UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

"Confronto e analisi di tre strumenti computazionali per l'analisi della comunicazione cellula-cellula in dati sul cancro del colon retto"

"Comparison and analysis of three computational tools for cell-cell communication in colorectal cancer data"

Relatore: Prof. Giacomo Baruzzo

Laureando: Federico Ballarini

Correlatore: Prof. Zlatko Trajanoski, Giulia Cesaro

ANNO ACCADEMICO 2022 – 2023

Data di laurea 30/11/2023

Alla mia famiglia, a Chiara e ai miei amici
per aver sempre creduto in me
e sostenuto nel mio percorso.

ABSTRACT

La bioinformatica prevede lo sviluppo e l'applicazione di metodi computazionali per l'analisi e l'interpretazione di grandi set di dati biologici, da quelli relativi al sequenziamento del RNA alle informazioni sulla struttura delle proteine fino ai dati clinici dei singoli pazienti.

L'avvento della tecnologia di sequenziamento dell'RNA a singola cellula ha rivoluzionato il mondo della bioinformatica, consentendo di studiare l'espressione genica cellula per cellula, e recentemente, tale tecnica è stata utilizzata nell'ambito della comunicazione cellula-cellula.

La prima parte di questo progetto ha come obiettivo confrontare tre diversi tools bioinformatici denominati scSeqComm, NicheNet e CellPhoneDB, per l'inferenza della comunicazione cellulare a partire da dati di sequenziamento dell'RNA a singola cellula.

Di ciascuno si è voluto osservare e comprendere le diverse metodologie utilizzate, i diversi aspetti biologici presi in considerazione e le prestazioni computazionali, in modo tale da poter fornire a future ricerche un criterio decisionale circa la metodologia da usare a seconda delle diverse esigenze.

In particolare, questi metodi sono stati applicati a dati relativi a 216 pazienti affetti da tumore al colon-retto, con lo scopo di inferire le interazioni tra cellule tumorali e cellule del sistema immunitario rilevanti in questa patologia.

La seconda parte del progetto, invece compara i risultati ottenuti, in output dai tre metodi con quelli provenienti da altre ricerche già presenti in letteratura riguardo a geni e/o fattori che sono coinvolti in questo tipo di malattia.

Il nostro lavoro mira quindi ad aiutare i ricercatori nella scelta più efficace di strumenti computazionali adeguati alle loro esigenze, per la quantificazione delle

interazioni cellulari a partire da un set di dati di sequenziamento dell'RNA a singola cellula.

1. Capitolo 1

1.1 Definizione di Gene

I geni sono considerati le unità base dell'ereditarietà, vengono trasmessi dai genitori alla prole e contengono le informazioni necessarie per specificare i tratti fisici e biologici. La maggior parte dei geni codifica per delle specifiche proteine, o segmenti di proteine, che hanno funzioni diverse all'interno del corpo. I geni codificanti per delle proteine negli esseri umani sono circa 20.000: è interessante notare che i geni codificanti proteine occupano solo l'1,5% dell'intero genoma umano (Toledo & Saltsman, 2012).

La definizione della parola gene è stata a lungo fonte di dibattito scientifico, quella comunemente più accettata è: "unità di trascrizione, ossia una sequenza di acidi nucleici (DNA, o più raramente RNA in alcuni virus) che portano le informazioni per produrre un particolare prodotto genico" (Rossi, 2010). Pertanto, sono geni tutti i segmenti del genoma suscettibili di essere trascritti; se volessimo utilizzare una semplice similitudine per avere un'idea più chiara potrebbe essere la seguente: "le proteine sono i mattoni e la malta che costituiscono le nostre cellule e i nostri tessuti, i geni sono la parte del nostro genoma che codifica le informazioni per produrre quelle proteine" (Salzberg, 2018).

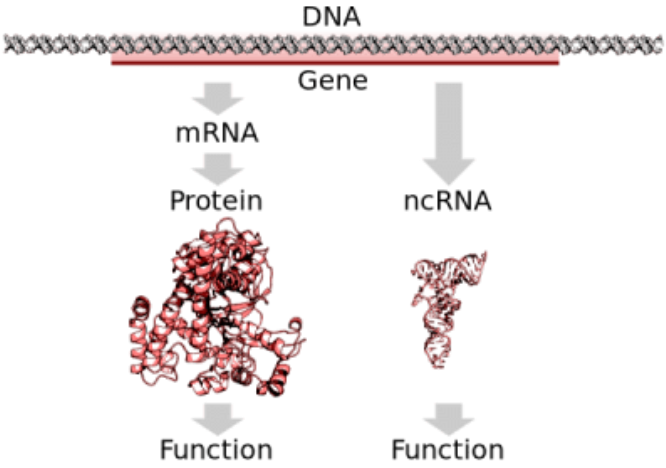


Figura 1, dal DNA alla proteina, (Toledo & Saltsman, 2012)

1.2 Sequenziamento RNA a singola cellula

Lo sviluppo di tecniche di sequenziamento di RNA e DNA ha permesso di osservare le cosiddette sequenze geniche, al cui interno sono codificati i geni di un determinato organismo, i quali altro non sono che le “istruzioni” che consentono di esprimere una caratteristica nel tempo e nello spazio.

Il “single cell RNA sequencing” (scRNA-seq) è una delle principali tecniche per il sequenziamento dell’RNA, sviluppata nel 2009 è diventato un approccio all’avanguardia per svelare l’eterogeneità e la complessità delle trascrizioni di RNA all’interno delle singole cellule, oltre a rivelare la composizione di diversi tipi e funzioni cellulari all’interno di strutture altamente organizzate, come tessuti e organi (Figura 3).

Più recentemente, scRNA-seq è stato applicato a coorti di pazienti che includono caratteristiche diverse, come ad esempio determinate sottocategorie di malattie, pazienti sani o malati, etc. ; questi set di dati multi-campione e multi-condizione consentono inferenze statistiche tra gruppi, al fine di trovare specifici geni marcatori. Ciò ha permesso di costruire strutture comunemente note con il nome di atlante (o atlas), i quali sono risorse chiave per comprendere possibili soluzioni nel trattamento di alcune condizioni o malattie.

1.3 Comunicazione cellulare

La comunicazione cellulare è caratterizzata dal tipo di geni espressi e utilizzando queste informazioni possiamo iniziare a rispondere a domande come "Il perché la sovra-espressione o la sotto-espressione di un determinato set di geni possa provocare lo sviluppo di determinate malattie".

Oltre a confrontare i geni espressi tra diversi tipi di cellule, possiamo anche studiare come questi modelli di espressione genica cambino nel tempo o in risposta a determinati stimoli.

Negli organismi pluricellulari, ciascuna cellula può ricevere o inviare informazioni ad altre cellule e tale meccanismo rende ciascuna cellula dipendente dalle altre. La comunicazione avviene essenzialmente attraverso l'interazione tra due molecole, il ligando e il recettore. Se la cellula "A" comunica con la cellula "B", il ligando sarà la molecola espressa da "A" e il recettore la molecola espressa dalla cellula "B".

I ligandi possono avere diversa natura biochimica (steroidei, peptidi, derivati di aminoacidi, ioni), in particolare possono essere o molecole solubili rilasciate dalle cellule che le producono oppure molecole transmembrana che rimangono associate alla cellula che le producono (Perroteau, s.d.).

I tipi di comunicazione cellulare per i ligandi solubili sono:

1. Comunicazione endocrina: i ligandi (ormoni) sono rilasciati dalle cellule endocrine nei vasi e attraverso la circolazione raggiungono, mantenendosi a distanza, le cellule bersaglio che esprimono specifici recettori;
2. Comunicazione paracrina: i ligandi sono rilasciati nello spazio intercellulare e per diffusione raggiungono le cellule bersaglio;
3. Comunicazione autocrina: le cellule che rilasciano il ligando esprimono anche il recettore;
4. Comunicazione sinaptica: il ligando solubile chiamato neurotrasmettitore è rilasciato dal neurone nello spazio circostante le sinapsi e non può diffondere al di fuori di queste;

Per quanto riguarda i ligandi transmembrana o associati alla matrice extracellulare, esistono due tipi di comunicazione:

1. Comunicazione giustacrina: il ligando è una molecola transmembrana oppure una molecola della matrice extracellulare e questa comunicazione richiede il contatto diretto cellula-cellula oppure cellula-matrice;
2. Adesione cellulare: si instaura fra cellula e cellula e fra cellule e matrice extra-cellulare, le adesioni fra cellule nell'embrione degli animali determinano la formazione dei tessuti e nell'adulto la struttura del corpo;

La comunicazione intercellulare è quindi definita dalle varie modalità e strutture che le cellule utilizzano per comunicare tra loro direttamente o attraverso l'ambiente che le circonda, mediante i ligandi e i recettori (Figura 2.).

L'interazione tra ligando e recettore attiva poi il dominio intracellulare del recettore che a sua volta attiva una cascata di reazioni che propagano il segnale all'interno della cellula (trasduzione del segnale intracellulare). L'effetto finale è l'osservazione di una risposta della cellula ricevente, che può avvenire tramite la regolazione trascrizionale di geni a valle.

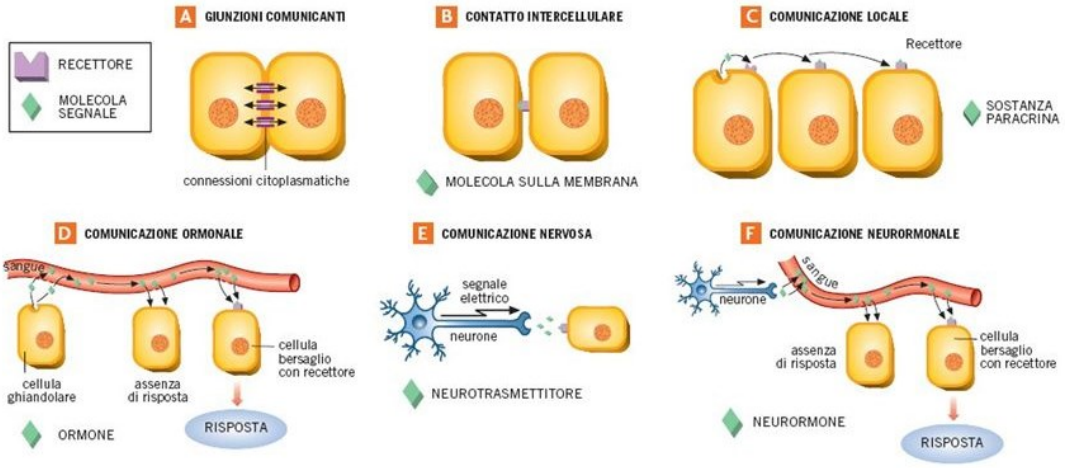


Figura 2, tabella riassuntiva riguardo i tipi di comunicazione cellulare, (Perroteau, s.d.)

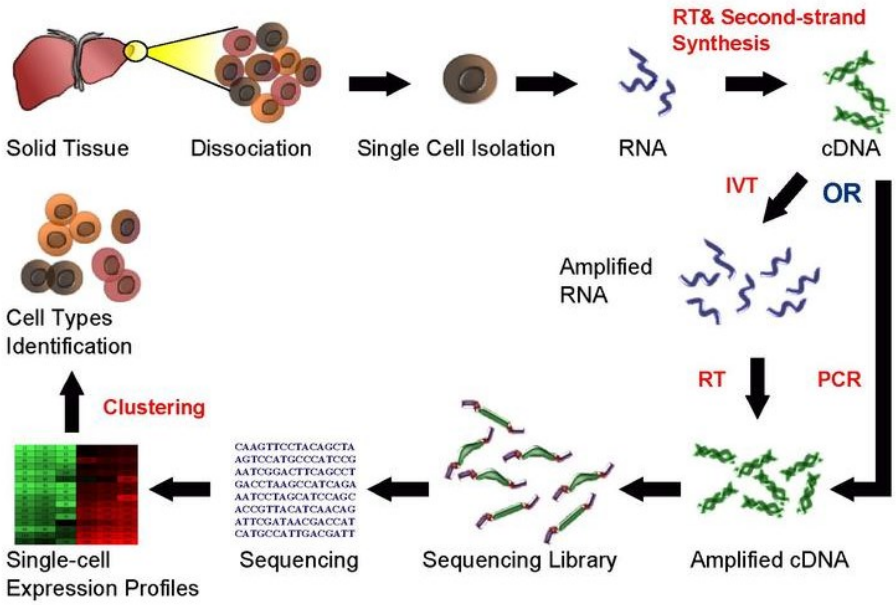
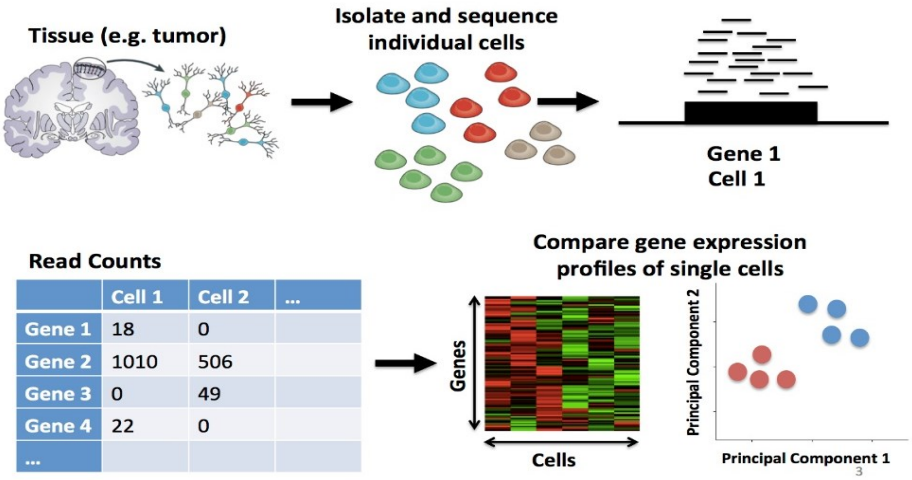


Figura 3, flusso di lavoro nel contesto del sequenziamento dell'RNA a singola cellula, (Haque, 2017)

Attualmente, scRNA-seq è diventato una scelta favorevole per studiare le questioni biologiche chiave dell'eterogeneità cellulare. Negli ultimi anni, scRNA-seq è stato applicato a varie specie, in particolare a diversi tessuti umani (compresi quelli normali e tumorali), e questi studi hanno rivelato una significativa variabilità dell'espressione genica da cellula a cellula (Figura 4). Negli ultimi anni sono stati proposti diversi protocolli scRNA-seq, che hanno ampiamente facilitato la comprensione dell'espressione genica a livello di singola cellula (Haque, 2017).



