frac{1}{T_1 - T_0} \ln(S^{B_{medio}}(T_1)) = \frac{1}{T_1 - T_0} \ln(e^{r_{i,B_1}(T_1 - T_0)} + e^{r_{i,B_2}(T_1 - T_0)} + e^{r_{i,B_3}(T_1 - T_0)})$$

Non corrisponde quindi alla media aritmetica, ma è ad essa approssimabile nel caso in cui i tre bulks siano simili, poiché tale formulazione è la combinazione dei contributi dei tre tassi associati, in cui quello più elevato vi contribuisce maggiormente. Avendo osservato, nel capitolo precedente, che i tre bulks presentano distribuzioni molto simili tra loro, è possibile dividere le letture del Bulk medio per 3:

$$S_i^{B_{medio}}(T_1) = \frac{S_i^{B_1}(T_1) + S_i^{B_2}(T_1) + S_i^{B_3}(T_1)}{3}$$

e da esso calcolare $r_{i,medio}$ come:

$$r_{i,medio} = \frac{1}{T_1 - T_0} \ln(S_i^{B_{medio}}(T_1))$$

essendo $S_0 = 1$ per ogni barcodes, per poter confrontare le distribuzioni dei tassi nel Bulk medio con quelle nei singoli bulks.

2.2 Distribuzione dei tassi intrinseci di crescita

Dopo aver stimato i tassi intrinseci di crescita associati a ciascun barcode nei tre bulks e nel Bulk medio, è utile studiarne la distribuzione in funzione del numero di barcodes distinti aventi lo stesso tasso. Poiché molti tassi differiscono solo per piccole quantità, è utile raggrupparli in intervalli, detti *bin*, centrati in punti discreti, così da facilitarne la visualizzazione. A tal fine vengono identificati il valore minimo e massimo dei tassi presenti nel Bulk medio e l'intervallo complessivo viene suddiviso in 49 bin equidistanti. Tutti i tassi che ricadono all'interno dello stesso bin vengono approssimati con il valore del centro di quel bin.

Nella figura 2.1 è illustrato l'andamento di tale distribuzione con una soglia pari a una lettura.

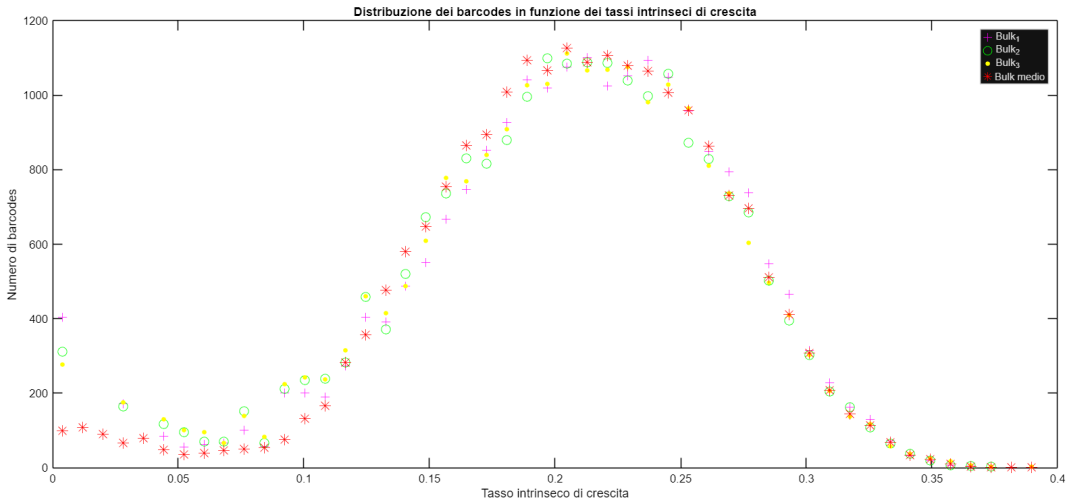


Figura 2.1: Distribuzione dei tassi intrinseci di crescita in relazione al numero di barcodes aventi un determinato tasso, nei tre bulks e nel Bulk medio.

Le distribuzioni ottenute mostrano un andamento complessivamente gaussiano con una coda più pronunciata verso sinistra. Sono presenti anche numerosi barcodes aventi un tasso prossimo allo zero. Tale comportamento potrebbe riflettere la presenza di valori estremi associati a un numero elevato di barcodes distinti e potenzialmente influenzati da errori di campionamento, come evidenziato nel capitolo precedente. Pertanto, considerando soglie di lettura maggiori, 5 e 10, si ottengono le distribuzioni riportate nelle figure 2.2, 2.3.

