UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA BIOMEDICA

"CONFRONTO DI METODI BIOINFORMATICI PER L'INFERENZA DELLA COMUNICAZIONE CELLULARE DA DATI DI TRASCRITTOMICA A SINGOLA CELLULA "

Relatore: Prof. Giacomo Baruzzo
Correlatore: Dott.ssa Giulia Cesaro

Laureando: Francesco Vendramin

ANNO ACCADEMICO 2022 – 2023
Data di laurea: 29 settembre 2023

Sommario

Abstract

La comunicazione cellulare ha un ruolo chiave nel funzionamento dell'organismo ed è quindi di fondamentale importanza riuscire a comprendere i meccanismi che ne stanno alla base; uno dei principali tipi di comunicazione è quello di tipo chimico, la quale sfrutta l'interazione tra due particolari tipologie di molecole, il ligando e il recettore.

I moderni metodi bioinformatici sono in grado, sfruttando le tecniche di sequenziamento dell'RNA a singola cellula, di fornire, attraverso l'utilizzo di score (o punteggi) intercellulari, una stima dell'evidenza della comunicazione intercellulare in corso tra due cellule, basandosi principalmente sui livelli di espressione dei geni corrispondenti ai relativi ligandi e recettori.

Dato che sono presenti diverse tipologie di score intercellulari, le quali prendono in considerazione differenti aspetti biologici della comunicazione cellulare, non esiste uno score unico che sia in grado di analizzare perfettamente ciascuno di questi aspetti; è utile, pertanto, avere a disposizione un set di punteggi da poter porre a confronto e tra cui scegliere in ogni situazione specifica. Per questo motivo sono stati selezionati cinque score da tre diversi articoli con lo scopo di integrarli all'interno di un tool già esistente, ovvero *scSeqComm*; tali score sono stati implementati in R e sono poi stati testati e posti a confronto sfruttando il set di dati *Tirosh* disponibile online.

Capitolo 1: Comunicazione cellulare

1.1 Cellule e metodi di comunicazione

Gli organismi complessi sono costituiti da un insieme di cellule che comunicano continuamente tra di loro, armonizzando la loro attività in modo da garantire il corretto svolgimento di tutte le funzioni dell'organismo stesso. Tale comunicazione ricopre infatti un ruolo di fondamentale importanza nel controllo e nel coordinamento delle attività delle cellule che lo costituiscono, come lo sviluppo e la funzionalizzazione di tessuti ed organi, azioni che sarebbero impossibili in assenza di un qualche tipo di interazione cellulare che permetta scambi di informazioni tra le divere parti.

La comunicazione si verifica grazie alla elaborazione e trasmissione di messaggi che raggiungono cellule bersaglio in grado di riconoscerli. I messaggi trasmessi sono caratterizzati dal contenuto dell'informazione, dalla destinazione e dalla velocità di trasmissione.

I meccanismi attraverso i quali avviene l'interazione cellulare sono molteplici, ma possono essere in generale ricondotti a due modalità principali, ovvero la comunicazione elettrica e la comunicazione chimica; esse prendono il nome dalla tipologia di messaggi che trasmettono, ovvero rispettivamente messaggi di tipo elettrico e messaggi di tipo chimico.

La comunicazione elettrica è solitamente caratterizzata da messaggi stereotipati, come ad esempio potenziali d'azione ("tutto o nulla") o potenziali elettrotonici, e da un'elevata velocità di trasmissione; affinché essa avvenga è però necessaria la presenza di linee di trasmissione, come ad esempio gli assoni nel sistema nervoso.[1]

La comunicazione chimica invece avviene per diffusione o flusso di massa, non richiedendo quindi alcuna linea di trasmissione dato che le molecole adibite al trasferimento del messaggio sono altamente specifiche e devono percorrere in generale brevi tratti prima di raggiungere la cellula bersaglio; questo tipo di trasmissione, se da un lato permette di comunicare una grande varietà di informazioni differenti, dall'altro comporta una riduzione nella velocità di trasmissione del messaggio [1].

1.1.1 Comunicazione chimica

La tipologia di comunicazione di interesse per le successive analisi è quella di tipo chimico; essa prevede il trasferimento del segnale mediante l'interazione tra due molecole: il messaggero e il recettore, dove la cellula che prevede l'emissione del messaggero è detta 'cellula sorgente', mentre la cellula che prevede l'espressione del recettore viene denominata 'cellula bersaglio'.
Le modalità di comunicazione chimica si possono classificare in base alla distanza d'azione in:

- Autocrina: il recettore si trova sulla stessa cellula che produce il messaggero;
- Paracrina: il recettore si trova su cellule adiacenti a quelle che producono il messaggero;
- Endocrina: il recettore è lontano dalla 'cellula sorgente' [1].

1.1.2 Comunicazione intercellulare ed intracellulare

La catena di comunicazioni chimiche che permette il trasferimento di un'informazione da una cellula ad un'altra può a sua volta essere suddivisa in due fasi.
La prima fase viene detta comunicazione intercellulare e prevede l'interazione tra una molecola detta ligando (o 'primo messaggero'), emessa dalla cellula sorgente, ed un'altra molecola, detta recettore, che si trova solitamente sulla membrana cellulare della cellula bersaglio, dando luogo ad un fenomeno chiamato trasduzione del segnale, ovvero un'alterazione chimica del recettore che innesca una reazione a catena all'interno della cellula bersaglio, attivando una via biochimica al suo interno e creando così una risposta cellulare [1].
La fase che segue è detta comunicazione intracellulare e prevede la formazione di un mediatore cellulare (o 'secondo messaggero') in grado di attivare la risposta cellulare specifica, la quale causa in genere una formazione a cascata di molecole segnale all'interno della cellula bersaglio con lo scopo di amplificare e trasmettere l'informazione; l'intero processo può avere ad esempio come conseguenza l'alterazione di metabolismo cellulare, forma, espressione genica e abilità di dividersi [1].

1.1.3 Tipologie di ligandi

I ligandi, data la loro peculiare funzione, sono una classe di molecole molto varia e che non presenta una classificazione univoca che permetta di catalogarli; è tuttavia possibile effettuare una distinzione in diverse tipologie secondo tre aspetti che contraddistinguono ogni ligando: il modo in cui agiscono, il punto in cui agiscono e la loro natura chimica.
In base al loro modo di agire essi possono infatti essere classificati in:

- ormoni, che vengono rilasciati da una cellula e agiscono su cellule bersaglio lontane;
- neurotrasmettitori, che vengono rilasciati da un neurone e agiscono su cellule bersaglio vicine;
- citochine, che vengono rilasciate da cellule del sistema immunitario e agiscono su altre cellule del sistema immunitario;
- fattori di crescita, che vengono rilasciati da cellule in crescita e agiscono su altre cellule in crescita;
- molecole segnale extracellulari, che vengono rilasciate da cellule di vari tessuti e agiscono su cellule di altri tessuti.