	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			

Figura 4, dal tessuto alla matrice contenente i "read counts", grazie al scRNA-seq (Haque, 2017)

Ciascun protocollo scRNA-seq presenta vantaggi e svantaggi, con il risultato che diversi approcci hanno caratteristiche distinte e prestazioni differenti nel condurre uno studio trascrittomico su singola cellula. Potrebbe essere necessario impiegare una specifica tecnologia scRNA-seq cercando di equilibrare tra obiettivo della ricerca e costo di sequenziamento.

A causa della bassa quantità del materiale di partenza, scRNA-seq presenta limitazioni, rispetto al bulk RNA-seq la quale è stata la tecnica principale di sequenziamento dell'RNA fino all'arrivo del scRNA-seq. Il sequenziamento dell'RNA a singola cellula produce dati più rumorosi e variabili: il rumore tecnico e la variazione biologica (ad esempio, la trascrizione stocastica) sollevano sfide importanti per l'analisi computazionale dei dati che si possono ottenere. Il controllo qualità (QC) è quindi un passo fondamentale per identificare e rimuovere i dati scRNA-seq di bassa qualità e per ottenere risultati affidabili e riproducibili; perciò, per gestire efficacemente l'elevata variabilità dei dati scRNA-seq, è necessario prestare attenzione al metodo scelto per "collezionarli".

1.4 Analisi bioinformatica della comunicazione cellulare

La bioinformatica è una scienza multidisciplinare, il cui obiettivo è analizzare l'informazione biologica al fine di formulare ipotesi su alcuni dei principali processi che permettono la vita; gli attuali ambiti di ricerca includono ad esempio l'allineamento di sequenze di basi nucleotidiche, la predizione delle strutture proteiche, l'annotazione genica, lo studio dell'espressione genica e l'analisi delle interazioni proteina-proteina.

L'avvento della tecnologia di sequenziamento dell'RNA a singola cellula ha rivoluzionato il mondo della bioinformatica, consentendo di studiare l'espressione genica, e recentemente, tale tecnica è stata utilizzata in particolare nell'ambito della comunicazione cellula-cellula. Tale tecnologia offre quindi un'opportunità unica nel comprendere il complesso funzionamento delle cellule ed essere in grado di

descrivere come avviene la trasduzione del segnale nelle cellule, concetto di fondamentale importanza per la medicina e la farmacologia.

I principali passaggi coinvolti nella analisi della comunicazione cellulare sono:

1. identificare molecole coinvolte nella trasmissione del segnale (ligandi e recettori)
2. misurare l'evidenza, a partire dai dati, di una comunicazione in atto attraverso una specifica coppia ligando-recettore tra una coppia di cellule (o una coppia di gruppi omogenei di cellule)

1.5 Database ligando recettore

Nel 1971, è stato segnalato per la prima volta l'importante ruolo dei ligandi e dei recettori nel cancro al seno, e da qui in poi la relazione tra le coppie ligando-recettore e malattie è stata gradualmente scoperta. Ligandi e recettori sono coinvolti nello sviluppo delle malattie principalmente per due motivi:

1. alterazioni strutturali o genetiche;
2. cambiamenti dell'espressione nelle coppie ligando-recettore.

La comprensione dell'interazione ligando-recettore è alla base degli attuali studi sulla comunicazione intercellulare e fornisce ai ricercatori una visione più approfondita dei processi dell'attività biologica cellulare e della progressione di determinate malattie, non sorprende infatti come siano stati effettuati diversi studi con gli approcci più disparati per la costruzione di database ligando-recettore, nell'uomo ma anche sugli animali quali soprattutto i topi, generalmente usati per studi sui farmaci (Ma, Zhang, Song, Wang, & Wei, 2021).

L'identificazione dei recettori che legano le proteine dei ligandi chiave fornisce preziose informazioni meccanicistiche riguardanti la trasduzione del segnale, l'azione dei farmaci o di altri effetti.

Per determinare i potenziali ligandi e recettori, si parte solitamente generando un ampio numero di dati sperimentali in ambiente controllato e classificando un certo numero di proteine umane o di topo, a seconda del database che si vuole configurare, secondo criteri uniformi in cui le proteine secrete o distribuite nello spazio extracellulare sono incluse come potenziali ligandi, mentre quelle distribuite o ancorate sulla membrana o sulla superficie cellulare sono considerate potenziali recettori. (Lopez-Garcia, Demiray, & Ru, 2018)

Con la crescente scoperta di recettori, ligandi e le loro interazioni, il numero di database contenenti interazioni ligandi-recettori è cresciuto sempre più negli anni, tuttavia, è noto ai ricercatori come ci siano da scoprire ancora moltissimi recettori, ligandi e le relazioni ad essi associati. Pertanto, dopo aver analizzato i complessi ligando-recettore esistenti, diversi ricercatori hanno sviluppato dei software per la simulazione e l'analisi predittiva delle loro interazioni, ad esempio DOCK, Autodock, AutoDock Vina, iGEMDOCK e RosettaDock. Questi numerosi database induttivi e strumenti di simulazione aiutano i ricercatori a migliorare lo studio sui complessi ligando-recettore e le loro interazioni, il che a sua volta contribuisce allo sviluppo di farmaci e al trattamento delle malattie.

La convalida delle interazioni ligando-recettore può, per ovvi motivi, essere impegnativa e richiedere molto tempo, si è anche scoperto come alcuni ligandi o alcuni recettori possano legarsi rispettivamente a più recettori o più ligandi a seconda del caso.

1.6 Tumore al colon retto

Il tumore del colon-retto è un cancro che si forma nei tessuti del colon (la parte più lunga dell'intestino crasso) o del retto (la parte dell'intestino crasso più vicina all'ano). Colon e retto fanno parte dell'intestino, l'organo che assorbe le sostanze nutritive assunte con il cibo. Ha l'aspetto di un tubo cavo la cui lunghezza varia da persona a persona tra i 4 e i 10 metri, ma in media è lungo 7 metri ed è suddiviso in due parti che hanno funzioni diverse:

1. l’intestino tenue ha la funzione di portare a termine la digestione iniziata in bocca e proseguita nello stomaco, a cui è collegato tramite il duodeno, ed è a sua volta suddiviso in tre parti distinte: duodeno, digiuno e ileo;
2. l’intestino crasso, la cui funzione principale è invece assorbire acqua per compattare le feci e comprende il colon, sigma e retto, il quale termina infine nel canale anale.

Come tutti i tumori, anche il tumore del colon-retto è dovuto alla crescita incontrollata di cellule, in questo caso di quelle epiteliali della mucosa che riveste internamente l’intestino. Questi tumori nascono soprattutto nel colon e nel retto, mentre i tumori del piccolo intestino e del canale anale sono molti rari (costituiscono il 2-3% di tutti i tumori del tratto digerente). I tumori del colon sono quasi tre volte più frequenti dei tumori del retto, e si manifestano con modalità diverse sia a livello clinico che molecolare, questo condiziona il tipo di trattamento che può essere: locale (chirurgia e/o radioterapia) o sistemico (chemioterapia, terapia biologiche o molecolari e immunoterapia), oltre alla sequenza in cui questi tipi diversi di cure vengono offerte al paziente.

Secondo le stime GLOBOCAN 2020 fornite dall’Agenzia internazionale per la ricerca sul cancro (AIRC, 2021), il tumore del colon-retto rappresenta il 10 per cento di tutti i tumori diagnosticati nel mondo, ed è terzo per incidenza dopo il cancro del seno femminile (11,7%) e del polmone (11,4%).

La malattia, abbastanza rara prima dei 40 anni, è maggiormente diffusa in persone fra i 60 e i 75 anni, con poche distinzioni fra uomini e donne. In Italia, le stime più recenti parlano di oltre 43.700 nuovi casi all’anno: circa 20.282 nelle donne e 23.420 negli uomini. Nell’ultimo ventennio, grazie principalmente allo screening di popolazione, l'incidenza è in diminuzione in Italia in entrambi i sessi: entrando nel dettaglio, i dati più aggiornati mostrano che dal 2008 al 2016 ogni anno l’incidenza si è ridotta del 3-4 % nella fascia di età sottoposta a screening, cioè nelle persone tra i 50 e 69 anni di età. Tuttavia, in controtendenza, dati più recenti mettono invece in luce un aumento annuo dello 0,4 % dei casi di tumore in

individui con meno di 50 anni di età e pertanto non coperti dallo screening. Questo aumento di casi riguarda in particolar modo soggetti molto giovani, fino all'età di 30 anni al momento della diagnosi, per cause non ancora conosciute e al momento oggetto di studio anche da parte di ricercatori in Italia e nel mondo. Inoltre, diversamente dalle attese, i tumori del colon-retto insorti in soggetti di giovane età, in più del 50% dei casi non sono dovuti a familiarità per questo tipo di cancro o a malattie genetiche predisponenti allo sviluppo di tumori, ma sembrano essere sporadici. La mortalità per il cancro del colon-retto è in forte calo, con tassi diminuiti di circa il 10 per cento nell'ultimo quinquennio, questi progressi sono attribuibili principalmente ai programmi di screening, alla diagnosi precoce e al miglioramento delle terapie, sia chirurgiche che mediche.

Negli anni più recenti anche l'immunoterapia con inibitori dei checkpoint immunitari si è aggiunta alle opzioni di trattamento disponibili per il tumore del colon-retto, da sola o in combinazione con altre terapie. Pembrolizumab, Nivolumab e Ipilimumab sono alcuni dei farmaci immunoterapici che hanno mostrato maggior efficacia contro questi tumori anche se il loro utilizzo è limitato a quei casi che presentano una caratteristica molecolare detta instabilità dei microsatelliti o MSI.

1.7 Obiettivi e organizzazione della tesi

Questo elaborato ha lo scopo di andare a confrontare tre diversi tools bioinformatici per l'inferenza della comunicazione cellulare in dati provenienti da un atlas contenente pazienti affetti da tumore al colon-retto:

1. scSeqComm
2. NicheNet
3. CellPhoneDB

Di ciascuno si è voluto osservare e comprendere le prestazioni in determinate condizioni di lavoro, in modo tale da poter fornire a future ricerche un criterio decisionale circa la metodologia da usare a seconda delle diverse esigenze.

I metodi qui analizzati utilizzano un set di dati contenente informazioni provenienti da sequenziamento di RNA a singola cellula. In particolare, l'atlas analizzato è composto da 216 pazienti malati, per un totale di 1.010.297 cellule ottenute analizzando campioni di tessuto sia sano che tumorale.

La seconda parte del progetto, mira invece a comparare i risultati ottenuti in output dai tre metodi con quelli provenienti da altre ricerche già presenti in letteratura riguardo a geni e/o fattori coinvolti in questo tipo di malattia.

Questo lavoro ha quindi lo scopo di andare ad aiutare i ricercatori nella scelta più efficace di strumenti computazionali adeguati alle loro esigenze, per la quantificazione delle interazioni cellulari a partire da un set di dati di sequenziamento dell'RNA a singola cellula.

2. Capitolo 2

2.1 Atlas tumore al colon retto

Il set di dati utilizzato per questa analisi è tutt'ora in via di sviluppo presso la Medizinische Universitat di Innsbruck (Innsbruck, s.d.).

Il dataset è composto da 216 pazienti affetti da tumore al colon-retto, dai quali, grazie a sequenziamento di RNA a singola cellula, sono state raccolte le informazioni riguardanti 1.010.297 cellule e 18.224 geni.

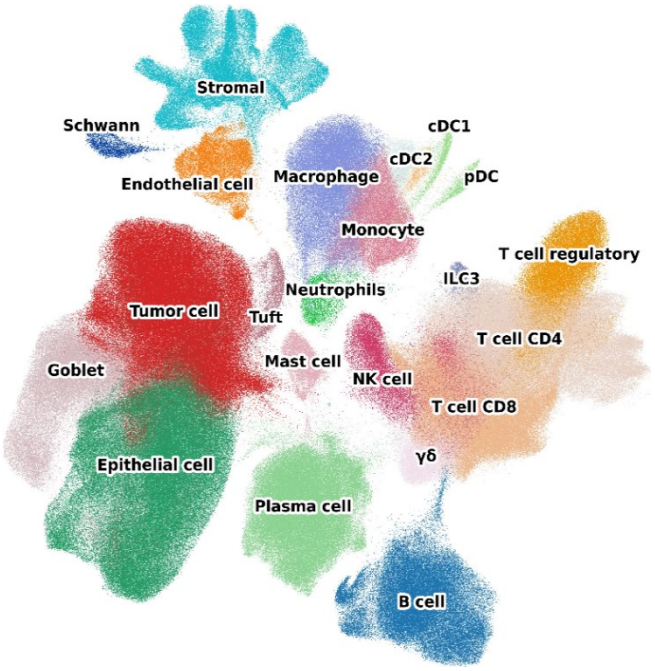


Figura 5, plot UMAP dell'atlas utilizzato

I 216 pazienti sono così suddivisi:

- 104 microsatellite stabile (MSS)
- 46 microsatellite instabile – alto (MSI-H)
- 1 microsatellite instabile (MSI)
- 65 non identificati

Tutti i campioni di tessuto provengono dal colon o dal retto e la ulteriore classificazione MSS, MSI-H o MSI è in riferimento ai tessuti dai quali la singola cellula è stata prelevata.

Il tumore del colon-retto mostra cambiamenti molecolari variabili dovuti a due principali meccanismi di instabilità genetica:

- instabilità cromosomica

- instabilità dei microsatelliti

Recentemente sono stati proposti due sistemi di classificazione patologica molecolare per il tumore del colon-retto (Müller MF, 2016). Grazie all'analisi molecolare integrata del progetto Cancer Genome Atlas, utilizzando tecnologie di sequenziamento si è potuto appunto classificare tale malattia in tre gruppi principali:

1. Microsatelliti instabili (MSI), caratterizzati da un differente numero di ripetizioni di sequenze brevi di DNA tra il tessuto tumorale e quello normale.
2. Microsatelliti instabili – alti (MSI – H), caratterizzati da mutazioni della DNA polimerasi.
3. Microsatelliti stabili (MSS), presenti in circa l'84% dei tumori al colon-retto e caratterizzati da un'alta frequenza di alterazioni nel numero di copie somatiche del DNA.