Le due distribuzioni mostrano differenze minime e i valori prossimi allo zero non sono più presenti. Passando da una soglia di 5 a una di 10 letture, si osserva la perdita di alcuni barcodes aventi piccoli e medi tassi di crescita; infatti, la seconda curva presenta un'altezza minore, suggerendo che alcuni barcodes sono cresciuti molto pur non riuscendo a superare la soglia di 10 letture. In tutti e tre i bulks, così come nel Bulk medio, emerge

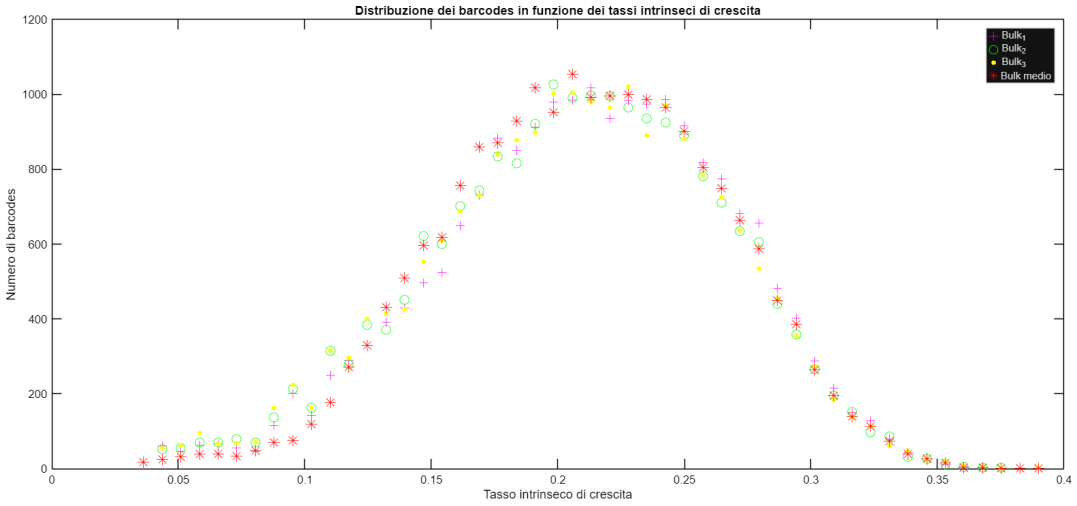


Figura 2.2: Distribuzione dei tassi intrinseci di crescita in relazione al numero di barcodes aventi un determinato tasso e almeno 5 letture nei tre bulks e nel Bulk medio.

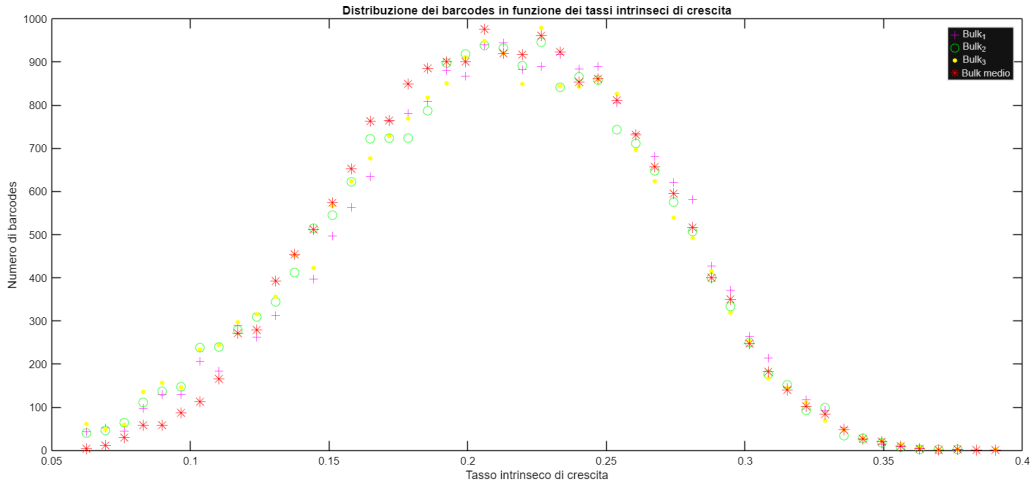


Figura 2.3: Distribuzione dei tassi intrinseci di crescita in relazione al numero di barcodes aventi un determinato tasso e almeno 10 letture nei tre bulks e nel Bulk medio.