In base al loro modo di agire invece vengono distinti in ligandi che agiscono:

- all'esterno della cellula, legandosi a recettori sulla membrana plasmatica;
- all'interno della cellula, penetrando nella cellula e legandosi a recettori intracellulari.

Infine, in base alla loro natura chimica, i ligandi possono essere:

- proteine, come gli ormoni, le citochine e i fattori di crescita;
- peptidi, come i neurotrasmettitori;
- lipidi, come gli ormoni steroidei;
- nucleotidi, come il cAMP;
- gas, come l'ossido nitrico [1].

1.1.4 Tipologie di recettori

Le principali categorie di recettori associate al processo di comunicazione chimica intercellulare sono:

- *recettori di tipo 1 o collegati a canali ionici o ionotropici*: in questa tipologia di recettori il ligando si lega al recettore, che è presente sulla membrana e va a modificare l'effettore, ovvero in questo caso il canale ionico; l'accoppiamento risulta diretto,

ovvero non necessita di un mediatore che trasformi il segnale da extracellulare ad intracellulare e il tempo di d'azione per ottenere una risposta è rapidissimo.

Una volta che il ligando si è legato al recettore, il canale ionico posto in prossimità si apre portando ad un flusso di ioni, che ha come conseguenza una depolarizzazione ('membrana eccitata') o un'iperpolarizzazione ('membrana inibita').

- *recettori di tipo 2 o accoppiati a proteine G o metabotropici*: questa tipologia di recettori è quella maggiormente presente all'interno del nostro organismo, ma anche una delle più complicate; questi recettori necessitano infatti di un intermediario per la trasduzione del segnale, ovvero la proteina G.

  Quando il ligando si lega al recettore, quest'ultimo attiva la proteina G, la quale a sua volta attiverà un canale ionico o un enzima; nel primo caso i processi che seguono sono i medesimi dei recettori di tipo 1, mentre nel caso in cui venga attivato un enzima si produrranno dei secondi messaggeri, quali nucleotidi ciclici o calcio intracellulare, che andranno a generare una serie di effetti cellulari; tali secondi messaggeri vanno ad innescare delle reazioni all'interno della cellula.

  La presenza di un intermediario rallenta il tempo d'azione da quasi istantaneo a qualche secondo, ma permette di operare anche un'inibizione dei canali ionici o dell'enzima, oltre alla loro attivazione.

- *recettori di tipo 3 o accoppiati a tirosinchinasi*: sono recettori di membrana accoppiati a delle chinasi, che generano risposte cellulare in genere dipendenti da fosforilazioni proteiche.

  Il recettore, una volta entrato in contatto con il ligando, attiva una chinasi che catalizza delle reazioni, le quali hanno come conseguenza la formazione di una serie di fosforilazioni proteiche, portando ad una modificazione dei geni a livello del DNA; dato che il bersaglio è proprio la trascrizione genica il tempo d'azione è molto lungo, ovvero di ore o addirittura giorni.

- *recettori di tipo 4 o citoplasmatici*: a differenza dei precedenti, questi recettori sono intracellulari o citoplasmatici e sono spesso utilizzati dagli ormoni steroidei.

  Dato che il meccanismo va a modificare l'espressione genica, esso richiede molto tempo per vedere delle risposte cellulari; devono infatti essere prodotte le proteine indotte dalla modificazione genica apportata dalla sostanza introdotta nella cellula.

  Un esempio di tale meccanismo è il seguente: l'ormone che si trova all'esterno della cellula abbandona la proteina che lo sta trasportando e si trasforma in una sostanza molto lipofila, riuscendo quindi ad attraversare la membrana ed entrare all'interno della

cellula; una volta all'interno la sostanza si lega ad una proteina di trasporto e raggiunge il nucleo, dove andrà ad espletare la sua attività di modificazione della trascrizione genica.

La risposta cellulare è costituita dalla produzione di un mRNA che andrà a sintetizzare delle proteine diverse.

Le 4 tipologie di recettori sopra citate sono riassunte secondo una serie di parametri nella tabella 1.1.

Recettore	**Localizzazione**	**Effettore**	**Accoppiamento**	**Tempo d'azione**
TIPO 1	Membrana	Canale ionico	Diretto	Millisecondi
TIPO 2	Membrana	Enzima o canale	Proteina G	Secondi
TIPO 3	Membrana	----------	Chinasi	Ore / giorni
TIPO 4	Intracellulare	----------	A vari recettori e a proteine di trasporto	Giorni / settimane

Tabella 1.1 Tabella riassuntiva tipologie di recettori

1.2 Stato dell'arte dell'analisi della comunicazione cellulare

Lo scopo principale delle analisi e dei metodi bioinformatici che sono ideati in questi ultimi anni è quello di capire con la minore incertezza possibile se è in corso la comunicazione cellulare tra due cellule dell'organismo; in particolare si vuole verificare che sia avvenuto lo scambio di una precisa informazione, la quale può essere trasferita esclusivamente tramite una o più coppie ligando – recettore, ben note nella letteratura scientifica.

L'idea alla base delle tecniche ad oggi implementate si basa sulla correlazione esistente tra l'espressione proteica e quella genica; la presenza di un determinato filamento di mRNA è infatti proporzionale e funge da indicatore dell'abbondanza di una specifica proteina; è pertanto possibile sfruttare l'enorme progresso fatto nel campo del sequenziamento dell'RNA a cellula singola (scRNA – seq), per verificare la presenza dei filamenti di mRNA corrispondenti ai relativi ligandi e recettori, in modo da ottenere indirettamente il livello di espressione delle singole proteine in ogni cellula di interesse.

Il sequenziamento a singola cellula è una tecnica nata per ovviare al principale problema dei metodi precedentemente utilizzati, ovvero quello di non essere in grado di tenere conto dell'eterogeneità all'interno dei gruppi cellulari, presente anche tra cellule molto simili tra loro; tali metodi, infatti, si basano sull'analisi della trascrittomica prodotta da un insieme di cellule, i cui valori vengono poi mediati e utilizzati come valori di riferimento di quel dato gruppo cellulare, considerato omogeneo. Questa semplificazione ha come conseguenza una perdita di informazione riguardo le specifiche cellule, le quali, in particolari contesti, come ad esempio in presenza di alcune patologie o di fenomeni di resistenza ai farmaci, potrebbero avere comportamenti anomali rispetto al gruppo di appartenenza. Il metodo di sequenziamento a singola cellula permette di rilevare queste anomalie e di utilizzare le informazioni così ottenute per avere una visione più dettagliata delle differenze nell'espressione genica tra le cellule di un dato cluster e per comprendere meglio da un lato la biologia e il funzionamento delle cellule individuali, dall'altro i complessi processi biologici che stanno dietro a questo tipo di comportamento [2].