I microsatelliti sono sequenze ripetitive di DNA distribuite lungo le regioni codificanti e non codificanti del genoma, le quali sono particolarmente sensibili agli errori di mancato appaiamento del DNA. Il nome di DNA “satellite” è stato conferito dopo le prime osservazioni, nelle quali si notò come centrifugando il DNA in una provetta si avesse la separazione in uno strato di DNA detto “massa” e in uno strato di DNA ripetitivo detto “satellite” (JR, 2007).

Lo status di instabilità microsatellitare rappresenta un fattore prognostico e predittivo per i tumori colorettali: pazienti con elevata instabilità microsatellitare (MSI-H) hanno una prognosi migliore di quelli con microsatelliti stabili (MSS).

Il test dell’instabilità dei microsatelliti ricopre, dunque, un ruolo importante nella pratica clinica.

Il test può costituire inoltre un fattore di valutazione per:

- l’identificazione delle forme ereditarie o sporadiche del cancro colorettale

- le terapie da utilizzare
- le possibili prognosi della malattia

All'interno dell'Atlas sono presenti diversi tipi di cellule e noi in particolare ci siamo voluti soffermare su quelle del sistema immunitario, oltre che ovviamente su quelle tumorali, di seguito un elenco proposto:

- *Cellule B*, fanno parte del sistema immunitario e hanno il compito di produrre gli anticorpi che reagiscono a un particolare antigene;
- *CAF*, fibroblasti attivati presenti nel microambiente del tumore. Essi assolvono a numerose funzioni che sono volte a garantire la maturazione e il trofismo di cellule tumorali;
- *ILC3,* cellule linfocitiche innate contribuiscono a regolare la risposta immunitaria ed a mantenere l'omeostasi dei tessuti;
- *Granulociti*, globuli bianchi caratterizzati dalla presenza nel citoplasma di grossi granuli;
- *Macrofagi/Monociti*, i macrofagi si sviluppano dai monociti che sono un tipo di globuli bianchi; tale trasformazione avviene nel momento in cui si spostano dal flusso sanguigno ai tessuti, cioè non appena insorge un'infezione;
- *Mastociti,* cellule immunitarie che sono generate nel midollo osseo e sono presenti in tutti i tessuti, soprattutto nella vicinanza di piccoli vasi e terminazioni nervose;
- *Cellule NK,* classe di cellule citotossiche del sistema immunitario, particolarmente importanti nel riconoscimento e distruzione di cellule cancerose o infette da virus;
- *Neutrofili*, i più numerosi globuli bianchi riscontrabili nel sangue circolante;
- *Plasmociti,* cellule del sistema immunitario che secernono grandi quantità di anticorpi, si differenziano dalla cellula B sotto stimolazione dei "linfociti T helper" e del riconoscimento diretto dell'antigene per cui sono specifiche.

- *Cellule T*, responsabili della produzione di anticorpi e della risposta cellulare ai virus;
- *Tuft,* cellule chemo-sensoriali del rivestimento epiteliale dell'intestino;
- *cDC,* sono anticorpi rivolti contro le cellule della mucosa gastrica;
- *pDC,* cellule dendritiche plasmocitoidi, un sottoinsieme delle cellule dendritiche, specializzate nella secrezione di alti livelli di interferoni di tipo I;

2.2 Database ligando recettore utilizzati

I database utilizzati sono 3:

1. Efremova
2. Ramilowski
3. Kumar

	Numero di coppie LR	**Proteine subunità**	**Riferimenti**
Efremova	881	Si	Efremova et al 2020
Ramilowski	2442	No	Ramilowski et al 2015
Kumar	1901	No	Kumar et al 2018

Tabella 1, principali caratteristiche database LR

Sono stati presi in considerazione rispettivamente in quanto: Efremova è il database con il quale è stato sviluppato e testato CellPhoneDB (Efremova, Vento-Tormo, & Teichmann, 2019), uno dei tre metodi bioinformatici analizzati in questa tesi. Ramilowski ha invece permesso di testare e sviluppare NicheNet (Browaeys,

Saelens, & Saeys, 2019) mentre per quanto riguarda Kumar abbiamo deciso di utilizzarlo perché presentava un numero di coppie ligando-recettore (LR) comprese fra quelle di Efremova e Ramilowski, permettendoci di osservare meglio quanto il numero di coppie ma anche di singoli ligandi e singoli recettori influenzasse il funzionamento dei metodi.

Di seguito sono invece riportate due tabelle riassuntive circa il numero di coppie LR presenti nei tre database, le coppie in comune fra i tre e alcune delle loro principali caratteristiche.

Numero di LR comuni fra i database	**Efremova**	**Ramilowski**	**Kumar**
Efremova	Ligandi: 396 Recettori: 358 Coppie LR: 881	Ligandi: 167 Recettori: 99 Coppie LR: 152	Ligandi: 161 Recettori: 98 Coppie LR: 145
Ramilowski	X	Ligandi: 695 Recettori: 652 Coppie LR: 2442	Ligandi: 642 Recettori: 589 Coppie LR: 1894
Kumar	X	X	Ligandi: 645 Recettori: 593 Coppie LR: 1901

Tabella 2, numeri ligandi, recettori e coppie LR per ogni database

2.3 I tre metodi:

I metodi utilizzati in questo confronto sono tre:

1. scSeqComm

2. NicheNet
3. CellPhoneDB

Come riassunto nella tabella 3, questi tool differiscono notevolmente nel modo in cui:

- le coppie ligando-recettore vengono trattate
- i livelli di espressione dei geni sono combinati
- tipo di output restituito

	Tipo di output	Tipo di dato in input	Proteine multi-subunità	Tipo di software
scSeqComm	Score numerico comunicazione inter e intra cellulare per ogni coppia LR	Matrice scRNAseq contente tutti i cluster	Si	Pacchetto R
NicheNet	Peso per ogni coppia LR	Matrice scRNAseq contente un cluster per volta	No	Pacchetto R
CellPhoneDB	P-value per ogni coppia LR	Matrice scRNAseq contente tutti i cluster	Si	Pacchetto Python

Tabella 3, riassunto tipo di output input e software dei 3 metodi

A ognuno dei metodi, per ogni database utilizzato, sono stati forniti i seguenti dati in input:

1. l'atlas completo

2. solo i campioni dei pazienti MSS
3. solo i campioni dei pazienti MSI-H

Queste operazioni di cambiamento, database e dati in input, sono state fatte per osservarne il comportamento in diverse situazioni, in modo tale da poter sviluppare una maggiore capacità di giudizio critico nella valutazione finale delle prestazioni di ogni metodo.

2.3.1 scSeqComm

Il primo metodo che vogliamo presentare è scSeqComm (Baruzzo , Cesaro , & Di Camillo, 2022), implementato in R, che permette di identificare e quantificare l'evidenza dell'attività intercellulare e intracellulare in corso fra due cluster di cellule, attraverso uno score.

La possibilità di quantificare l'evidenza della comunicazione in corso aiuta a prioritizzare ai risultati, mentre l'evidenza combinata della segnalazione sia intercellulare che intracellulare aumenta l'affidabilità della comunicazione dedotta.

Il calcolo dello score intercellulare scSeqComm richiede in input una matrice di espressione genica scRNA-seq normalizzata con indicazione sui cluster cellulari ai quali ogni cellula appartiene, e un database contenente le indicazioni sulle coppie LR. Successivamente, il primo passo è quello di andare a identificare e quantificare i segnali delle coppie LR in due passaggi chiave:

1. Si assegna uno score S a ogni ligando e ogni recettore, compreso tra zero e uno, espressi in uno specifico cluster.
2. Per ogni coppia LR conosciuta tra ogni coppia di cluster di cellule di interesse, viene dedotto uno score che identifica la comunicazione intercellulare in corso in funzione dello score del ligando e del recettore calcolati al punto 1.

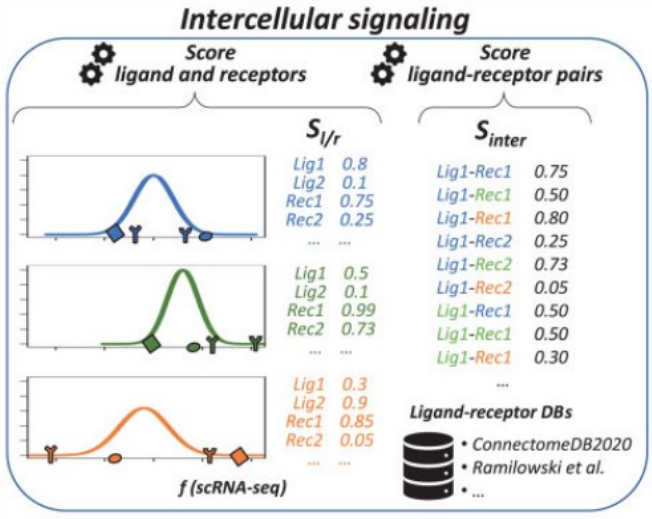


Figura 6, grafico riassuntivo dei due passaggi principali per il calcolo dello score intercellulare in scSeqComm, (Baruzzo , Cesaro , & Di Camillo, 2022)

Più in particolare, lo score S per il ligando e il recettore descritto al punto 1 sono volti a misurare quanto il livello medio di espressione del ligando/recettore osservato sia elevato rispetto ai livelli medi di espressione osservabili per caso all'interno dello stesso cluster. La distribuzione del livello medio di espressione genica osservabile casualmente è stata ottenuta, utilizzando un approccio di permutazione, come segue:

1. permutare casualmente righe/geni nella matrice indipendentemente per ciascuna colonna/cella;
2. calcolare i livelli medi di espressione genica in tale versione mescolata;
3. ripetere i passaggi 1 e 2 più volte;

Per il teorema del limite centrale, la distribuzione descritta sopra può essere approssimata da una distribuzione normale, anche se le variabili originali non sono distribuite normalmente.

Pertanto, il punteggio relativo al ligando o al recettore è stato calcolato come la probabilità di osservare valori inferiori al livello medio di espressione del ligando o del recettore quando si campionano valori da una distribuzione normale.

Tale formulazione tiene conto anche intrinsecamente della variabilità dell'espressione genica e del numero di cellule per cluster durante il calcolo dei valori dello score. Ad esempio, lo stesso livello di espressione medio del ligando /recettore sarà considerato meno affidabile (cioè, avrà un punteggio S inferiore) quando osservato in un cluster con poche cellule o con una grande varianza, rispetto alla sua osservazione in un cluster con molte cellule e meno rumoroso.

Per gestire la possibilità che un ligando (recettore) sia costituito da più subunità, il punteggio S è stato calcolato come media geometrica dei punteggi delle subunità. L'uso della media geometrica implementa un'operazione "AND" logica tra le subunità, il che implica che una subunità non attiva (cioè, con punteggio uguale a zero) porterebbe ad un punteggio S = 0 anche le altre subunità coinvolte.

Lo schema di punteggio descritto nel paragrafo precedente consente di caratterizzare in modo indipendente ligandi e recettori in diversi cluster di cellule. Tuttavia, la comunicazione cellulare può essere considerata attiva solo se sono attivi sia il ligando che il suo recettore specifico. È stato definito il punteggio intercellulare tra due cluster di cellule come il valore minimo fra lo score del ligando e quello del recettore, in modo tale che sia il segnale più debole quello che definisce l'intensità della comunicazione intercellulare in corso.

Nel secondo passo il tool va a quantificare la segnalazione intracellulare misurando l'evidenza di una risposta trascrizionale nei geni bersaglio regolati da fattori di trascrizione noti. L'associazione tra fattori di trascrizione e recettori a monte è ponderata in base alle conoscenze biologiche disponibili provenienti da database a priori.

Come prima cosa vengono convertiti i percorsi di segnalazione noti in grafi di geni e utilizzando la topologia dei grafi viene calcolato un punteggio di associazione a priori tra un recettore e un “transcription factor” (TF) (il TF è una proteina che si lega con specifiche sequenze di DNA e regola la trascrizione dei geni) utilizzando l'algoritmo Personalized PageRank (PPR). In secondo luogo, per ciascun cluster cellulare e grafo di segnalazione, viene valutata l'attività dei TF all'interno del

percorso misurando i cambiamenti nei livelli di espressione dei geni regolati. In terzo luogo, per ciascun insieme di recettori, cluster cellulari e grafi di segnalazione, vengono combinati i punteggi dei recettori-TF e i punteggi di attività dei TF in modo tale da calcolare quanto tale recettore sia coinvolto nella comunicazione in corso.

L'output finale (di nostro interesse in questo specifico progetto) di scSeqComm sono quindi due score: uno score relativo alla attività intercellulare per ogni coppia ligando-recettore fra due cluster di cellule di partenza, e l'altro score relativo a quella intracellulare nel cluster che esprime il recettore.

2.3.2 NicheNet

NicheNet (Browaeys, Saelens, & Saeys, 2019) è anch'esso implementato in un pacchetto R e richiede in input dei dati molto simili a quelli richiesti da scSeqComm, in particolare: una matrice di conteggio scRNAseq normalizzata, cluster di appartenenza delle singole cellule e database LR.

Per prima cosa NicheNet cerca di classificare i ligandi, in base alla loro attività, attraverso il loro livello di espressione, quantificandolo attraverso indici come auroc, aupr, coefficiente di correlazione di Pearson, i quali informano su quanto bene un ligando sia in grado di predire i cosiddetti geni bersaglio. Tali indici sono calcolati a partire dai dati di espressione delle cellule interagenti nel dataset definiti dagli sviluppatori "Ligand activity scores". Sono stati, cioè, definiti due set di geni: ligandi potenzialmente attivi nelle cellule mittenti e un set di geni di interesse nelle cellule riceventi. I "Ligand activity scores" sono stati quindi calcolati come coefficiente di correlazione di Pearson tra un vettore contenente i livelli di attività dei ligandi e dei corrispondenti geni target di ciascun ligando selezionato e il "target indicator vector", il quale è un vettore che indica se un gene appartiene o meno al set di geni di interesse.

Per non perdere l'appartenenza di ogni coppia LR ai suoi cluster di riferimento è necessario dare in input a NicheNet un cluster di cellule riceventi e mittenti per

volta, e solo in seguito andare a riunire gli output ottenuti, in quanto altrimenti NicheNet va a perdere questo dato fondamentale.