un pronunciato andamento di tipo gaussiano: questo significa che i tassi si distribuiscono simmetricamente attorno a un valore medio, con una minoranza di barcodes che cresce molto rapidamente o molto lentamente. Inoltre, si osserva che le differenze tra i tassi presenti nei tre bulks si concentrano nella coda sinistra della distribuzione, corrispondente ai barcodes meno proliferanti. Questi barcodes, essendo caratterizzati da abbondanze molto basse, risultano maggiormente sensibili alle variazioni nella profondità di campionamento, alla stocasticità legata ai processi biologici e al sequenziamento, che amplificano le differenze tra i bulks.

Successivamente, analizzando quali dei barcodes presenti nel Bulk medio siano anche presenti in una delle tre linee, associandovi il corrispondente tasso intrinseco di crescita calcolato sulla base delle abbondanze nel Bulk medio, si ottengono le distribuzioni illustrate nella figura 2.4, se considerati tutti i barcodes presenzi, senza il filtro sulla soglia di lettura.

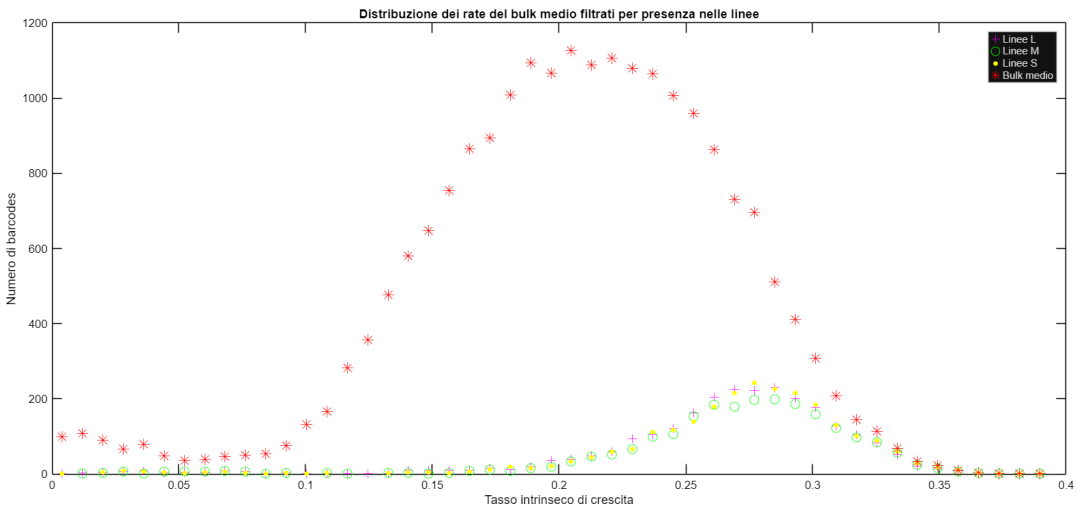


Figura 2.4: Distribuzione dei tassi intrinseci di crescita in relazione al numero di barcodes aventi un determinato tasso, nelle tre linee e nel Bulk medio.

Si evince che i barcodes maggiormente presenti nelle tre linee sono quelli associati a un tasso più elevato, e per questo le loro distribuzioni si concentrano nella parte destra del grafico. Inoltre, i barcodes meno proliferanti, pur essendo numerosi nel Bulk medio, nelle linee risultano quasi assenti. Tuttavia, nelle tre linee è presente anche un'abbondanza di barcodes associati ai tassi minori, evidenziata dalla presenza di una lunga coda sinistra, ma questo potrebbe essere nuovamente imputabile a errori di campionamento. Per questo motivo, risulta necessario analizzare le distribuzioni ristrette ai barcodes aventi almeno 5 e 10 letture in almeno una delle linee considerate. Esse sono riportate nelle figure 2.5, 2.6.