Nonostante questa nuova tecnica di sequenziamento abbia portato notevoli vantaggi, come la capacità di rilevazione dell'eterogeneità cellulare e della presenza di popolazioni rare, che le permettono, in combinazione anche con altre tecniche come la genomica e la proteomica, di avere numerose applicazioni in ambito medico, essa presenta anche alcuni svantaggi e limitazioni; tra questi sicuramente sono da evidenziare i problemi relativi alle lunghe tempistiche, ai costi elevati e alla variabilità sperimentale che rende difficile la riproducibilità dei risultati; inoltre nell'utilizzo di questa tecnica vengono trascurate due dimensioni molto importanti, ovvero quella spaziale e quella temporale; sarebbe infatti da tenere in considerazione da un lato il fatto che la vicinanza o lontananza tra le cellule ha un ruolo chiave nella comunicazione intercellulare, dall'altro che potrebbe risultare significativo in alcuni casi, come ad esempio durante la differenziazione cellulare o una risposta immunitaria, analizzare come l'interazione tra le cellule evolve nel tempo [3].

L’intero processo atto ad ottenere i dati relativi all’espressione genica nelle singole cellule è costituito da diversi passaggi: la prima fase è quella di isolamento delle singole cellule, necessaria per isolare le singole cellule dal tessuto di interesse; tale isolamento si può ottenere in diversi modi, ognuno dei quali ha i propri pregi e i propri difetti, sia in termini di efficacia che di rapidità della tecnica; alcuni esempi sono la diluizione limitata, la micromanipolazione, la selezione delle cellule attivate dal flusso (FACS) e la microdissezione guidata dal laser [2]. Una volta completato questo primo passaggio segue la preparazione delle librerie di RNA, ovvero un procedimento in cui viene estratto l’RNA, per poi convertirlo in cDNA (DNA complementare) e amplificarlo in modo da ottenere abbastanza materiale genetico da sequenziare. Una volta ultimata questa fase preparatoria è possibile procedere con il sequenziamento vero e proprio, durante il quale vengono lette le sequenze di nucleotidi che compongono il cDNA di ciascuna cellula, in modo da ottenere informazioni dettagliate sulle sequenze geniche attive in ognuna di esse. Infine, i dati di sequenziamento così ottenuti vengono sottoposti ad analisi attraverso l’utilizzo di strumenti bioinformatici [2].
In generale i dati ottenuti con il procedimento appena presentato possono essere raccolti in vari formati e visualizzazioni a seconda dell’obiettivo dell’analisi, ma il formato più comune utilizzato in questo ambito è quello della matrice di espressione genica. Essa è una tabella in cui ogni riga rappresenta un gene, ogni colonna rappresenta una cellula e i valori nella tabella rappresentano l'intensità dell'espressione genica per ciascun gene in ciascuna cellula; tale matrice può essere utilizzata per identificare i geni differentemente espressi tra le cellule e per eseguire analisi di clustering per identificare gruppi di cellule simili.

1.2.1 Attuali metodi bioinformatici

A partire dai dati ricavati tramite il sequenziamento di RNA a singola cellula esistono una grande varietà di metodi che hanno lo scopo di ottenere informazioni inerenti alle comunicazioni intercellulari, la maggior parte dei quali si basa solitamente non sui singoli livelli di espressione, bensì su livelli medi di espressione; questi sono calcolati suddividendo le cellule in gruppi cellulari e calcolando per ognuno di essi la media dei livelli di espressione di ligandi e recettori.
In genere questi metodi bioinformatici tendono ad assegnare ad ogni coppia ligando – recettore in due cluster cellulari un cosiddetto ‘punteggio (o score) intercellulare’, ovvero una misura dell'evidenza di una comunicazione cellulare in corso, basata sul livello di espressione medio del ligando e del recettore così calcolati.

Tuttavia, non esiste un metodo univoco per calcolare tale punteggio intercellulare, bensì i vari metodi proposti differiscono notevolmente sotto diversi aspetti, quali:

- modo in cui i livelli di espressione medi di ligandi e recettori sono combinati;
- caratteristiche del punteggio risultante, il quale può essere espresso ad esempio come valore binario (0 = interazione non avvenuta, 1 = comunicazione avvenuta) oppure come un valore continuo (solitamente normalizzato tra 0 e 1);
- database ligando – recettore utilizzato; i database disponibili, infatti, differiscono ampiamente nel numero di coppie ligando – recettore riportate e nel livello di convalida sperimentale; in aggiunta recentemente sono stati introdotti database che includono anche la struttura multi-subunità di ligandi e recettori, ma solo pochi schemi di punteggio intercellulare sono in grado di sfruttare queste informazioni aggiuntive.

L'aspetto che tra i tre appena elencati sicuramente incide di più nel differenziare i vari punteggi intercellulari è sicuramente il primo; infatti, il modo in cui tali livelli di espressione sono combinati va ad incidere sia sul range in cui saranno definiti i valori, sia sul tipo di dipendenza (es. lineare, esponenziale, logaritmica) che lo score avrà da essi.

Oltre a sfruttare i dati forniti dai geni che codificano per ligandi e recettori, alcuni score tengono in considerazione altri fattori che concorrono o sono influenzati dall'interazione cellulare che si sta analizzando; alcuni esempi sono molecole agoniste o antagoniste che influenzano la comunicazione tra le cellule oppure semplicemente il numero di cellule presenti all'interno di un dato gruppo cellulare, il quale in combinazione con il livello medio di espressione può essere un indice di comunicazione cellulare in corso.

Oltre a quanto appena citato, alcuni metodi, quali ad esempio NicheNet [4], scMLnet [5], CytoTalk [6] e scSeqComm [7], sfruttano il fatto che una volta avvenuta l'interazione tra ligando e recettore, questa generi una reazione a catena all'interno della cellula coinvolgendo una serie di altre molecole; tali metodi tentano di dedurre quindi non solo la segnalazione intercellulare, bensì l'intero processo di comunicazione cellulare, prendendo quindi in esame anche la segnalazione intracellulare.