Viene assegnata uno score di priorità ai ligandi delle cellule mittenti, ovvero la probabilità di influenzare l'espressione genica nelle cellule riceventi con le quali interagiscono. Questo processo prende vita nella “ligand activities prediction” (Figure 8).

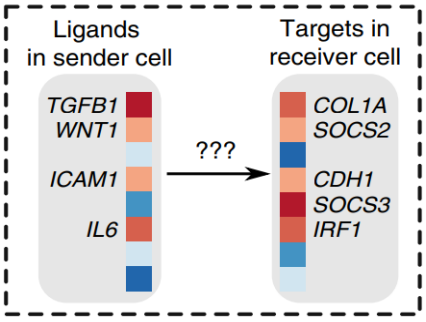


Figura 7, NicheNet dai ligandi nelle cellule mittenti ai target nelle cellule riceventi, (Browaeys, Saelens, & Saeys, 2019)

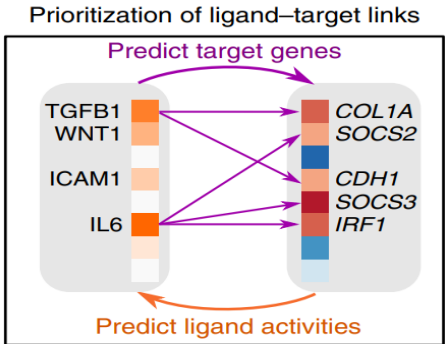


Figura 8, grafico della "ligand activities prediction", (Browaeys, Saelens, & Saeys, 2019)

NicheNet si concentra poi nell’identificare i geni bersaglio attivati dai ligandi nelle cellule mittenti cercando di comprendere al meglio quali siano i ligandi maggiormente coinvolti nella comunicazione cellula-cellula.

E solo infine ci è permesso di ottenere anche i recettori associati ai migliori ligandi.

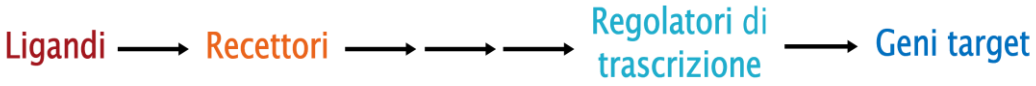


Figura 9, step chiave dei passaggi di NicheNet, per trovare i geni target.

L’output principale di NicheNet non sono quindi le coppie LR come per scSeqComm o CellPhoneDB ma piuttosto le potenziali interazioni fra ligando e geni bersaglio (Figura 7).

Solo, in seguito NicheNet fornisce anche la possibilità di costruire un elenco delle potenziali interazioni LR, grazie ai database forniti inizialmente; tuttavia, le coppie

LR e il loro relativo peso, che è un concetto assimilabile allo score di scSeqComm legato all'attività della coppia, non sono il vero scopo per il quale è stato creato questo metodo.

2.3.3 CellPhoneDB

CellPhoneDB (Efremova, Vento-Tormo, & Teichmann, 2019) si basa su un pacchetto Python a differenza dei primi due che, come abbiamo visto, sono basati su R.

Come per gli altri due metodi viene richiesto in input una matrice di espressione genica e i file relativi alle cellule e ai loro cluster di appartenenza. Come primo passo CellPhoneDB rimuove i ligandi/recettori espressi in meno del 10% delle cellule in ogni specifico cluster.

La specificità dell'interazione LR tra due cluster cellulari viene calcolata come la media fra l'espressione media del ligando e del recettore presi singolarmente. I punteggi di specificità LR statisticamente significativi vengono identificati tramite un test di permutazione andando a permutare casualmente le cellule nella matrice iniziale. Il test statistico di permutazione ("permutation test"), si può riassumere nei tre passaggi seguenti:

1. Divide i campioni in vari gruppi, determina e calcola ogni volta la statistica test, i componenti dei gruppi vengono poi mescolati e si ricalcola la statistica test, così per un numero di volte definito inizialmente;
2. Viene costruita una distribuzione statistica approssimata, in quanto è evidente come fare tale procedimento anche solo per un gruppo di poche unità possa rivelarsi computazionalmente molto dispendioso;
3. Si calcola il p-value;
4. Si osserva il valore del p-value e si traggono le considerazioni finali riguardo l'ipotesi nulla iniziale;

Innanzitutto, permutiamo casualmente le etichette dei cluster di tutte le cellule (1.000 volte sotto consiglio degli sviluppatori) e determiniamo la media del livello di espressione del ligando in un cluster e del livello di espressione del recettore nel cluster interagente. In questo modo generiamo una distribuzione nulla per ciascuna coppia ligando-recettore in ogni confronto tra due diverse cellule.

Otteniamo un valore P di probabilità relativo a ciascun complesso ligando-recettore calcolando la proporzione delle medie che sono pari o superiori alla media effettiva. Sulla base del numero di coppie significative, viene quindi data la possibilità all'utente di selezionare manualmente quelle biologicamente rilevanti. Per i complessi multi-subunità, viene invece richiesto che tutte le subunità del complesso siano espresse sopra una certa soglia, decisa dall'utente.

2.4 Elementi comuni e non comuni nei tre metodi

Per rendere il nostro confronto il più omogeneo possibile abbiamo cercato, dove consentito, di implementare i metodi con le medesime caratteristiche.

In input abbiamo utilizzato sempre come cluster di cellule mittenti quelle tumorali e come cluster di cellule riceventi quelle del sistema immunitario ottenendo in output solo le coppie LR attive fra questi gruppi di cellule. Abbiamo deciso di non effettuare le analisi con tutte le possibili combinazioni di cluster principalmente per tre motivi:

1. utilizzare una quantità di dati ragionevole, in quanto il dispendio a livello computazionale altrimenti sarebbe stato troppo alto;
2. per osservare il comportamento dei software in condizioni di lavoro più specifiche rispetto ad utilizzare tutti i possibili dati;
3. per dare un risvolto anche biologico e concentrarci su quelle interazioni proprie del microambiente tumorale e non tutte le possibili interazioni.

Infine, i confronti fra il numero di coppie LR ottenute in output sono stati fatti utilizzando gli stessi DB ligando recettore.

2.4.1 Implementazione scSeqComm

In scSeqComm la decisione da prendere di maggior rilevanza è stata quella riguardante lo score intercellulare, il quale è un numero compreso tra 0 e 1, restituito per ogni coppia ligando recettore dove il valore zero indica interazione minima, mentre il valore 1 è segnale di interazione massima.

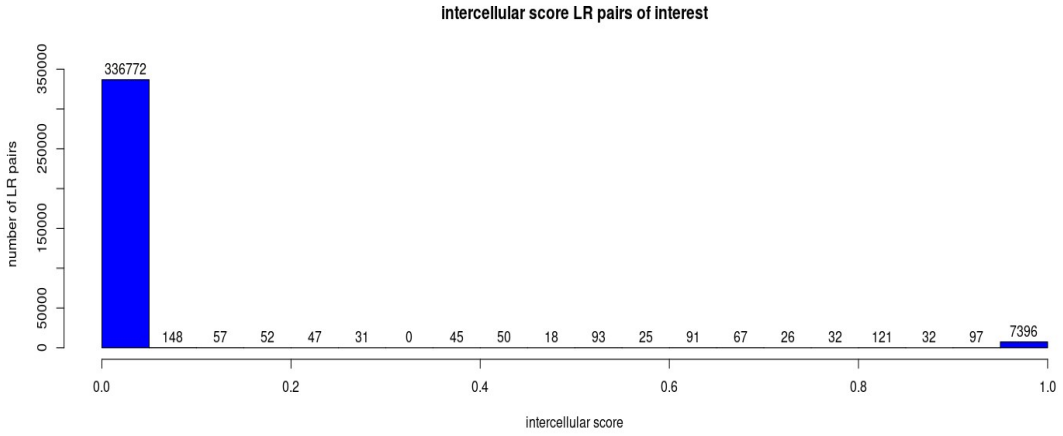


Figura 10, istogramma raffigurante la distribuzione degli score intercellulari con DB Efremova considerando in input tutto l'atlas

Vista la distribuzione degli score nell'istogramma (Figura 10) e sotto consiglio degli sviluppatori, abbiamo deciso di prendere in considerazione solo le coppie con score maggiore uguale a 0.8 in modo da selezionare le comunicazioni con le interazioni più alte. Inoltre, le coppie che avremmo perso prendendo una soglia più bassa, come per esempio 0.5 (default), non erano molte rispetto alle coppie individuate con soglia più alta.

2.4.2 Implementazione CellPhoneDB

Abbiamo innanzitutto creato il "metadafile" richiesto in cui le singole cellule sono associate ai relativi cluster, sotto poi consiglio degli sviluppatori abbiamo deciso di rimuovere i ligandi/recettori espressi in meno del 10% di cellule per ogni specifico cluster.

Anche CellPhoneDB come scSeqComm permette l'utilizzo di database multi-subunità LR, i quali tuttavia non sono stati presi in considerazione in questo progetto in quanto NicheNet non ne permetteva la possibilità.

Il vantaggio di tale procedura è che esistono test di permutazione per qualsiasi statistica di test, indipendentemente dal fatto che la sua distribuzione sia nota o meno. Si è quindi sempre liberi di scegliere la statistica che meglio discrimina tra ipotesi nulla e alternativa e che minimizza le perdite.

Dal file di output restituito (un file .txt) abbiamo tenuto solo le coppie ligando recettore che mostravano un valore di p-value minore o uguale a 0.05 e mantenuto solo le coppie che venivano annotate in un database a priori come veri ligandi.

2.4.3 Implementazione NicheNet

Il primo passo per NicheNet consiste nel creare la "ligand_target_matrix", la quale è specifica per ogni singolo database, creata grazie ad una funzione integrata fornita dagli sviluppatori.

Successivamente per trovare il set di geni target di interesse abbiamo:

- Considerato solo i geni espressi in almeno il 10% delle cellule per ogni specifico cluster (come precedentemente fatto anche per CellPhoneDB).
- Trovato i geni differenzialmente espressi nelle cellule immunitarie usando come condizione campioni di tessuto tumorali vs campioni di tessuto sani.

In seguito, abbiamo utilizzato l'indice di Pearson con soglia 0.15, per andare a classificare l'attività dei ligandi in modo da poter classificare quelli più attivi rispetto a quelli meno attivi.

In NicheNet è necessario anche, se si vuole mantenere l'appartenenza delle coppie LR ai cluster, eseguire il codice per ogni cluster di cellule mittenti o riceventi separatamente, e solo in seguito unire gli output, in quanto altrimenti NicheNet va a perdere l'appartenenza delle coppie LR al cluster.

2.5 Elementi osservati e misurati

I nostri sforzi si sono concentrati sul misurare:

1. Tipo di coppie LR ottenute in output;
2. Numero di coppie LR ottenute in output;
3. "Agreement" fra i diversi metodi, database e le coppie LR ottenute in output;
4. Tempistiche e richieste a livello computazionale;

Siamo andati ad osservare grazie a dei diagrammi di Eulero Venn il "grado di condivisione" dei risultati, più nel particolare il numero totale e la percentuale di coppie LR comuni e non nei diversi casi.

Abbiamo effettuato questo procedimento di confronto sia utilizzando lo stesso database LR per ogni gruppo (dataset completo, MSS e MSI-H) ma anche nel caso di DB LR diversi.

Avendo tipologie di output così diverse, come principale metro di paragone ci siamo voluti concentrare in particolare sul numero di coppie LR restituite.

Abbiamo poi confrontato i metodi in base alle richieste a livello computazionale e tempistico, per osservarne il comportamento al variare della quantità di dati forniti in input.

Infine, grazie ad una "Gene ontology" (GO) abbiamo voluto dare un ulteriore risvolto, all'analisi da noi effettuata, osservando il significato a livello biologico delle coppie comuni ottenute nelle diverse situazioni.

La GO è un'importante iniziativa bioinformatica volta a unificare la rappresentazione degli attributi dei geni e dei prodotti genici in tutte le specie. Mentre la nomenclatura si concentra sui geni, l'ontologia genetica si concentra sulla funzione dei geni e sui loro prodotti. La GO descrive la nostra conoscenza del dominio biologico rispetto a tre aspetti:

1. Funzione molecolare, che consiste nell'attività a livello molecolare svolte dai prodotti genetici. I termini delle funzioni molecolari nella GO rappresentano attività (come catalisi o trasporto) piuttosto che entità (molecole o complessi) e non specificano dove, quando o in quale contesto l'azione abbia luogo.
2. Componente cellulare, le posizioni relative alle strutture cellulari in cui un prodotto genetico svolge una funzione, siano essi compartimenti cellulari (come il mitocondrio) o complessi macromolecolari di cui fanno parte (ad esempio, il ribosoma).
3. Processo biologico, il quale può essere rappresentato da termini ampi relativi ai processi biologici come la riparazione del DNA o termini più specifici. Si noti che un processo biologico non è equivalente a un percorso.

3. Capitolo 3

Non disponendo di una “ground truth” e nemmeno di un “gold standard” per questo tipo di dati, per i quali altrimenti avremmo potuto effettuare un lavoro di benchmarking, che ci avrebbe permesso di fare una valutazione approfondita e completa delle capacità degli algoritmi utilizzando un set di dati di riferimento (ciò di cui siamo sprovvisti), siamo andati a valutare i metodi secondo:

- limitazioni dei metodi;
- tempo di esecuzione;
- facilità di utilizzo;
- numero e qualità delle interazioni fornite in output.

In questo capitolo, si riportano le considerazioni e i risultati relativi ai primi tre punti che sono principalmente legati ad aspetti tecnici e metodologici. L’analisi dettagliata dell’output dei diversi metodi in termini di coppie ligando-recettore individuate è trattata invece nel capitolo 4.

3.1 Principali limitazioni nei tre metodi

Come evidenziato anche da alcune recenti revisioni (Almet, Cang, Jin, & Nie , 2021) (Armingol, Officer, Harismendy, & Lewis, 2021) la validazione completa e sistematica di un metodo di comunicazione cellulare scRNA-seq non è possibile.