Anche in questo caso, le due figure mostrano differenze minime; inoltre, molti barcodes

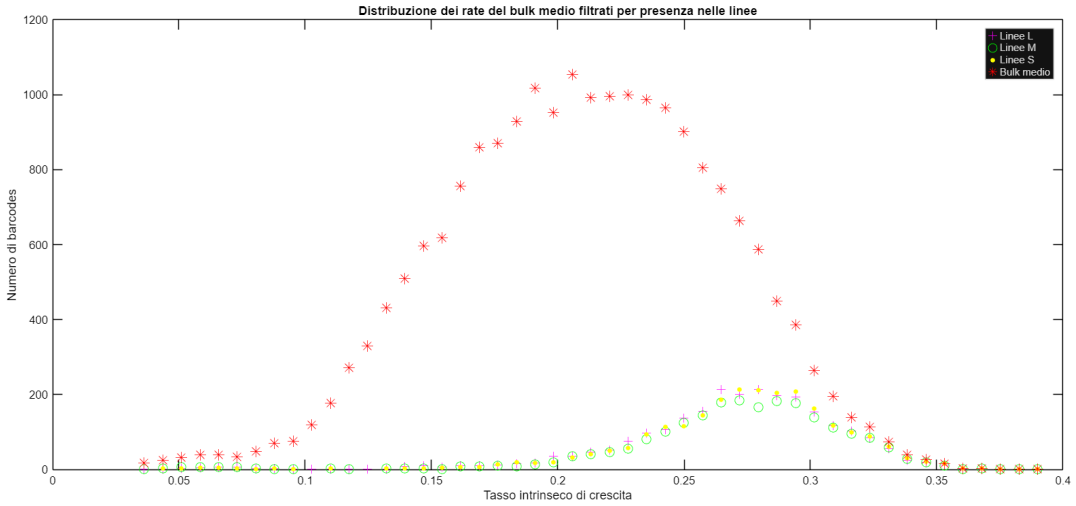


Figura 2.5: Distribuzione dei tassi intrinseci di crescita in relazione al numero di barcodes aventi un determinato tasso e almeno 5 letture nelle tre linee o nel Bulk medio.

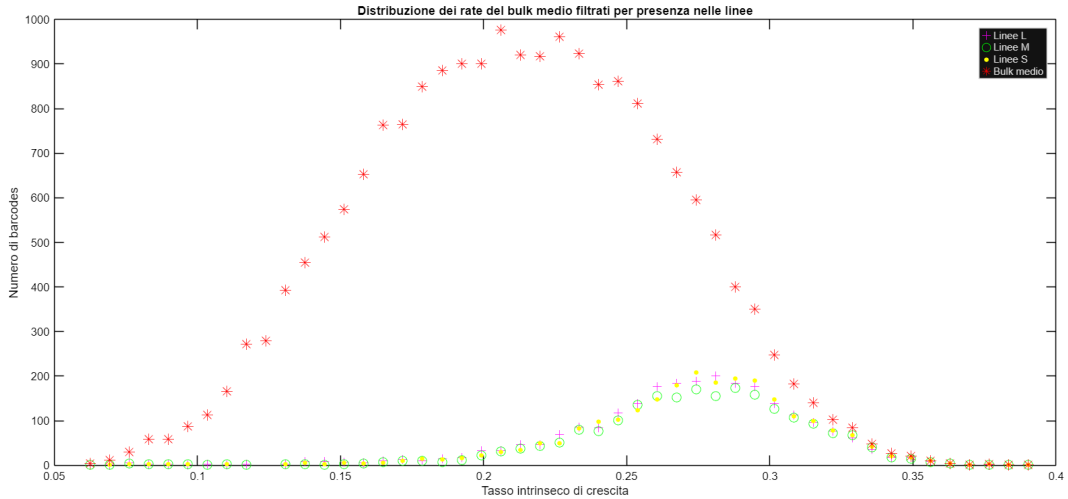


Figura 2.6: Distribuzione dei tassi intrinseci di crescita in relazione al numero di barcodes aventi un determinato tasso e almeno 10 letture nelle tre linee o nel Bulk medio.

a basso tasso risultano ora assenti. Questo suggerisce che i cloni dominanti si trovano molto al di sopra della soglia di 10 letture. Inoltre, nonostante la differenza di dimensione tra il Bulk medio, che contiene circa $6M$ di cellule, e le tre linee, con circa $2M$ di cellule ciascuna, le distribuzioni dei tassi intrinseci di crescita mantengono una forma gaussiana e risultano approssimativamente simili. Questo suggerisce che i cloni aventi tassi maggiori rimangono presenti anche dopo il processo di crescita-divisione che affrontano le tre linee. Dal punto di vista biologico, ciò evidenzia una selezione robusta dei cloni più competitivi, mentre quelli marginali contribuiscono solo in parte alla forma complessiva della distribuzione.