Per quanto riguarda la differenza tra punteggi continui e binari, i primi sono sicuramente maggiormente informativi rispetto alle differenze tra le varie coppie e permettono di avere una maggiore comprensione dell'attendibilità della previsione; se infatti ad esempio un punteggio continuo presenta un range che va da 0 a 1, sarà sicuramente più probabile che una coppia ligando – recettore con un valore di 0.9 abbia effettivamente interagito rispetto ad una con punteggio di 0.6, ma un punteggio binario le avrebbe ugualmente indicate come attive, ovvero con il simbolo 1.

Il punteggio binario però presenta un notevole vantaggio, ovvero è di facile comprensione e non ha bisogno di interpretazione, cosa che gli permette di essere fruito anche da non esperti del campo; tale vantaggio è talmente significativo che la maggior parte dei punteggi continui prevede un livello di soglia, detto anche threshold, che permette di trasformare tale score nella propria versione binaria, attribuendo valore 0 alle coppie sotto tale livello e valore 1 a quelle al di sopra di esso.

1.2.2 Stato attuale

Attualmente esistono quindi un gran numero di metodi atti a studiare ed analizzare la comunicazione cellulare basandosi sul sequenziamento dell'RNA a singola cellula, ma, nonostante ciò, questa rimane un'area di ricerca ancora piuttosto giovane e in rapida evoluzione. Come infatti viene posto in evidenza in alcuni studi [3,8], l'utilizzo che viene fatto attualmente dei dati scRNA-seq è principalmente quello di generatore di ipotesi, le quali necessitano però di ulteriori validazioni sperimentali; al fine di svolgere tale compito è necessario però che i punteggi intercellulari siano in grado di identificare un quantità ragionevole di obiettivi robusti, fatto da cui deriva la necessità di avere score aventi da un lato un comportamento di tipo conservativo, ovvero in grado di fornire il minor numero possibile di falsi positivi anche a discapito della riduzione degli obiettivi segnalati, dall'altro di avere una semplice interpretabilità.
Un altro aspetto che è importante sottolineare è come non esiste e non esisterà mai uno score considerabile il migliore in senso assoluto, poiché ogni score modella e misura diversi aspetti molecolari che indirettamente supportano l'evidenza di una comunicazione cellulare in atto, e valorizza aspetti della comunicazione diversi (esempio: specificità vs intensità della comunicazione).
Da queste osservazioni si evince che avere a disposizione un unico score, per quanto efficace esso sia nello svolgere il compito per il quale è stato concepito, non permette di avere uno strumento completo e versatile per effettuare una ampia e variegata serie di analisi in ambito biologico, rendendo quindi necessario l'utilizzo simultaneo di un maggior numero di punteggi intercellulari o l'uso di punteggi intercellulari diversi per analisi diverse.

1.2.3 Obiettivo della tesi

Al fine di ovviare, almeno in parte, al problema della limitatezza applicativa che contraddistingue tutti gli score intercellulari, può essere di grande aiuto la realizzazione di un programma che sia in grado di fornire all'utente che lo utilizza non solo un numero rilevante di score che abbiano strutture diverse e che mirino a descrivere diversi aspetti biologici, ma anche avere una conoscenza delle differenze tra i risultati ottenibili con i diversi score intercellulari.
Per realizzare questo obiettivo è stato scelto come punto di partenza un tool bioinformatico [7] in cui è già stato implementato uno score di interazione cellulare, ovvero software chiamato scSeqComm. scSeqComm accetta in input un dataset scRNA-seq e un database ligandi-recettori e restituisce in output una tabella contenente sulle righe le diverse coppie ligando – recettore per ogni possibile combinazione di cluster di cellule e sulle colonne diverse informazioni relative ad esse, compreso lo score intercellulare *scSeqComm*.
A partire da scSeqComm, si sono estese le sue funzionalità andando ad implementare diversi punteggi intercellulari presi dalla letteratura, e arricchendo l'output di scSeqComm con tali punteggi.
Infine, si è condotta una analisi comparativi dei diversi punteggi su un dataset scRNA-seq reale, al fine di studiare le differenze nei risultati ottenuti con i diversi score.

Capitolo 2: Score di interazione intercellulare

2.1 ScSeqComm

Come detto nel capitolo precedente, è stato scelto come punto di riferimento per tutte le analisi effettuate riguardanti la validità e l'efficacia dei diversi score intercellulari, proposti nei vari articoli presi in considerazione, il metodo computazionale *scSeqComm* [7].
Tale metodo va ad implementare, sfruttando R come linguaggio di programmazione, un codice che è in grado, a partire da un dataset scRNA-seq, di calcolare l'evidenza dell'interazione tra una particolare coppia ligando – recettore.

2.1.1 Score del ligando e score del recettore

Per andare ad ottenere una stima dell'interazione cellulare tra diversi gruppi di cellule, sono stati seguiti due passaggi distinti: in primo luogo è stato assegnato un punteggio a ciascun ligando e a ciascun recettore espressi in uno specifico gruppo di cellule; successivamente, per ogni coppia ligando – recettore nota, sia tra due gruppi di cellule che all'interno dello stesso gruppo, è stato dedotto un punteggio intercellulare, espressione della comunicazione intercellulare in corso in funzione del punteggio del ligando e del punteggio del recettore.
Nella prima fase viene calcolato un punteggio S, funzione sia del singolo ligando (recettore), sia del gruppo cellulare preso in considerazione, indicato quindi con il simbolo S(g,k), dove g rappresenta lo specifico gene, mentre k lo specifico cluster; tale punteggio fornisce dei risultati in un range continuo di valori che va da 0 a 1 ed è mirato a misurare quanto il livello di espressione medio del ligando (recettore) osservato è alto rispetto ai livelli di espressione medi osservabili casualmente nello stesso cluster k.
In particolare, la distribuzione di questo livello di espressione media osservabile casualmente è stata ottenuta attraverso un approccio di tipo permutativo, seguendo questi passaggi:

1. per ogni cluster k selezionare la sottomatrice X^k, costituita dalle sole colonne che si riferiscono a cellule appartenenti a tale gruppo cellulare;
2. permutare casualmente le righe di tale sottomatrice in modo indipendente per ogni colonna;
3. calcolare i livelli medi di espressione genica in tale versione mescolata;
4. ripetere le fasi 2 e 3 più volte.