3.1.1 Limiti scSeqComm

Partendo da scSeqComm possiamo osservare come, sebbene, i risultati concordino ampiamente con la biologia conosciuta, siano ugualmente presenti alcune limitazioni: il punteggio restituito in output non è distribuito uniformemente tra 0 e 1 ma, quando i cluster di cellule nel dataset in input presentano un'ampia cardinalità o una variabilità dell'espressione genica molto bassa, il punteggio tende rispettivamente a 1 o a 0, poiché aumenta l'evidenza di una vera attivazione o inibizione. Tuttavia, anche in questi casi, il punteggio per i ligandi e i recettori presi singolarmente e il conseguente punteggio a livello intercellulare delle coppie LR

consentono di dare priorità e identificare i segnali intercellulari maggiormente rilevanti. Una seconda possibile limitazione riguarda invece l'utilizzo del minimo per calcolare il punteggio a livello intercellulare ed è data dal fatto che in tal modo viene data una debole priorità ai segnali forti. Tuttavia, l'uso contestuale del secondo punteggio proposto, ovvero il punteggio di segnalazione intracellulare, potrebbe aiutare nei momenti in cui la funzione min() perde la sua capacità di discriminazione.

Per quanto riguarda invece il punteggio della segnalazione intracellulare proposto, il principale limite è la semplice interpretabilità: sebbene il punteggio intracellulare si sia rivelato una misura affidabile dell'evidenza di una segnalazione intracellulare in corso e un modo efficace per classificare la segnalazione intracellulare dedotta, assegnare un significato chiaro ai valori di tale punteggio e identificare gli intervalli di punteggio rilevanti non è mai una questione banale. Come qualsiasi schema di punteggio che utilizzi caratteristiche espresse in modo differenziale, questo schema di punteggio può essere limitato nell'identificare tendenze biologiche rilevanti se condivise dalla grande maggioranza dei tipi di cellule. Infatti, l'identificazione dei fattori di trascrizione attivi in un certo cluster può essere distorta se la maggior parte dei geni bersaglio ha un livello di espressione molto simile nei cluster cellulari, ciò può verificarsi in set di dati scRNA-seq composti da popolazioni cellulari simili (ad esempio un set di dati composto solo da cellule immunitarie), mentre è più raro in set di dati che includano un'ampia varietà di tipi di cellule (ad esempio microambiente tumorale) lo stesso problema si ripresenta in CellPhoneDB e NicheNet.

3.1.2 Limiti CellPhoneDB

I limiti riguardanti CellPhoneDB sono dovuti al fatto che: la priorità è data solamente alle interazioni cellulari biologicamente importanti che sono specifiche della coppia di tipi cellulari comunicanti. Pertanto, un valore di p-value non significativo non indica il fatto che l'interazione non sia presente, ma solo che quella interazione non è altamente specifica tra quei due tipi di cellule. Inoltre, utilizzare

un test statistico di permutazione per generare un'ipotesi nulla, in un dataset delle dimensioni di 1 010 297 cellule, come si può osservare al paragrafo 3.2, richiede molto tempo e molte risorse a livello computazionale. Tuttavia, per risolvere questo problema, gli sviluppatori hanno introdotto un approccio di sotto campionamento che preserva l'eterogeneità del set di dati e riduce i requisiti di velocità e memoria.

Infine, proprio come scSeqComm questo schema di punteggio può essere limitato nell'identificare tendenze biologiche rilevanti se condivise dalla grande maggioranza dei tipi di cellule.

3.1.3 Limiti NicheNet

Trattando infine l'ultimo metodo, NicheNet, controllando l'espressione dei ligandi maggiormente attivi, possiamo osservare come alcuni di questi ligandi e/o i loro recettori siano davvero poco espressi, questo è dovuto al fatto che la definizione delle priorità dei ligandi da parte di NicheNet (analisi dell'attività dei ligandi nel programma) avviene sulla sola base dell'insieme di geni bersaglio che sono espressi in modo differenziale nella cellula ricevente, quindi non esiste alcuna priorità basata sulla forza di espressione del ligando nella cellula mittente o sulla forza di espressione del recettore nella cellula ricevente. L'espressione nelle cellule mittenti viene utilizzata solo per determinare quali ligandi siano espressi in una cellula mittente, mentre l'espressione nelle cellule riceventi viene utilizzata per determinare quali recettori siano espressi nella cellula ricevente.

L'identificazione dei fattori di trascrizione attivi in un certo cluster può essere distorta, proprio come in scSeqComm e CellPhoneDB, se la maggior parte dei geni bersaglio ha un livello di espressione molto simile nei cluster cellulari, ciò può verificarsi in set di dati scRNA-seq composti da popolazioni cellulari simili.

Nella versione attuale di NicheNet, la forza di espressione non è quindi inclusa direttamente perché si è trovato difficile formalizzare il compromesso tra attività del ligando ed espressione ma poiché il livello di espressione è importante, viene raccomandato di controllare l'espressione dei ligandi e dei loro recettori dopo la

definizione delle priorità di NicheNet. Per fare un esempio: se un ligando classificato 7° su 200 fosse espresso molto più forte di un ligando classificato 1° o 2° in base alla sua attività, questo ligando candidato, più fortemente espresso, potrebbe essere più interessante.

Un'altra limitazione importante di NicheNet è il fatto che non possa dire quali popolazioni cellulari interagiscono fra loro e quali no, a meno che non si provveda come nel nostro caso ad utilizzare il programma ogni volta inserendo manualmente i dati relativi ai cluster di interesse, in quanto altrimenti NicheNet restituisce una semplice lista di coppie LR considerate attive senza nessuna appartenenza ai cluster di riferimento.

Inoltre, NicheNet non restituisce tutte le interazioni significative ligando-recettore, in quanto viene eseguita un'analisi dell'attività del ligando per classificare i ligandi secondo la loro attività, e in base a ciò, i primi n (di default sono 20) ligandi vengono selezionati per ulteriori analisi e visualizzazioni.

Per quanto riguarda poi i valori di correlazione di Pearson (dai quali poi estrapoliamo gli n ligandi più importanti) bisogna tenere in considerazione che questi sono sensibili alla soglia scelta (0,1 in questo studio). I punteggi intorno allo zero (come 0,008) significano che i principali geni bersaglio previsti dai ligandi non sono di interesse per l'analisi, questi ligandi avranno tipicamente un AUROC intorno a 0,50. Ciò significa che NicheNet non trova prove che la cellula mittente stia regolando i cambiamenti dell'espressione genica nella cellula ricevente (sebbene l'assenza di prove non dimostri la mancanza).

3.2 Tempo di esecuzione

Per il progetto, grazie all'università di medicina di Innsbruck abbiamo potuto utilizzare l'HPCC ("high performance computing cluster") le cui caratteristiche software e hardware sono:

- Basato su Linux

- 1 x Head Node: zeus.icbi.local
  - 64 CPU cores / 3.0 TB RAM
  - 10 GBit Ceph storage network, 1 x 1 Gbit cluster network
  - 2 x 480 GB SSD RAID for system
  - 2 x 1.6 TB SSD RAID for local scratch, OS/Tools mirror, backup
- 10 x Compute Nodes: apollo-01 … apollo-10
  - 44 CPU cores, / 1.0 TB RAM
  - 2 x 10 Gbit ceph storage network, 1 x 1 Gbit cluster network
  - 2 x 480 GB SSD RAID for system
  - 2 x 800 GB SSD RAID for local scratch

Per scSeqComm e CellPhoneDB, dove era possibile sceglierlo sono stati utilizzati 20 core per velocizzare il programma (mentre questa scelta non era possibile in NicheNet). Come si può osservare dalle tabelle 4, 5 e 6, CellPhoneDB è il metodo che richiede maggior tempo con le caratteristiche di implementazione scelte, proprio a causa della lunga analisi statistica (test di permutazione) che effettua.

È importante tenere in considerazione osservando le tempistiche di utilizzo dei metodi, che scSeqComm e NicheNet vanno a valutare sia la comunicazione intracellulare che intercellulare, mentre CellPhoneDB solo quella intercellulare.

NicheNet risulta essere il metodo che richiede meno tempo in tutti e tre i casi, tuttavia, come osservato anche al paragrafo precedente, il numero 3.1.3, è anche quello con più limiti per quanto riguarda questo tipo di analisi.

Infine, abbiamo scSeqComm che con 20 core ha richiesto circa 3 ore per completare l'analisi su tutto l'atlas e per quello che vedremo anche nei capitoli successivi vista la facilità di utilizzo e i limiti relativamente contenuti che ha dimostrato si è rivelato essere il più adatto nel nostro tipo di lavoro.

Database = Efremova	scSeqComm	NicheNet	CellPhoneDB
Atlas intero	04:17:22	00:57:11	11:09:04
MSS	03:19:34	00:37:49	09:55:14
MSI - H	00:59:42	00:20:36	08:35:44

Tabella 4, tempi di utilizzo metodi con database Efremova

Database = Ramilowski	scSeqComm	NicheNet	CellPhoneDB
Atlas intero	04:45:28	01:10:18	13:19:14
MSS	03:04:56	01:04:35	11:56:44
MSI - H	00:43:12	00:54:54	10:35:34

Tabella 5, tempi di utilizzo metodi con database Ramilowski

Database = Kumar	scSeqComm	NicheNet	CellPhoneDB
Atlas intero	04:09:45	01:48:17	12:26:56
MSS	03:06:41	00:43:39	11:22:37
MSI - H	00:41:33	00:24:16	10:59:44

Tabella 6, tempi di utilizzo metodi con database Kumar

3.3 Facilità di utilizzo

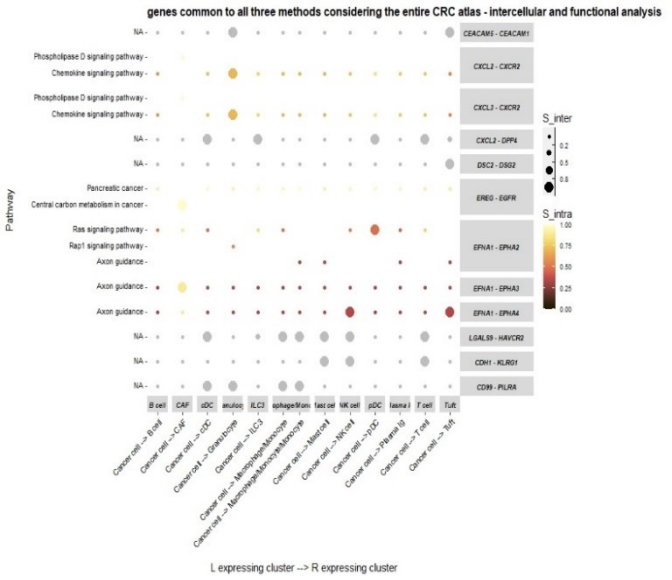
Considerando che NicheNet prevede che i dati forniti in input siano quanto più possibile costituiti da coppie LR che regolino l'insieme dei geni coinvolti (poiché

più questo insieme è costituito da geni che ci si aspetta siano regolati dal microambiente tumorale, meglio è) se questo set di geni fosse costituito anche da geni non necessariamente influenzati microambiente tumorale, NicheNet può comunque fornire buone previsioni, ma solo se una sostanziale parte di questi geni sono influenzati dalle interazioni cellula-cellula, pertanto, bisogna essere cauti col tipo di dati forniti in input.

Per quanto riguarda CellPhoneDB si è rivelato essere il metodo meno "user friendly" in quanto risulta avere: tempi di utilizzo molto lunghi con la versione 2.0.0 (è per ora impossibile utilizzare l'ultima versione 4.1.0 per problemi di sviluppo), l'approccio statistico proposto poi oltre ai tempi di utilizzo lunghi presenta, come descritto al paragrafo precedente, diversi limiti in output.

Infine, il metodo che ha rivelato avere maggior facilità di utilizzo, possibilità di implementazione, possibilità di analisi diverse, quantità di plot forniti di default in output e precisione biologica ci sentiamo di dire sia scSeqComm, in quanto nonostante in assenza di un certo tipo di potenza computazionale anche con tempi lunghi permette comunque di fare analisi semplici ma accurate integrando sia la comunicazione a livello intercellulare che intracellulare andando a colmare quelle lacune che un solo indice altrimenti potrebbe avere. Il pacchetto, inoltre, non si ferma poi al solo livello di attività cellulare ma fornisce già integrata al suo interno anche la possibilità di effettuare una prima analisi a livello biologico dei dati ottenuti in output grazie a delle funzioni che permettono di implementare la "Gene Ontology", la quale è già stata spiegata più nel dettaglio al paragrafo 2.3.

Ecco qui un esempio (Figura 11) di output finale restituito da scSeqComm raffigurante il punteggio intercellulare e intracellulare nelle diverse coppie LR, le quali sono elencate a destra con ogni cluster preso in analisi (in questo caso cellule tumorali come mittenti e cellule del sistema immunitario come riceventi) indicato in basso, i relativi "pathway" cellulari e la "gene ontology" con breve spiegazione dei termini a cui corrispondono i "GO_ID".



	GOJD	Term
1	GO:0002181	cytoplasmic translation
2	GO:0043066	negative regulation of apoptotic process
3	GO:0031663	lipopolysaccharide-mediated signaling pathway
4	GO:0009410	response to xenobiotic stimulus
5	GO:0042542	response to hydrogen peroxide
6	GO:0071356	cellular response to tumor necrosis factor
7	GO:0006954	inflammatory response
8	GO:0032355	response to estradiol
9	GO:0048661	positive regulation of smooth muscle cell proliferation
10	GO:0043123	positive regulation of I-kappaB kinase/NF-kappaB signaling
11	GO:0098586	cellular response to virus
12	GO:0045766	positive regulation of angiogenesis
13	GO:0042981	regulation of apoptotic process
14	GO:0031295	T cell costimulation
15	GO:0071347	cellular response to interleukin-1
16	GO:0043065	positive regulation of apoptotic process
17	GO:0032743	positive regulation of interleukin-2 production
18	GO:0071260	cellular response to mechanical stimulus
19	GO:0033209	tumor necrosis factor-mediated signaling pathway
20	GO:0071346	cellular response to interferon-gamma

Figura 11, output finale scSeqComm

4. Capitolo 4

In questo capitolo si analizzano gli output dei tre metodi e si vanno a confrontare le coppie ligando-recettore identificate come rilevanti per il dataset in esame. Inoltre, si riporta una preliminare e sintetica interpretazione biologica dei risultati prodotti dei metodi.