Capitolo 3

Conclusioni e possibili sviluppi

L'esperimento condotto dai ricercatori dell'IRCC di Candiolo mira a determinare se i meccanismi di resistenza alle terapie osservati nelle cellule cancerogene abbiano origine epigenetica o genetica, con l'obiettivo finale di sviluppare strategie terapeutiche mirate, capaci di identificare ed eliminare in modo selettivo le cellule resistenti. Prima di introdurre l'utilizzo di farmaci nello studio, è tuttavia essenziale comprendere a fondo i processi che regolano la dinamica cellulare.

Questo elaborato, attraverso l'analisi dei dati raccolti dai ricercatori, si propone quindi di fornire informazioni utili a orientare e supportare futuri sviluppi sperimentali.

L'analisi interpretativa dei dati ha permesso di formulare alcune ipotesi sulla dinamica clonale che ha interessato la popolazione cellulare considerata. Avendo assunto che il numero di letture sia proporzionale al numero di cellule, si è inizialmente osservato che i tre bulks possono essere considerati come rappresentativi della popolazione coinvolta dall'esperimento, poiché mostrano comportamenti molto simili sia nella distribuzione delle letture sia nel rango occupato da ciascun barcodes. Inoltre, il numero di barcodes esclusivi di ogni campione è molto limitato. In aggiunta, la quasi totalità dei barcodes sopra soglia presenti nelle tre linee è presente anche nei bulks e i barcodes che compaiono nei bulks ma non nelle linee si collocano quasi esclusivamente nella coda del Bulk medio, indicando che si tratta di cloni poco proliferanti, la cui assenza nelle linee è quindi coerente con la loro scarsa competitività. Va anche considerato che la maggioranza dei barcodes presenti nelle linee ma assenti nei bulks potrebbe derivare da errori di campionamento. Inoltre, confrontando il Bulk medio con ciascuna linea, si osserva che i barcodes condivisi sono principalmente concentrati nella parte iniziale della distribuzione del bulk medio e, perciò, sono i barcodes con frequenza maggiore. Inoltre, una parte dei barcodes condivisi tende a mantenere il loro rango delle letture: i cloni più abbondanti nei bulks risultano generalmente più abbondanti anche nelle linee, mentre quelli rari rimangono tali o scompaiono. Questi comportamenti, osservati in tutti i confronti, rafforzano l'idea che i bulks rappresentino fedelmente la popolazione di partenza e che le loro letture siano proporzionali all'abbondanza reale dei cloni.

Tuttavia, dall'analisi delle tre linee si evidenziano dinamiche clonali differenti rispetto a quelle dei tre bulks, che riflettono le loro dimensioni iniziali e il processo di crescita-divisione che ciascuna di esse deve affrontare.La linea S è quella che presenta la maggiore varietà di barcodes con una soglia di lettura pari a 5, mentre mostra una diversità intermedia quando la soglia viene aumentata a 10: un comportamento che può essere attribuito alla fase iniziale di forte espansione, durante la quale anche cloni inizialmente rari hanno avuto la possibilità di proliferare notevolmente, mentre la competizione ridotta ha limitato la perdita di barcodes per effetti stocastici. La linea L, al contrario, non attraversa una fase di espansione iniziale e tende quindi a conservare maggiormente la gerarchia clonale di partenza: un maggior numero di cloni abbondanti rimane tale, mentre quelli meno proliferanti si estinguono progressivamente nel corso delle numerose divisioni. La linea M, che parte da dimensioni intermedie, sembra collocarsi in ultima posizione dal punto di vista della diversità clonale, mostrando una popolazione eterogenea ma con un numero inferiore di barcodes capaci di superare le soglie di lettura rispetto alle linee S e L, probabilmente a causa di una competizione più marcata e della fase di espansione iniziale. In aggiunta, è plausibile pensare che i barcodes comuni identifichino cloni dotati di una maggiore probabilità di sopravvivenza e, di conseguenza, di essere campionati. Inoltre, il fatto che il numero totale di barcodes con almeno 5 e 10 letture sia simile nelle tre linee suggerisce che solo un numero limitato di cloni riesca a crescere oltre una certa soglia.