Per il teorema del limite centrale, la distribuzione così calcolata può essere approssimata da una distribuzione normale, caratterizzata da una propria media e una propria deviazione standard, anche se le variabili originali non sono normalmente distribuite; pertanto nell'effettivo, il punteggio S(g,k) è stato calcolato come probabilità di osservare valori inferiori al livello medio di espressione del ligando (recettore) g nella sottomatrice X^k quando si campionano valori da una distribuzione normale, con media pari al valor medio della sottomatrice X^k e standard deviation pari al rapporto tra la deviazione standard della sottomatrice X^k e la radice quadrata del numero di cellule presenti nel cluster k.
Un aspetto importante di tale formulazione è che essa tiene conto anche della variabilità dell'espressione genica e del numero di cellule per cluster durante il calcolo dei valori di punteggio del ligando (recettore). Ad esempio, lo stesso livello di espressione media del ligando (recettore) sarà considerato meno affidabile (cioè avrà un punteggio S più basso) se osservato in un cluster con poche cellule o una grande varianza, rispetto alla sua osservazione in un cluster con molte cellule e un'espressione genica meno rumorosa (assumendo la stessa espressione media di base nei cluster).

2.1.2 Score intercellulare

La seconda fase invece riguarda il modo in cui i due punteggi S(g,k) della coppia ligando – recettore vengono combinati al fine di ottenere una stima dell'intensità della comunicazione intercellulare. Dato che, affinché una comunicazione possa essere considerata in corso, è necessario che siano attivi contemporaneamente sia il ligando che il suo recettore specifico, è stato proposto il punteggio di segnalazione intercellulare $S_{inter}(l, r, k1, k2)$ tra il cluster cellulare k1 e il cluster cellulare k2 attraverso la coppia ligando – recettore (l, r), con l espresso da k1 e r espresso da k2, come segue:

$$S_{inter}(l, r, k1, k2) = min\big(S(l, k1), S(r, k2)\big), \qquad S_{inter} \in [0,1].$$

É stato scelto il minimo come operatore per stabilire il punteggio intercellulare, poiché esso modella l'idea che il segnale più debole è quello che definisce l'intensità della comunicazione intercellulare in corso. In effetti, a differenza di altre funzioni che sono state proposte e che sono comunemente utilizzate per effettuare questa stima, in questo caso il minimo non risulta essere distorto in presenza di geni interagenti con livelli di espressione ampiamente diversi tra

loro; in altre funzioni infatti, come il prodotto, nell'eventualità vi sia un gene che domina l'intero segnale di interazione, ad esempio nel caso di un ligando espresso in modo ridotto che interagisce con un recettore espresso in modo elevato, si osserverà un punteggio intercellulare rilevante seppur il livello molto basso di espressione del ligando indichi una mancata comunicazione tra le cellule.
D'altro canto invece, il minimo, seppur perda informazione riguardo al livello più alto di espressione genica, risulta avere un comportamento di tipo maggiormente conservativo, che, come già detto in precedenza, risulta essere uno dei requisiti cardine che un modello di questo tipo dovrebbe avere.

2.2 Altri score intercellulari

2.2.1 CellChat score

Il primo articolo preso in analisi, ovvero quello in cui è presentato lo score *CellChat* [9], propone una serie di passaggi volti a calcolare l'inferenza delle comunicazioni intercellulari, che possono essere schematicamente riassunti come segue:

1. identificazione di geni di segnalazione espressi in modo differenziale; per dedurre le comunicazioni specifiche dello stato cellulare, è necessario innanzitutto identificare i geni di segnalazione espressi in modo differenziale in tutti i gruppi cellulari all'interno di un dato set di dati scRNA-seq, utilizzando il Wilcoxon rank sum test con un livello di significatività di 0.05. Quest'ultimo è un potente test non parametrico che permette di verificare, in presenza di valori ordinali provenienti da una distribuzione continua, se due campioni statistici provengono dalla stessa popolazione; in particolare viene utilizzato in presenza di campioni indipendenti;
2. calcolo della 'ensemble average expression', ovvero dell'espressione media d'insieme; al fine di tenere in conto e ridurre al minimo gli effetti dovuti al rumore, risulta maggiormente efficace calcolare l'espressione media d'insieme dei geni di segnalazione all'interno di un dato gruppo cellulare sfruttando il seguente metodo statistico per il calcolo della media:

$$EM = \frac{1}{2}\,Q2 + \frac{1}{4}\,(Q1 + Q3)$$

dove Q1, Q2 e Q3 sono rispettivamente il primo, il secondo e il terzo quartile dei livelli di espressione di un gene di segnalazione all'interno di un cluster cellulare;

3. calcolo della probabilità di comunicazione intercellulare; al fine di modellare l'interazione cellulare mediata da coppie proteiche ligando – recettore, risulta efficace utilizzare la legge dell'azione di massa, la quale afferma che la velocità di una reazione chimica è proporzionale alla concentrazione delle sostanze partecipanti. Utilizzando quindi una tecnica di propagazione basata sul random walk, è possibile ricavare, facendo affidamento sui profili proiettati di ligandi e recettori, la seguente formula, che esprime la probabilità di comunicazione $P_{i,\,j}$ tra due gruppi cellulari i e j, per una particolare coppia ligando – recettore, indicata col simbolo k:

$$P_{i,j}^{k} = \frac{L_i R_j}{K_h + L_i R_j} \times \left(1 + \frac{AG_i}{K_h + AG_i}\right) \cdot \left(1 + \frac{AG_j}{K_h + AG_j}\right) \times \frac{K_h}{K_h + AN_i} \cdot \frac{K_h}{K_h + AN_j} \times \frac{n_i n_j}{n^2}$$

dove L_i e R_j rappresentano il livello di espressione medio del ligando L e del recettore R rispettivamente all'interno del cluster i e del cluster j; per modellare l'interazione tra il ligando L ed il recettore R è stata utilizzata una funzione di Hill (primo termine), utilizzando un parametro costante K_h, che dovrebbe essere posto in linea teorica pari al valore intermedio del range dei dati; esso infatti è assunto di default pari a 0.5, considerando un intervallo di valori di ingresso compreso tra 0 e 1.

Come noto dalla letteratura scientifica, agonisti ed antagonisti extracellulari, appartenenti sia alla cellula sorgente sia alla cellula bersaglio, sono in grado di andare a modificare anche in modo significativo l'interazione ligando – recettore, sia agendo in modo diretto che in modo indiretto; pertanto, è necessario calcolare l'espressione media di tali agonisti (indicata con AG) ed utilizzare una funzione di Hill per modellare la loro modulazione positiva sull'interazione ligando – recettore (secondo e terzo termine); occorre agire allo stesso modo per quanto riguarda gli antagonisti (indicati con AN) al fine di modellarne l'effetto negativo esercitato sull'interazione (quarto e quinto termine).