4.1 Coppie ligando recettore ottenute in numero e percentuale in relazione al database e al metodo utilizzato

In modo tale da visualizzare al meglio la sovrapposizione dei risultati in output fra i tre diversi metodi siamo andati a costruire dei diagrammi di Eulero Venn, i quali ci permettono con facilità di visualizzare numero e percentuale di coppie LR comuni e non nei diversi casi.

Confrontando le interazioni ligando-recettore prioritarie restituite in output di scSeqComm, NicheNet e CellPhoneDB troviamo "poca" sovrapposizione. Questo si può spiegare, come già ribadito più volte in precedenza, in quanto tutti e tre gli strumenti hanno in mente obiettivi diversi e quindi producono misure diverse.

Come prevedibile, vista l'analisi fatta ai paragrafi 3.1.1, 3.1.2 e 3.1.3, la sovrapposizione maggiore si ha fra scSeqComm e CellPhoneDB, è minore invece quella con NicheNet e in particolare fra NicheNet e CellPhoneDB (Figure 12, 13 e 14).

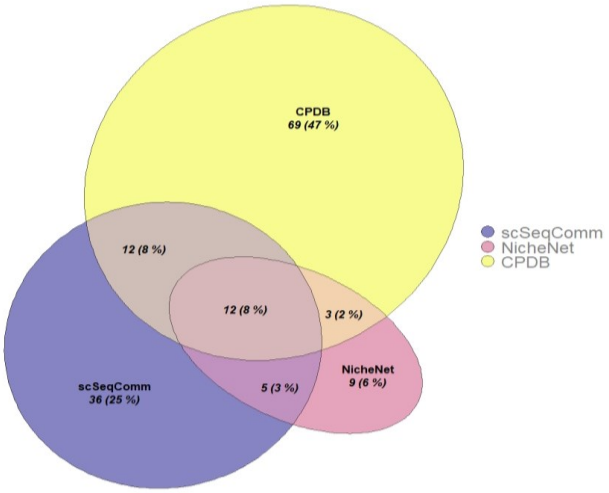


Figura 12, diagramma di Eulero Venn: sovrapposizione coppie LR usando tutto l'atlas con database Efremova

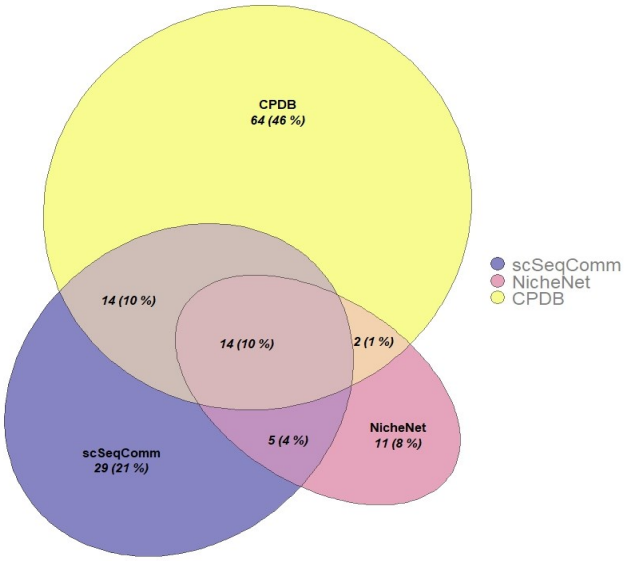


Figura 13, diagramma di Eulero Venn: sovrapposizione coppie LR usando campioni MSS con database Efremova

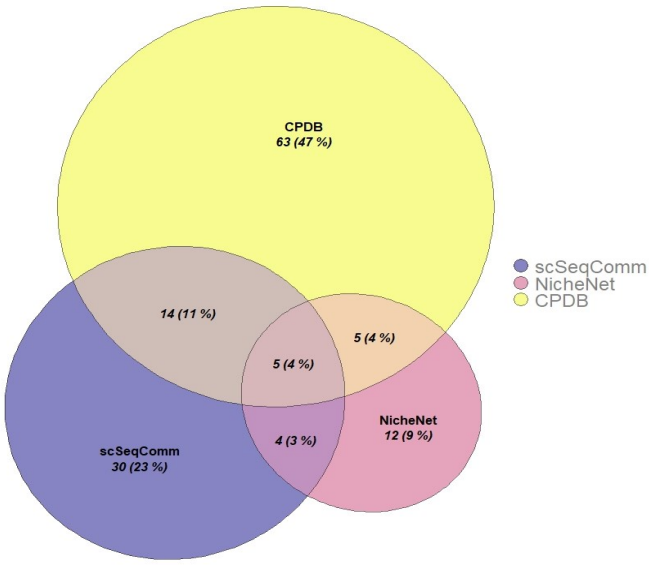


Figura 14, diagramma di Eulero Venn: sovrapposizione coppie LR usando campioni MSI - H con database Efremova

Minori sono anche le coppie restituite da NicheNet rispetto a CellPhoneDB e scSeqComm, proprio per il fatto che NicheNet mira prima di tutto a trovare i legami ligandi – geni target a differenza degli altri due.

Ad influenzare le analisi non sono nemmeno (in questo caso) il numero di cellule prese in considerazione in input in quanto nonostante le cellule passino da 1 010 297 per l'intero atlas a due casi più specifici come MSS e MSI – H, con rispettivamente 501 190 e 100 342 cellule, i numeri di coppie LR restituite in output non sono poi così diversi; quindi, ad influenzare gli output, se fornite in quantità adeguate con adeguata variabilità fra i tipi di cellule, non sono nemmeno quest'ultime.

Successivamente per meglio osservare i dati ottenuti per ogni caso con ogni singolo DB abbiamo costruito i seguenti diagrammi, i quali ci permettono di apprezzare il reale numero di interazioni restituite in output.

Si identifica subito come sia sempre CellPhoneDB quello in grado di trovare il maggior numero di coppie; come prevedibile poi il maggior numero di coppie restituite in output si ha per tutti e tre i metodi con il database Ramilowski, il quale è quello costituito dal maggior numero di coppie LR in partenza, segno che il

numero di interazioni restituito in output dipende fortemente dal database ligando – recettore scelto in input e dal numero di coppie LR dalle quali è composto (Figure 15, 16 e 17).

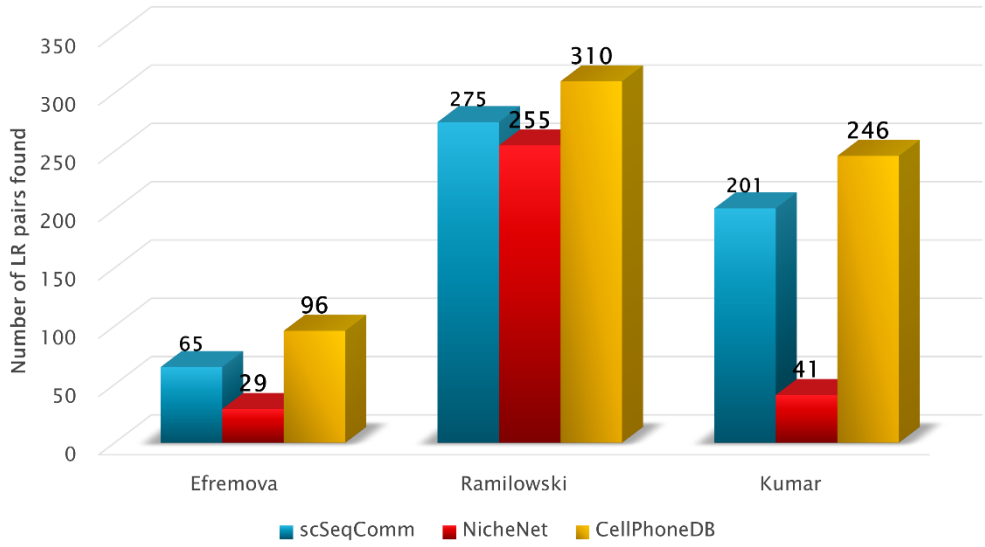


Figura 15, diagramma a colonne con rappresentante il numero di coppie LR in output da ogni metodo con i diversi database usando tutto l'atlas

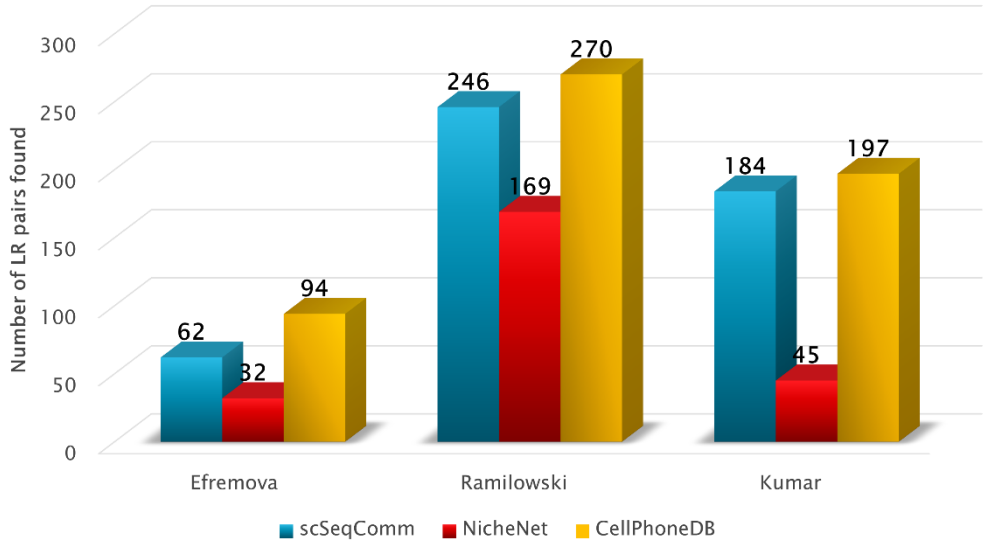


Figura 16, diagramma a colonne del numero di coppie LR in output da ogni metodo con i diversi database usando i campioni MSS

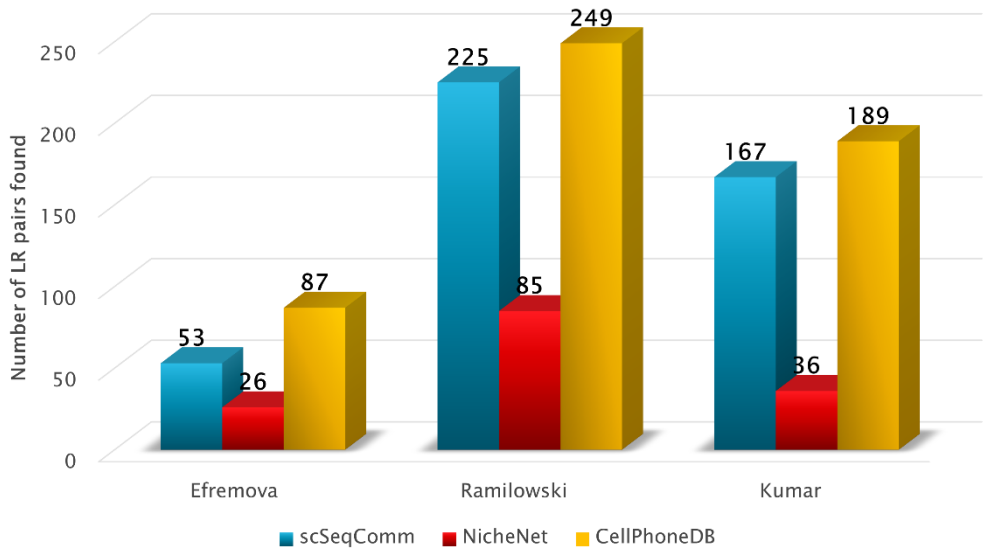


Figura 17, diagramma a colonne del numero di coppie LR in output da ogni metodo con i diversi database usando i campioni MSI – H

Si può infatti osservare una tendenza comune in tutti e tre i grafici, nei quali crescono le coppie LR restituite al crescere delle coppie LR fornite in input.

Ancora una volta in tutti i casi è NicheNet il metodo che restituisce meno interazioni LR ma di fatto questo era prevedibile per i limiti elencati precedentemente e per la soglia di n = 30 ligandi prioritari, classificati secondo l'indice di Pearson e da noi scelto.

Si rivela un buon compromesso fra i due ancora una volta scSeqComm, il quale ha in output coppie ligandi recettori che sono in accordo con la letteratura già esistente (Baruzzo , Cesaro , & Di Camillo, 2022) (Almet, Cang, Jin, & Nie , 2021)

4.2 Valutazione coppie ligando - recettore e gene ontology nel tumore al colon retto

Nella pratica clinica attuale, lo screening e la diagnosi del tumore al colon-retto si basano principalmente sul test del sangue occulto fecale (FOBT), sulla colonscopia e sulla rilevazione dell'antigene carcino - embrionario (CEA). Ciascun metodo presenta degli svantaggi: la sensibilità del FOBT e del CEA è limitata e, sebbene la colonscopia sia il "gold standard" per la diagnosi, la preparazione richiesta e le

occasionali complicanze gravi che si verificano ne limitano l'applicazione. Inoltre, la stratificazione clinica, il trattamento e la prognosi dipendono dalla localizzazione del tumore e dalla stadiazione di quest'ultimo; i risultati del trattamento variano e rimangono talvolta insoddisfacenti, suggerendo che questi indicatori non forniscono informazioni prognostiche ottimali (Bormann, Stinzing, & Tierling, 2018).

Vi sono prove sempre più evidenti che la patogenesi, la progressione, la risposta al trattamento e la prognosi del tumore al colon retto siano tutte significativamente influenzate da una complessa interazione tra le cellule tumorali e quelle del sistema immunitario all'interno del microambiente tumorale. Lo studio di nuove firme di espressione genica immuno-correlate potrebbe aiutare con il trattamento e la prognosi di tale malattia in futuro.

Per meglio osservare e visualizzare le coppie LR comuni ai tre metodi abbiamo costruito dei "circos plot" (Figure 18, 19 e 20) nei quali per tutti e tre i casi si è potuto osservare come, nonostante le differenze nel dato in input, le coppie LR siano quasi sempre le stesse.