Le rappresentazioni grafiche fornite da heatmap e scatter plot confermano che il processo di crescita-divisione tende a non preservare la distribuzione iniziale dei cloni. Le heatmap mostrano infatti diagonali deboli e percentuali elevate fuori dalla diagonale, segno che i barcodes cambiano quartile. Gli scatter plot evidenziano una forte dispersione rispetto alla bisettrice, indicando variazioni sostanziali nelle abbondanze relative. Nel complesso, emerge che la fase di crescita-divisione non solo aumenta la diversità clonale tra le linee, ma comporta anche una riorganizzazione della gerarchia dei barcodes, con espansione di alcuni cloni e perdita di altri. Tuttavia, confrontando ciascuna linea con il Bulk medio, si osserva che i cloni estremamente rari o estremamente abbondanti tendono a mantenere la propria posizione. La linea L presenta la percentuale più alta sulla diagonale; pertanto, è quella che mantiene più stabilmente il proprio ranking, coerentemente con l'assenza di una fase di espansione iniziale che potrebbe alterare maggiormente le gerarchie. La linea M mostra un comportamento molto simile, mentre la linea S è quella che presenta la maggiore riorganizzazione, probabilmente proprio a causa della forte espansione iniziale. Nel complesso, sembra quindi che la stabilità del rango clonale diminuisca al diminuire delle dimensioni iniziali del campione da cui la linea ha avuto origine.

In aggiunta, dallo studio della legge di potenza emerge che, poiché nelle tre linee è necessario considerare un numero ridotto di barcodes per ottenere una buona approssimazione della distribuzione, il processo di crescita-divisione determina una marcata concentrazione

clonale. Nelle linee, infatti, un insieme ristretto di cloni domina la distribuzione finale, mentre nei bulks la stessa frazione di letture risulta distribuita tra un numero molto più elevato di barcodes, riflettendo una popolazione iniziale più eterogenea e bilanciata. Si deduce quindi che il processo di crescita-divisione trasforma una popolazione ampia ed equilibrata in una struttura più gerarchica e sbilanciata.

La corretta interpretazione dei dati di barcoding richiede anche l'impiego di modelli matematici adeguati. A tal fine, come spesso avviene in contesti biologici, si è applicato il modello di crescita malthusiana alla popolazione cellulare. In questo modo, è stato possibile determinare un tasso intrinseco di crescita per ciascun barcode e studiarne la distribuzione nei tre bulks. Tale analisi mostra che i tassi intrinseci di crescita si distribuiscono in modo complessivamente gaussiano, con una coda più pronunciata verso sinistra, soprattutto quando non viene applicata alcuna soglia di lettura. Applicando soglie più stringenti, come 5 o 10 letture, i valori prossimi allo zero scompaiono e le curve ottenute risultano molto simili tra loro, indicando che i cloni dominanti si collocano ben al di sopra di tali soglie.

In tutti e tre i bulks, così come nel Bulk medio, la forma della distribuzione è sostanzialmente gaussiana. Pertanto, i tassi si distribuiscono attorno a un valore medio, con una minoranza di barcodes che cresce molto rapidamente o molto lentamente. Le differenze tra i bulks si concentrano soprattutto nella coda sinistra, dove si trovano i barcodes meno proliferanti: proprio questi risultano maggiormente influenzati dalla profondità di campionamento e dalle fluttuazioni stocastiche, che amplificano le variazioni tra i tre campioni. Nel complesso, i bulks mostrano quindi una distribuzione dei tassi stabile e coerente: la forma globale della distribuzione, dominata da cloni con tassi intermedi, è sovrapponibile nei diversi bulks, mentre una maggiore variabilità emerge soltanto nella coda dei barcodes più rari.

Tuttavia, il modello di crescita malthusiano presenta diversi limiti, in quanto nasce come modello deterministico e, pertanto, un barcode non può mai estinguersi; inoltre, mantiene inalterate le gerarchie relative nel tempo. Di conseguenza, non può riprodurre le dinamiche di selezione che emergono nelle tre linee durante il processo di crescita-divisione che affrontano. Ciononostante, risulta utile per verificare quali tassi comportino la sopravvivenza e l'espansione di cloni anche all'interno delle tre linee.