Infine, è necessario tenere in considerazione anche il contributo dato dalla quantità di cellule presenti in ciascun gruppo cellulare, dato che si stanno considerando dati trascrittomici di singole cellule non ordinati; tale effetto è rappresentato dall'ultimo termine della formula, dove n_i e n_j sono rispettivamente il numero di cellule presenti

all'interno del gruppo cellulare i e del gruppo cellulare j, mentre n rappresenta il numero totale di cellule prese in considerazione all'interno del set di dati;

4. Identificazione delle comunicazioni cellulari statisticamente rilevanti. Per identificare tali interazioni significative viene utilizzato un test di permutazione, il quale prevede la permutazione casuale delle etichette dei gruppi cellulari e il ricalcolo della probabilità di interazione tra il cluster i e il cluster j attraverso la coppia ligando L e recettore R; il *p-value* di ogni probabilità $P_{i,j}$ è calcolato nel modo seguente:

$$p - value = \frac{\left\{\#m \middle| P_{i,j}^{(m)} \leq P_{i,j},\ m = 1, 2, \ldots, M\right\}}{M}$$

ovvero è pari al rapporto tra la somma del numero di permutazioni che hanno un valore di probabilità minore o uguale a quello di riferimento e il numero di permutazioni M, che viene posto come default pari a 100; il simbolo $P_{i,j}^{(m)}$ indica la probabilità di comunicazione nella m-esima permutazione.

L'interazione è considerata statisticamente rilevante solo nel caso in cui essa abbia un *p-value* minore di 0.05.

2.2.2 NATMI score

Nel secondo articolo [10] preso in considerazione in questa analisi non viene proposto un vero e proprio unico punteggio intercellulare, indice dell'avvenuta comunicazione tra ligando e recettore e quindi tra le due cellule; vengono bensì proposti tre differenti score che hanno lo scopo di andare a valutare la forza delle connessioni tra due gruppi cellulari per quanto riguarda una coppia ligando – recettore specifica.

Il primo punteggio proposto è il 'mean-expression weight', o peso dell'espressione media, il quale esprime la forza della connessione tra due tipi di cellule basandosi sui livelli medi di espressione di ligandi e recettori; esso è calcolato come il prodotto tra il livello medio di espressione del ligando L in uno specifico cluster i, e il livello medio di espressione del recettore R in uno specifico cluster j:

$$MEW = L_i \times R_j$$

Il secondo punteggio suggerito è lo 'specificity weight', o 'peso di specificità'; esso fornisce indicazioni su quanto un gene è specifico per un particolare tipo di cellula ed è quindi una misura della forza della connessione esistente tra il ligando (recettore) e il gruppo cellulare.
Tale punteggio si calcola come prodotto tra due termini:

- il rapporto tra il livello medio di espressione del ligando L nel cluster i e la somma dei livelli medi di espressione di tale ligando in tutti i tipi cellulari del set di dati a disposizione;
- il rapporto tra il livello medio di espressione del recettore R nel cluster j e la somma dei livelli medi di espressione di tale recettore in tutti i tipi cellulari del set di dati a disposizione.

La formula matematica è la seguente:

$$SW = \frac{L_i}{\sum_{i=1}^{n} L_i} \times \frac{R_j}{\sum_{j=1}^{n} R_j}$$

È presente infine un terzo punteggio, il 'total-expression weight', o peso totale di espressione, il quale ha come obiettivo quello di misurare la forza di una connessione tra due tipi di cellule, prendendo in considerazione l'abbondanza di ogni tipo di cellula nella rete; in questo caso quindi il 'peso' della connessione è dato dal prodotto tra la somma dei livelli di espressione del ligando L nel cluster cellulare i e la somma dei livelli di espressione del recettore R all'interno del cluster j.
La formula matematica è la seguente:

$$TEW = \sum_{a=1}^{m} l_a \times \sum_{b=1}^{n} r_b$$

dove l_a e r_b sono rispettivamente il livello di espressione del ligando all'interno della a-esima cellula del cluster i e il livello di espressione del recettore all'interno della b-esima cellula del cluster j.
Da notare come il livello medio di espressione del ligando L all'interno del cluster i è pari a:

$$L_i = \frac{\sum_{a=1}^{m} l_a}{m}$$

con m pari al numero di cellule del cluster i.
In modo analogo si osserva che il livello medio di espressione del recettore R all'interno del cluster j è pari a:

$$R_j = \frac{\sum_{b=1}^{n} r_b}{n}$$

con n pari al numero di cellule del cluster j.
A questo punto è quindi possibile riscrivere la formula del peso totale di espressione nel modo seguente:

$$TEW = (L_i \times m) \times (R_j \times n)$$

Da cui, ricordando la formula utilizzata per il calcolo del peso dell'espressione media, è possibile ricavare la seguente relazione tra i due punteggi:

$$TEW = MEW \times (m \times n)$$

ovvero il total-expression weight risulta essere pari al prodotto tra il mean-expression weight e il prodotto tra il numero di cellule nei due rispettivi cluster cellulari.

Questo ultimo punteggio proposto risulta essere utile nel caso si voglia conoscere in termini assoluti ad esempio quale gruppo cellulare produce maggiormente una data tipologia di ligando (recettore), indipendentemente dal livello medio di espressione; possono infatti esserci dei cluster con livelli di espressione medi molto bassi, ma che, avendo una grande abbondanza cellulare, producono comunque maggiori quantità di ligando (recettore) rispetto ad altri gruppi cellulari con livelli di espressione medi relativamente più alti.

2.2.3 SoptSC score

Il terzo articolo [11] scelto per le analisi presenta, differentemente dal secondo e in analogia con il primo, un unico punteggio intercellulare, che tiene conto non solo dei livelli medi di espressione di ligandi e recettori, ma bensì anche di altri fattori che possono influenzare o essere influenzati dall'interazione.

L'idea alla base di questo metodo è quella di andare ad individuare delle particolari vie di comunicazione tra le varie cellule, che non comprendono solo il ligando e il recettore, ma anche i geni della cellula ricevente che vengono up-regolati o down-regolati in caso di avvenuta comunicazione; pertanto, data una coppia ligando – recettore, con L e R indicanti le distribuzioni delle espressioni geniche, rispettivamente del ligando e del recettore, in tutte le cellule, si considerano la matrice $Y = [Y_{i,j}]$ (di dimensioni $m_1 \times n$), che rappresenta gli m_1 geni che vengono up-regolati dalla via di comunicazione, e la matrice $Y^* = [Y^*_{i,j}]$ (di dimensioni $m_2 \times n$), che denota invece gli m_2 geni che vengono down-regolati in caso di interazione tra ligando e recettore.