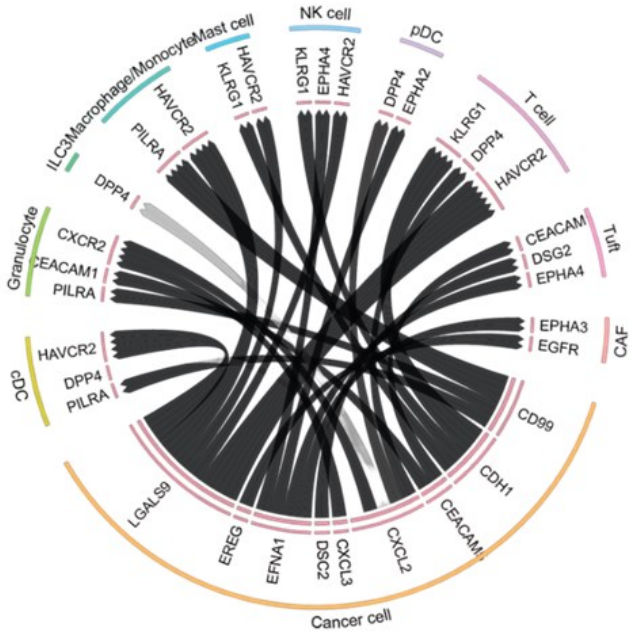


Figura 18, "circos plot" coppie LR comuni a tutti e tre i metodi con il database Efremova usando tutto l'atlas

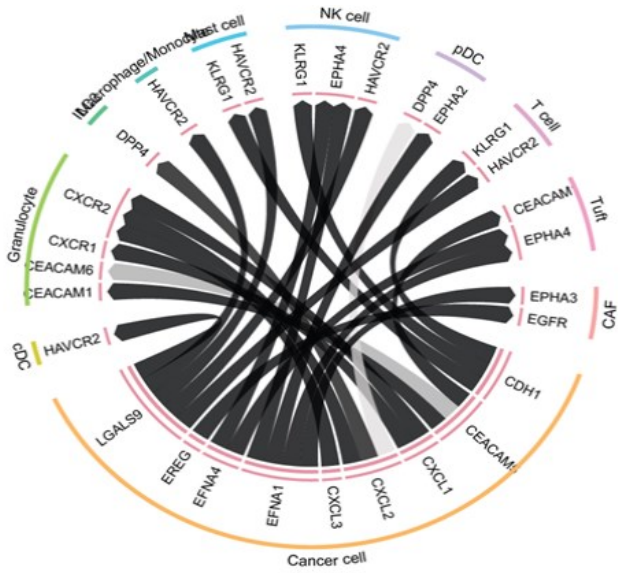


Figura 19, "circos plot" coppie LR comuni a tutti e tre i metodi con il database Efremova usando i campioni MSS

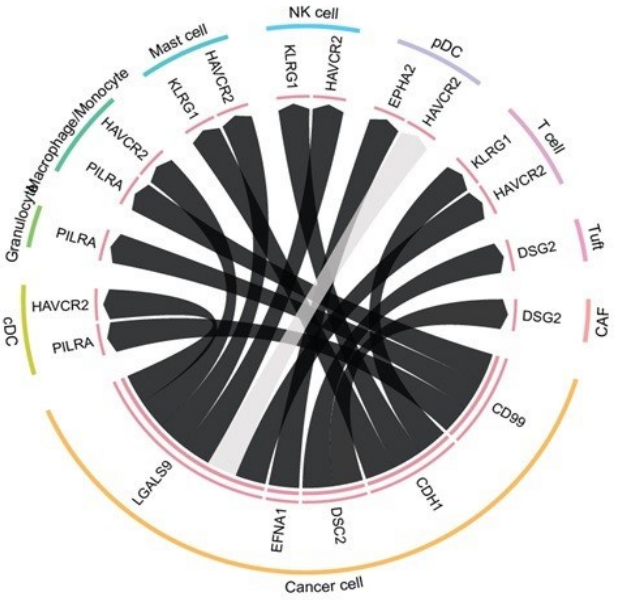


Figura 20, "circos plot" coppie LR comuni a tutti e tre i metodi con il database Efremova usando i campioni MSI - H

Osservando i grafici ottenuti notiamo come gran parte delle coppie siano già state confermate in letteratura (Beauchamp & W, 2011) (QH, YZ, Tu , & Liu , 2020) (Thomas , Klebanov , John , Miller , & Vegesna, 2019) (Ager , Zhang, & Chaimowitz, 2023) (Can, Rui , Feng, Pei, & Sun, 2022), ed è proprio su queste che possibili ricerche future dovrebbero concentrarsi.

Per meglio capire di cosa stiamo parlando, partiamo per esempio dalla coppia EREG – EGFR, della quale si è visto in particolare come: l'espressione dei ligandi del fattore di crescita epidermico epiregulina (EREG) siano positivamente correlati con una risposta immunitaria efficace nel cancro del colon-retto. I ligandi del fattore di crescita epidermico, EREG, si sono perciò rivelati biomarcatori predittivi utili nel tumore al colon retto, tuttavia la regolazione della loro espressione genica in questo tipo di tumore è poco conosciuta e andrebbe approfondita.

I tumori del colon-retto hanno mostrato livelli di espressione di EPHA2 significativamente più elevati rispetto al tessuto normale corrispondente (Beauchamp & W, 2011): EPHA2 è uno dei recettori presenti in ognuno dei casi da noi osservato. È un membro della famiglia delle tirosin – chinasi, le quali sono enzimi che regolano processi cellulari, quali la proliferazione e la differenziazione cellulare, e possono essere coinvolti nel processo di oncogenesi, poiché catalizzano la fosforilazione dei residui di tirosina di alcuni recettori cellulari, tra cui appunto EGFR; tuttavia, il suo ruolo nella progressione del cancro del colon-retto non è ancora del tutto chiaro.

Altre ricerche hanno poi evidenziato EPHA2 come un marcatore prognostico sfavorevole nel tumore al colo-retto in stadio II/III, il che potrebbe essere dovuto alla sua capacità di promuovere la migrazione e l'invasione cellulare, ciò potrebbe fornire un punto di partenza per ulteriori indagini su EPHA2 come nuovo biomarcatore prognostico e bersaglio terapeutico (Bormann, Stinzing, & Tierling, 2018).

È stato riportato poi anche che EPHA2 si lega all'EGFR nelle linee cellulari tumorali umane, rendendo perciò di fatto il complesso “EPHA2, EGFR”

clinicamente rilevante: infatti l'espressione di EGFR e EPHA2, come nel nostro caso è sovra-regolata nei tumori al colon-retto. (QH, YZ, Tu , & Liu , 2020).

I CEACAM sono molecole di segnalazione intracellulare e intercellulare, con diverse funzioni, dalla differenziazione e trasformazione cellulare alla modulazione delle risposte immunitarie associate a infezioni o infiammazioni nel cancro. Le conoscenze attuali su CEACAM1, CEACAM5 e CEACAM6, evidenziano il loro significato patologico nelle aree della biologia del cancro, dell'immunologia e delle malattie infiammatorie (Thomas , Klebanov , John , Miller , & Vegesna, 2019).

È stato osservato come nella modulazione immunitaria, CEACAM1 sia fortemente sovraregolato nelle cellule T in seguito all'attivazione da parte delle citochine. Sulla base della sua cinetica nelle cellule T e del ruolo inibitorio nelle cellule B, CEACAM1 probabilmente svolge un ruolo inibitorio nella funzione immunitaria delle cellule T. La sua segnalazione è complessa e dipende dal tipo e dallo stadio dei tessuti coinvolti, con ruoli di soppressione della crescita in alcuni tipi di tessuti ma anche ruoli proliferativi e stimolativi in altri tipi di tessuti (Thomas , Klebanov , John , Miller , & Vegesna, 2019).

I pazienti affetti da tumore al colon retto i quali presentano un CEACAM1 e CEACAM6 elevato in combinazione con una bassa espressione di EPHA2 hanno beneficiato di un tempo più lungo alla prima recidiva/metastasi rispetto a quelli con un'elevata espressione di EPHA2.

Si capisce quindi come il ruolo di CEACAM1 sia molto importante nel microambiente tumorale, ma debba essere considerato in presenza di altri membri della famiglia CEACAM (Thomas , Klebanov , John , Miller , & Vegesna, 2019).

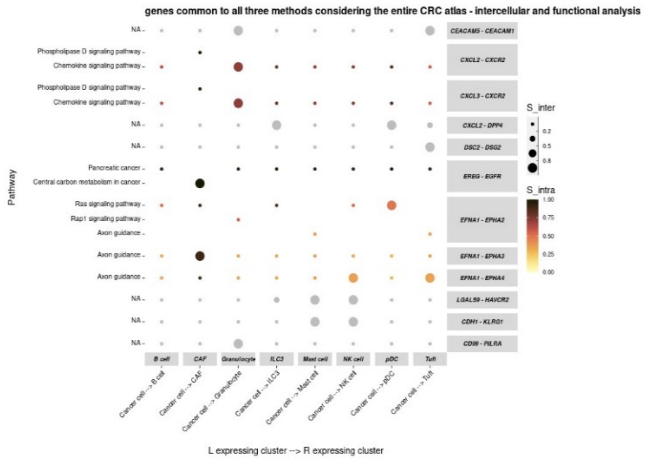
KLRG1 è stato identificato in modo univoco nelle popolazioni di cellule T intra-tumorali, studi di validazione hanno confermato che le firme KLRG1 nelle cellule T infiltranti il tumore umano sono associate alla progressione della malattia dimostrando l'utilità della scoperta di KLRG1 come biomarcatori dinamici (Ager , Zhang, & Chaimowitz, 2023).

Le chemochine CXC appartengono a una famiglia unica di citochine chemiotattiche che influenzano l'inizio, la progressione e l'esito clinico di molti tipi di tumore. È

stato osservato come ci sia una associazione del ligando CXCL3 con la progressione del tumore e una prognosi sfavorevole della malattia (Can, Rui , Feng, Pei, & Sun, 2022); inoltre, è stato scoperto che i livelli di CXCL3 distinguono il tumore al colon-retto da altri tipi di tumori gastrointestinali.

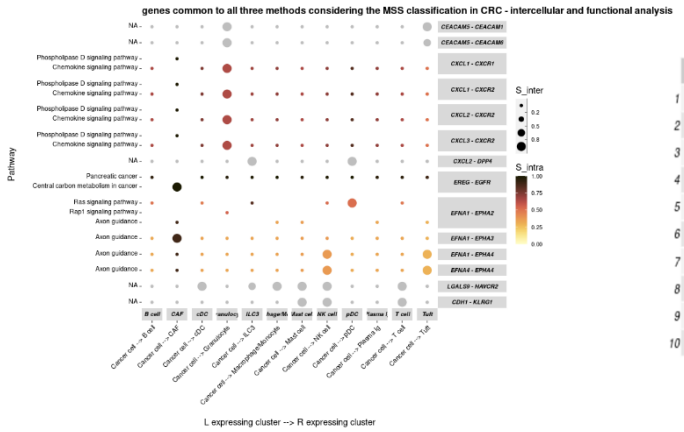
Anche DSG2 è stato identificato come possibile biomarcatore prognostico per i pazienti affetti da cancro al colon-retto, tuttavia, ancora troppi pochi studi hanno valutato l'espressione di DSG2 e la sua funzione in questo tipo di tumore (Tingting , Xuan , Lizhou , & Jiaojiao , 2021).

Successivamente per i tre casi siamo andati a eseguire, grazie al pacchetto di funzioni già implementato in scSeqComm, la "GO" (Figure 21, 22 e 23) dei geni comuni trovati nei tre casi dai tre metodi con Efremova come DB LR, osservando come tali coppie LR fossero tutte coinvolti nella reazione di risposta immunitaria nel microambiente tumorale segno che i metodi identificano coppie ligando recettore ragionevolmente corrette.



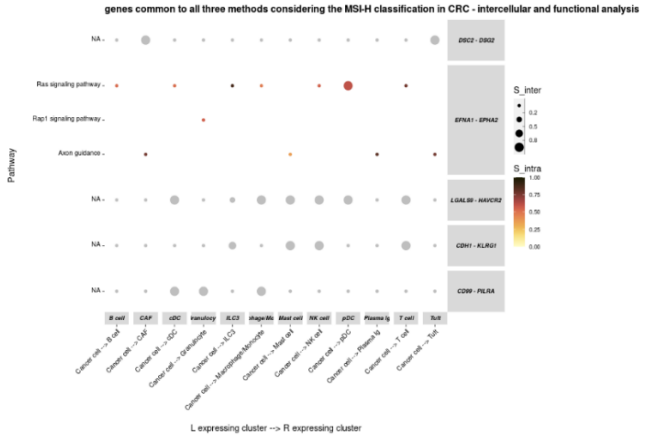
	GO.ID	Term
1	GO:0002181	cytoplasmic translation
2	GO:0043066	negative regulation of apoptotic process
3	GO:0009410	response to xenobiotic stimulus
4	GO:0032355	response to estradiol
5	GO:0071356	cellular response to tumor necrosis factor
6	GO:0098586	cellular response to virus
7	GO:0042542	response to hydrogen peroxide
8	GO:0007568	aging
9	GO:0044344	cellular response to fibroblast growth factor stimulus
10	GO:0045766	positive regulation of angiogenesis

Figura 21, "Gene Ontology" coppie LR comuni ai tre metodi utilizzando database Efremova usando tutto l'atlas



	GO.ID	Term
1	GO:0002181	cytoplasmic translation
2	GO:0031663	lipopolysaccharide-mediated signaling pathway
3	GO:0043066	negative regulation of apoptotic process
4	GO:0009410	response to xenobiotic stimulus
5	GO:0043123	positive regulation of I-kappaB kinase/NF-kappaB signaling
6	GO:0048661	positive regulation of smooth muscle cell proliferation
7	GO:0098586	cellular response to virus
8	GO:0006954	inflammatory response
9	GO:0009612	response to mechanical stimulus
10	GO:0071356	cellular response to tumor necrosis factor

Figura 22, "Gene Ontology" coppie LR comuni ai tre metodi utilizzando database Efremova usando campioni MSS



	GO.ID	Term
1	GO:0006954	inflammatory response
2	GO:0031663	lipopolysaccharide-mediated signaling pathway
3	GO:0045944	positive regulation of transcription by RNA polymerase II
4	GO:0032731	positive regulation of interleukin-1 beta production
5	GO:0009410	response to xenobiotic stimulus
6	GO:0032722	positive regulation of chemokine production
7	GO:0071222	cellular response to lipopolysaccharide
8	GO:0043123	positive regulation of I-kappaB kinase/NF-kappaB signaling
9	GO:1904646	cellular response to amyloid-beta
10	GO:0070374	positive regulation of ERK1 and ERK2 cascade
11	GO:0031295	T cell costimulation
12	GO:0032743	positive regulation of interleukin-2 production
13	GO:0030890	positive regulation of B cell proliferation
14	GO:0048661	positive regulation of smooth muscle cell proliferation
15	GO:0032760	positive regulation of tumor necrosis factor production
16	GO:1902895	positive regulation of miRNA transcription
17	GO:0071347	cellular response to interleukin-1
18	GO:0032755	positive regulation of interleukin-6 production
19	GO:0050729	positive regulation of inflammatory response
20	GO:0045766	positive regulation of angiogenesis

Figura 23, "Gene Ontology" coppie LR comuni ai tre metodi utilizzando database Efremova usando campioni MSI - H

In futuro potrebbe essere utile e necessario studiare la correlazione tra i livelli di MSI e MSS delle cellule tumorali e i profili di espressione genica di tutte le altre cellule nei campioni MSS e MSI e non solo le relazioni fra le cellule tumorali e quelle del sistema immunitario.