L'analisi delle tre linee mostra che la distribuzione dei tassi intrinseci di crescita dei barcodes differisce in modo sostanziale da quella osservata nei bulks, nonostante anche le linee mantengano una forma approssimativamente gaussiana. Queste ultime presentano una distribuzione spostata verso i tassi più elevati: i cloni maggiormente rappresentati nelle linee sono infatti quelli associati ai tassi di crescita più alti nel Bulk medio, mentre i barcodes meno proliferanti, pur numerosi nella popolazione iniziale, risultano quasi del tutto assenti dopo il processo di crescita-divisione. Ciò suggerisce una selezione marcata dei cloni più competitivi, mentre quelli caratterizzati da tassi minori contribuiscono solo

marginalmente alla struttura finale della popolazione.

Alla luce di queste considerazioni, diventa quindi naturale chiedersi come tali limitazioni possano essere superate e quali strumenti o approcci possano essere sviluppati per descrivere in modo più realistico le dinamiche osservate, aprendo la strada a possibili approfondimenti e sviluppi futuri.

L'analisi svolta mostra chiaramente che un modello di crescita puramente deterministico non è sufficiente a descrivere in modo accurato le dinamiche osservate; pertanto, si deduce la necessità di introdurre una componente stocastica. Nella fase iniziale dell'esperimento, infatti, partendo da una popolazione estremamente ridotta, la crescita e l'estinzione dei singoli barcodes sono dominate dalla componente stocastica: in questo contesto, le differenze biologiche tra i cloni sono trascurabili e la loro sopravvivenza dipende soprattutto dal caso. Applicare fin da subito i tassi stimati nei bulks non permette né l'estinzione dei barcodes rari né la crescita di quelli presenti nella popolazione totale ma non campionati nei bulks, la cui assenza riflette i limiti del campionamento e non una reale scomparsa biologica. Inoltre, anche quando la popolazione raggiunge dimensioni tali da rendere sensato introdurre differenze nei tassi di crescita, le popolazioni clonali potrebbero essere ancora soggette a fluttuazioni casuali che contribuiscono alla variabilità osservata tra i campioni sperimentali. Per questi motivi, si deve prendere in considerazione l'introduzione di una componente stocastica sia nella fase iniziale sia in quella successiva, così da riprodurre in modo più realistico estinzioni, dispersioni e differenze tra cloni che emergono nei dati sperimentali.

In aggiunta, non avendo a disposizione dati relativi ai barcodes presenti nelle tre linee al momento della loro formazione e avendo assunto i bulks come rappresentativi della popolazione iniziale, un possibile sviluppo futuro consiste nel ricostruire le tre linee a partire dalla distribuzione del Bulk medio e simulare successivamente il processo di crescita-divisione, così da valutare se sia possibile riprodurre sperimentalmente le configurazioni osservate.

Bibliografia

[1] Istat (2025) *Cause di morte in Italia, anno 2022*. Statistiche report.

[2] Aiom, Airtum, Aiom Fondazione, Osservatorio Nazionale Screening, PASSI, PASSI d'Argento, SIAPEC-IAP (2024) *I numeri del cancro in Italia, 2024*. Intermedia editore.

[3] Bhang Hyo-eun C. et al. (2015) *Studying clonal dynamics in response to cancer therapy using high-complexity barcoding*. Nature medicine.

[4] Kebschull Justus M., Zador Anthony M. (2018) *Cellular barcoding: lineage tracing, screening and beyond*. Nature Methods: vol. 15, pp. 871–879.

[5] Naik Shalin H., Schumacher Ton N., Perié Leïla (2014) *Cellular Barcoding: A technical appraisal*. Experimental Hematology: vol. 42, pp. 598-608.

[6] Wang Xiaoyin, McManus Michael (2009) *Lentivirus Production*. Journal of Visualized Experiments: vol. 32, pp. e1499.

[7] Leibovich Nava, Goyal Sidhartha (2025) *Limitations and optimizations of cellular lineage tracking*. PLOS Computational Biology

[8] Corre Guillame, Galy Anne (2023) *Evaluation of diversity indices to estimate clonal dominance in gene therapy studies*. Molecular Therapy: Methods & Clinical Development: vol. 29.

[9] Clauset Aron, Rohilla Shalizi Cosma, Newman M. E. J. (2009) *Power-Law Distributions in Empirical Data*. SIAM Review: vol. 51, no. 4, pp. 661-703

[10] Newman M. E. J. (2005) *Power laws, Pareto distributions and Zip's law*. Contemporary Physics: vol. 46, pp. 323-351

[11] Chasnov Jeffrey R., (2025) *Mathematical Biology*. Hong Kong University of Science and Technology.

[12] Hastings Alan (1997) *Population biology. Concepts and Models*. Springer.