Con queste premesse la probabilità che un segnale sia stato trasferito da una cellula i ad una cellula j attraverso la particolare via di segnalazione considerata, può essere espressa dalla seguente formula:

$$P_{i,j} = \frac{\exp\left(-\frac{1}{L_i R_j}\right) K_{i,j} \exp\left(-\frac{m_1}{\sum_{v=1}^{m_1} Y_{v,j}}\right) \Lambda_{i,j} \exp\left(-\frac{\sum_{v=1}^{m_2} Y^*_{v,j}}{m_2}\right)}{\sum_k \alpha_{i,k} K_{i,k} \beta_k \Lambda_{i,k} \gamma_k}$$

dove i seguenti valori sono così definiti:

$$K_{i,j} = \frac{\alpha_{i,j}}{\alpha_{i,j} + \beta_j}, \quad \Lambda_{i,j} = \frac{\alpha_{i,j}}{\alpha_{i,j} + \gamma_j}, \quad \alpha_{i,j} = \exp\left(-\frac{1}{L_i R_j}\right),$$

$$\beta_j = \exp\left(-\frac{m_1}{\sum_{v=1}^{m_1} Y_{v,j}}\right), \quad \gamma_j = \exp\left(-\frac{\sum_{v=1}^{m_2} Y^*_{v,j}}{m_2}\right)$$

Come si può notare, all'interno della formula sono presenti tre termini esponenziali, i quali esprimono tre contributi differenti; il primo stima la probabilità di interazione tra le cellule i e j basandosi esclusivamente sui livelli di espressione all'interno di esse rispettivamente del ligando e del recettore; la scelta di utilizzare questa forma dell'esponenziale è funzionale ad ottenere valori alti in presenza di entrambi i livelli di espressione elevati, mentre di ottenere valori molto bassi o addirittura nulli se uno dei livelli di espressione risulta essere prossimo a zero.

Il secondo termine esponenziale quantifica invece il numero di geni che vengono up-regolati all'interno della cellula j a seguito dell'interazione tra ligando e recettore; tale termine risulta pesato tramite il coefficiente $K_{i,j}$, il quale fa in modo che la presenza di questi geni attivati dalla

cascata di reazione, vada a modificare in positivo il punteggio intercellulare solo ed esclusivamente nel caso in cui il primo esponenziale abbia un valore sufficientemente elevato.
Il terzo ed ultimo esponenziale al contrario esprime la quantità di geni che vengono down-regolati all'interno della cellula j a seguito della comunicazione intercellulare; in modo analogo al secondo esponenziale è presente un coefficiente $\Lambda_{i,j}$ con lo scopo di pesare il suo contributo all'interno della formula.
Infine, è presente un termine al denominatore che ha lo scopo di normalizzare le probabilità di comunicazione dividendole per la somma di tutte le probabilità di segnalazione per quella particolare via di comunicazione.

Capitolo 3: Analisi ed implementazione degli score di interazione intercellulare

3.1 Analisi scSeqComm

Tenendo in considerazione il fatto che l'obiettivo principale è quello di fornire strumenti aggiuntivi validi per la valutazione delle comunicazioni intercellulari, è stato inizialmente preso in analisi il codice scritto in R in cui è stato realizzato lo score scSeqComm, al fine di comprendere meglio la struttura e la logica del programma in cui sarebbero stati poi inseriti gli altri punteggi; una volta ricavate le informazioni necessarie, è stato possibile implementare gli score presentati nei diversi articoli ed inserirli all'interno del codice in modo da poter effettuare confronti e valutazioni sia sulla validità e l'efficacia dei singoli score, sia sulle diversità presenti tra questi ultimi.

3.1.1 Struttura programma

Il fatto di utilizzare il codice scritto di *scSeqComm* come base di partenza per l'implementazione degli altri score e per gli studi di correlazione, seppur da un lato abbia permesso di non partire da zero nella scrittura del codice, dall'altro ha comportato alcune difficoltà; comprendere e fare proprie centinaia di righe di codice scritte da altri programmatori è infatti sicuramente un compito non da poco, ma di fondamentale importanza per capire come strutturare le funzioni atte ad implementare gli score selezionati. Per maggiore chiarezza e comprensione, è quindi di seguito riportata la struttura di base del programma *scSeqComm*.
Il programma si occupa di effettuare una serie di operazioni al fine di calcolare i punteggi intercellulari a partire da dati forniti dall'utente, sotto forma di una matrice. In particolare, si assume che la matrice di partenza contenga dati già normalizzati, che i cluster cellulari siano già stati definiti e che essa abbia la struttura mostrata nella Figura 3.1.

	Cy71_CD45_D08_S524_comb	Cy81_FNA_CD45_B01_S301_comb	Cy80_II_CD45_B07_S883_comb
RPS11	12.632885	11.2040347	11.646985
ELMO2	5.115184	0.0000000	0.000000
CREB3L1	0.000000	0.0000000	0.000000
PNMA1	0.000000	0.0000000	0.000000
MMP2	2.941108	0.0000000	0.000000
TRAF3IP2-AS1	2.664487	3.3950571	4.230317
C10orf90	0.000000	6.6009047	4.190636
ZHX3	0.000000	2.3950520	0.000000
ERCC5	0.000000	0.0000000	5.314352
APBB2	0.000000	3.6633461	0.000000
PDCL3	0.000000	8.0398138	6.138475
AEN	0.000000	0.0000000	5.350441
DECR1	8.373910	0.0000000	6.484010
RPS18	14.307288	12.9374738	13.301943
BRIX1	0.000000	8.8256915	4.904974

Figura 3.1 Matrice contenente i livelli di espressione dei geni (inclusi ligandi e recettori) in tutte le cellule del set di dati. In figura solo un piccolo sottoinsieme delle righe (geni) e colonne (cellule) è riportato.

Ovvero si assume che abbia come indici delle righe i diversi nomi dei geni relativi a ligandi e recettori, mentre come indici delle colonne i nomi delle diverse cellule sequenziate; all'interno della matrice sono contenuti i livelli di espressione dei singoli geni nelle relative singole cellule. I gruppi cellulari a cui appartengono le singole cellule invece sono contenuti in una variabile separata.

Il programma comprende nel suo complesso un gran numero di processi e funzioni, ma la maggior parte di essi non è di interesse per la trattazione fatta in seguito e non verrà pertanto presa in considerazione in questa breve panoramica; in particolare è stata analizzata solo la prima parte del codice, ovvero quella riguardante il calcolo del punteggio intercellulare.

La prima operazione effettuata dal programma, una volta caricata la matrice dei dati, è quella di effettuare una selezione sulle righe, in modo tale da mantenere solo i geni che sono contenuti nel database delle coppie ligando – recettore a disposizione.