5. Capitolo 5

5.1 Conclusioni

Ricapitolando i vari temi affrontati possiamo osservare come nonostante alcune differenze fra i metodi, tutti e tre cerchino di misurare l'evidenza di una comunicazione intercellulare in corso tra due cluster cellulari attraverso la definizione di un "punteggio di interazione". Il punteggio viene calcolato in funzione dei livelli di espressione genica, nei gruppi di cellule mittenti e riceventi, del ligando e del recettore. Tuttavia, i metodi differiscono principalmente per due motivi: nel modo in cui viene calcolato il punteggio di interazione e nel modo in cui i livelli di espressione e il caratteristico punteggio risultante associato alla coppia ligando-recettore vengono combinati fra loro.

Abbiamo poi visto come il principale vantaggio nel calcolare un valore di attività per una coppia LR sia quello di poter riordinare le coppie ligando recettore secondo una classifica, per risolvere l'importante problema del dare una priorità alla comunicazione cellulare da convalidare poi a livello biologico.

I diversi modi di utilizzare i dati di espressione del ligando-recettore per calcolare il punteggio hanno anche un effetto diretto su ciò che il punteggio misura intrinsecamente.

L'obiettivo di scSeqComm è la progettazione del punteggio S in cui i ligandi/recettori rilevanti sono identificati principalmente sulla base del loro livello di espressione genica e il numero di coppie restituite in output si basa sulla scelta dell'utente di una certa soglia considerata ragionevole. Le soglie solitamente non sono specifiche del cluster, i loro valori dipendono dalla scala dei dati e dal metodo di normalizzazione del conteggio utilizzato e generalmente non esiste un modo per definirle in modo affidabile.

In ogni caso, scSeqComm consente agli utenti di selezionare quali geni includere/omettere dalla procedura di permutazione, consentendo così di mitigare alcuni dei possibili problemi con un approccio come quello descritto fino ad ora.

L'obiettivo di CellPhoneDB è trovare coppie ligando-recettore specifiche per il tipo di cellula; quindi, CellPhoneDB è l'ideale se si vuole studiare il potenziale di interazione differenziale di diversi tipi di cellule. CellPhoneDB cerca innanzitutto le coppie ligando-recettore espresse tra le cellule interagenti, osservando l'espressione sia del ligando che del recettore: più ligandi/recettori forti e specifici sono espressi nelle popolazioni di cellule mittenti/riceventi, migliore sarà la priorità tramite un'analisi di CellPhoneDB.

L'obiettivo di NicheNet è complementare a quello di CellPhoneDB o scSeqComm: NicheNet mira a trovare le coppie ligando-recettore che hanno maggiori probabilità di regolare l'espressione genica nella cellula ricevente. Quindi cerca coppie ligando-recettore per le quali nei dati esistano prove di interazione: questo potrebbe fornire più informazioni funzionali e anche alcuni indizi su quali interazioni potrebbero essere realmente attive in quanto l'espressione del ligando e del recettore a livello di RNA (come riscontrato da CellPhoneDB), non significa necessariamente una interazione nella realtà. Quindi, proprio come CellPhoneDB, NicheNet inizia a cercare tutte le coppie ligando - gene target e poi le coppie ligando - recettore espresse tra le cellule interagenti di interesse ma invece di dare priorità a questi osservando la forza di espressione, NicheNet darà loro la priorità in base ai geni bersaglio osservati nella cellula ricevente, come già evidenziato al capitolo 3 paragrafo 3.1. Un possibile svantaggio di questo approccio è che possono essere restituite alcune coppie ligando-recettore con un'espressione relativamente bassa, pertanto, consigliamo anche di controllare l'espressione dei ligandi e dei loro recettori dopo l'operazione di "prioritizzazione" di NicheNet.

Pertanto, quando si confronta l'output di NicheNet con quelli di CellPhoneDB e scSeqComm, ci si può aspettare che siano presenti le seguenti differenze: le coppie di recettori ligandi rilevate da NicheNet ma non da CellPhoneDB o scSeqComm potrebbero essere espresse in generale o espresse in modo piuttosto basso, ma hanno alcune "prove di segnalazione". Le coppie selezionate da CellPhoneDB e scSeqComm ma non da NicheNet saranno invece espresse in modo forte e specifico per il tipo di cellula, ma non vi è alcuna prova basata sulle conoscenze pregresse

che queste coppie siano effettivamente funzionali, (naturalmente la mancanza di prove non significa che non ci siano segnali in corso).

In conclusione, l'affidabilità dei diversi approcci dipende fortemente dalle caratteristiche dei dati in input e da come questi siano utilizzati per calcolare l'output.

5.2 Possibili sviluppi e applicazioni allo studio del "tumor microenviroment"

Il TME (tumor microenvironment) rappresenta una rete intricata comprendente diversi tipi di cellule impegnate nella comunicazione, la quale gioca un ruolo fondamentale nel modellare il panorama tumorale. Una comunicazione efficace tra le cellule immunitarie e le cellule tumorali riveste un grande significato nella determinazione delle risposte dei pazienti all'immunoterapia e contribuisce alla resistenza al trattamento e alla variabilità inter-paziente nelle risposte alle cure. Lo studio della comunicazione cellula-cellula nelle cellule immunitarie in condizioni normali e patologiche fornisce informazioni cruciali sui meccanismi dell'immunoterapia antitumorale, sulle risposte dei pazienti, sulla progressione della malattia e sullo stato del TME. Esistono vari approcci sperimentali e computazionali per chiarire le interazioni intercellulari patologiche, sia direttamente che indirettamente, con l'obiettivo di identificare le popolazioni cellulari comunicanti. La comprensione completa, la modellizzazione e la scoperta delle interazioni cellula-cellula all'interno del TME hanno un immenso potenziale per l'identificazione dei fattori critici e delle strategie che influenzano la risposta immuno-terapica, la resistenza al trattamento e lo stato del TME.

In conclusione, la comunicazione cellula-cellula è un processo criptico che controlla i segnali e la capacità delle cellule di mantenere l'omeostasi nel microambiente tumorale. La comunicazione intercellulare comprende diversi fattori e il sequenziamento dell'RNA a singola cellula fornisce trascrittomi in grado di definire l'eterogeneità cellulare, nuovi sottotipi di cluster cellulari, e di identificare i profili di interazione ligando-recettore. Sebbene la comunicazione cellula - cellula sia stata studiata per decenni, le nuove scoperte tecnologiche hanno

portato a una nuova comprensione dei suoi meccanismi che offrono opportunità per l'identificazione e lo sviluppo di biomarcatori e bersagli terapeutici specifici della malattia.

Possibili nuove domande per ricerche future riguardo l'immunoterapia antitumorale utilizzando questi metodi per la ricerca delle interazioni cellula-cellula potrebbero essere:

- Quali sono le sottopopolazioni di cellule immunitarie che interagiscono con la cellula tumorale durante una specifica immunoterapia?
- Quali sono le caratteristiche molecolari di queste cellule immunitarie interagenti e come sono correlate le caratteristiche molecolari in risposta all'immunoterapia?
- Le cellule che hanno interagito con le cellule tumorali migrano attraverso le lesioni metastatiche del tumore?

In tempi recenti, i progressi nell'ingegneria cellulare e nella bioinformatica sono emersi come potenti strumenti per svelare nuovi meccanismi e relazioni cellulari, aprendo così la strada a migliori opzioni di immunoterapia per i pazienti. Sfruttando queste metodologie in modo sinergico, diventa possibile colmare le lacune delle conoscenze esistenti, acquisire una comprensione completa della resistenza al trattamento e progettare immunoterapie contro il cancro più potenti.

In definitiva, le risposte a queste domande possono svelare le azioni farmacologiche dell'immunoterapia antitumorale e rivelare il meccanismo molecolare sottostante della resistenza a determinate cure.

Riferimenti

Ma, F., Zhang, S., Song, L., Wang, B., & Wei, L. (2021, Luglio 3). Applications and analytical tools of cell communication based on ligand-receptor interactions at single cell level. Tratto da https://cellandbioscience.biomedcentral.com/articles/10.1186/s13578-021-00635-z

Shao , X., Liao, J., Li, C., & Lu, X. (2020, Novembre 4). CellTalkDB: a manually curated database of ligand–receptor interactions in humans and mice . Tratto da https://doi.org/10.1093/bib/bbaa269

Ager , C., Zhang, M., & Chaimowitz, M. (2023). KLRG1 marks tumor-infiltrating CD4 T cell subsets associated with tumor progression and immunotherapy response. doi:10.1136/jitc-2023-006782

AIRC. (2021, Ottobre 1). AIRC. Tratto da Tumore del colon-retto: https://www.airc.it/cancro/informazioni-tumori/guida-ai-tumori/tumore-colon-retto#:~:text=In%20Italia%20le%20stime%20pi%C3%B9,Italia%20in%20entrambi%20i%20sessi.

Almet, A., Cang, Z., Jin, S., & Nie , Q. (2021, Marzo 26). The landscape of cell–cell communication through single-cell transcriptomics. doi:https://doi.org/10.1016/j.coisb.2021.03.007

Armingol, E., Officer, A., Harismendy, O., & Lewis, N. (2021, Febbraio). Deciphering cell-cell interactions and communication from gene expression. doi:10.1038/s41576-020-00292-x

Baruzzo , G., Cesaro , G., & Di Camillo, B. (2022, Marzo 28). Identify, quantify and characterize cellular communication from single-cell RNA sequencing data with scSeqComm. doi:10.1093/bioinformatics/btac036

Beauchamp , A., & W, D. (2011, Ottobre 25). Ephs and ephrins in cancer: ephrin-A1 signalling. doi:10.1016/j.semcdb.2011.10.019

Bormann, F., Stinzing, S., & Tierling, S. (2018, Settembre 25). Epigenetic regulation of Amphiregulin and Epiregulin in colorectal cancer. doi:https://doi.org/10.1002/ijc.31892

Browaeys, R., Saelens, W., & Saeys, Y. (2019, Marzo 5). doi:https://doi.org/10.1038/s41592-019-0667-5

Can, C., Rui , Z., Feng, G., Pei, Y., & Sun, L. (2022, Maggio 26). Plasma CXCL3 Levels Are Associated with Tumor Progression and an Unfavorable Colorectal Cancer Prognosis. Tratto da https://www.hindawi.com/journals/jir/2022/1336509/

Efremova, M., Vento-Tormo, M., & Teichmann, S. (2019, Giugno 24). CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. doi:https://doi.org/10.1038/s41596-020-0292-x

Haque, A. E. (2017, Agosto 18). Genome Medicine. Tratto da A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications: https://doi.org/10.1186/s13073-017-0467-4

Innsbruck, C. (s.d.). Tratto da https://atlas-protocol.readthedocs.io/en/latest/

Jeeho, K., In-Youb, C., & Ho Jin, Y. (2022, Giugno 6). Interactions between EGFR and EphA2 promote tumorigenesis through the action of Ephexin1. Tratto da https://www.nature.com/articles/s41419-022-04984-6

JR, J. (2007, Gennaio). National Library of Medicine. doi:10.1111/j.1365-2559.2006.02549.x

Lopez-Garcia, L., Demiray, L., & Ru, S. (2018). Validation of extracellular ligand–receptor interactions by Flow-TriCEPS. Tratto da https://doi.org/10.1186/s13104-018-3974-5

Müller MF, I. A. (2016). Molecular pathological classification of colorectal cancer. Tratto da https://pubmed.ncbi.nlm.nih.gov/27325016/

Perroteau, I. (s.d.). Comunicazione cellulare, Trasduzione del segnale. Tratto da Università di Torino : chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://biologia.i-learn.unito.it/pluginfile.php/4892/mod_resource/content/0/lezioni-10/18_comunicazione_apoptosi.pdf

QH, L., YZ, W., Tu , J., & Liu , C. (2020, Giugno 23). Anti-EGFR therapy in metastatic colorectal cancer: mechanisms and potential regimens of drug resistance. Gastroenterol Rep. doi:10.1093/gastro/goaa026

Rossi, N. (2010). Treccani. Tratto da Dizionario di Medicina Treccani: https://www.treccani.it/enciclopedia/gene_%28Dizionario-di-Medicina%29/

Salzberg, S. (2018, Agosto 20). How many genes do we have? doi:https://doi.org/10.1186/s12915-018-0564-x

Thomas , J., Klebanov , A., John , S., Miller , L., & Vegesna, A. (2019, Febbraio 1). CEACAMS 1, 5, and 6 in disease and cancer: interactions with pathogens. doi:10.18632/genesandcancer.230

Tingting , Y., Xuan , G., Lizhou , J., & Jiaojiao , G. (2021). DSG2 expression is low in colon cancer and correlates with poor survival. Tratto da https://bmcgastroenterol.biomedcentral.com/articles/10.1186/s12876-020-01588-2

Toledo, C., & Saltsman, K. (2012, Giugno 12). Genetics by the Numbers. Tratto da National Institute of General Medical Sciences: https://www.nigms.nih.gov/education/Inside-Life-Science/Pages/Genetics-by-the-Numbers.aspx#:~:text=20%2C000&text=That's%20the%20approximate%20number%20of,about%20as%20many%20human%20genes.