3.1.2 Analisi score del ligando e del recettore

Una volta effettuato il filtraggio dei dati come descritto alla fine del precedente paragrafo, viene invocata la funzione *compute_score_S_ligand_receptor_all_cluster*, la quale ha lo scopo di calcolare il punteggio di tutti i ligandi e recettori all'interno di ogni cluster, seguendo quanto detto nel paragrafo 2.1.1.

La prima parte della funzione è osservabile nella Figura 3.2.

```
# name of each cell cluster
cluster_names <- names(cell_cluster)

# list that will contain, for each cluster, the values of ligands and receptors scores
ligands_receptors_S_score <- list()

# identify row indices of ligands and receptors
lig_id <- which(row.names(exprMatr) %in% ligands)
rec_id <- which(row.names(exprMatr) %in% receptors)
```

Figura 3.2 funzione compute_score_S_ligand_receptor_all_cluster (prima parte)

In essa la funzione si occupa di:

- salvare in una variabile i nomi dei vari gruppi cellulari;
- inizializzare una variabile che conterrà gli score da restituire in output;
- identificare gli indici delle righe in cui sono presenti i ligandi e i recettori contenuti nella lista data in input e ottenuta con l'intersezione dei dati menzionata in precedenza.

```
# each cell cluster
for(cl in cluster_names){

  cell_id <- which(colnames(exprMatr) %in% cell_cluster[[cl]])

  # compute the ligands and receptors S score for the subset of the count matrix related to the current cell cluster
  ligands_receptors_S_score[[cl]] <- compute_score_S_ligand_receptor (exprMatr = exprMatr,
                                                                      cells = cell_id,
                                                                      ligands = lig_id,
                                                                      receptors = rec_id,
                                                                      H0_genes = H0_genes,
                                                                      verbose = verbose, debug_result = debug_result,
                                                                      exprMatr_type = exprMatr_type)
}
```

Figura 3.3 funzione compute_score_S_ligand_receptor_all_cluster (seconda parte)

Successivamente, come è possibile vedere nella Figura 3.3 essa opera un ciclo che itera per ogni cluster presente nella lista la medesima operazione, ovvero invocare la funzione *compute_score_S_ligand_receptor*, la quale si occupa effettivamente di calcolare il punteggio cluster per cluster.

Tale funzione è a sua volta strutturata come mostrato nella Figura 3.4.

```
    res <- list()
    # compute ligands and receptors expression levels (as mean of their normalized counts across cells)
    res$lig_avg_expr <- Matrix::rowMeans(exprMatr[ligands, cells])
    res$rec_avg_expr <- Matrix::rowMeans(exprMatr[receptors, cells])

    # compute mean and sd of gaussian approximation
    res$H0_mean <- Matrix::mean(exprMatr[H0_genes,cells])
    res$H0_sd <- sd(exprMatr[H0_genes,cells]) / sqrt(length(cells))
}
```

Figura 3.4 funzione compute_score_S_ligand_receptor (prima parte)

Per prima cosa viene inizializzata una lista vuota che conterrà i risultati delle operazioni successive, dopodiché vengono calcolati:

- il livello di espressione medio dei ligandi, attraverso la media sulle righe della sottomatrice avente come righe solo quelle corrispondenti ai ligandi dati input e come colonne le cellule di quel particolare cluster cellulare;
- il livello di espressione medio dei recettori, calcolato in modo analogo a quello dei ligandi;
- la media, calcolata su tutti i valori della sottomatrice;
- la deviazione standard, sempre su tutti i valori della sottomatrice.

```
score_L <- pnorm(res$lig_avg_expr, mean = res$H0_mean, sd = res$H0_sd)
score_R <- pnorm(res$rec_avg_expr, mean = res$H0_mean, sd = res$H0_sd)

# imposing zero score when average expression levels is zero
if (length(which(res$lig_avg_expr ==0 ))>0) score_L[which(res$lig_avg_expr == 0)] <- 0
if (length(which(res$rec_avg_expr ==0 ))>0) score_R[which(res$rec_avg_expr == 0)] <- 0

results <- list(ligand_score = score_L, receptor_score = score_R,
                lig_avg_expr = res$lig_avg_expr, rec_avg_expr = res$rec_avg_expr)
```

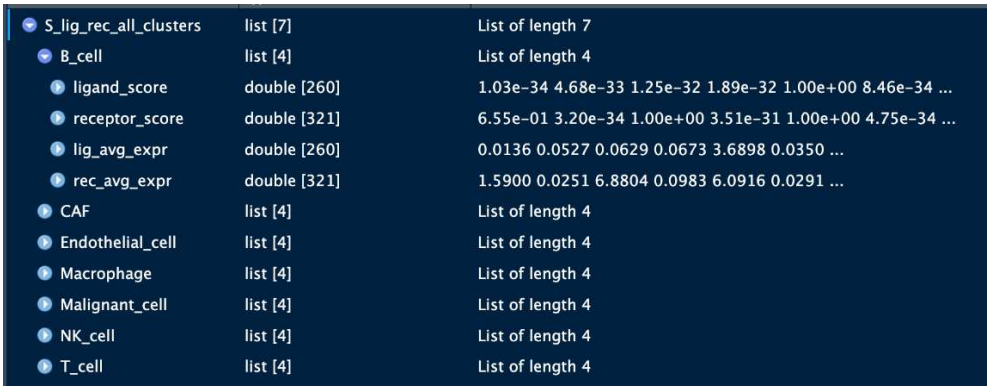
Figura 3.5 funzione compute_score_S_ligand_receptor (seconda parte)

A questo punto, con tutti i dati a disposizione, è finalmente possibile, come mostrato nella Figura 3.5, calcolare gli score cellulari per i ligandi e i recettori, come già detto nel paragrafo 2.1.1, ovvero come probabilità di osservare valori inferiori al livello medio di espressione del ligando (recettore) g nella sottomatrice X^k quando si campionano valori da una distribuzione normale.

Il caso in cui il livello medio di espressione sia pari a zero viene trattato separatamente, dato che la funzione *pnorm* in questo caso non opera correttamente.

Infine, vengono restituiti in output gli score e i livelli medi di espressione; tali valori vengono quindi raccolti dalla funzione di partenza nella lista *ligands_receptors_S_score*, inizializzata in precedenza, poi restituita in output al programma principale.

Tale lista apparirà quindi come mostrato nella Figura 3.6.



S_lig_rec_all_clusters	list [7]	List of length 7
B_cell	list [4]	List of length 4
ligand_score	double [260]	1.03e-34 4.68e-33 1.25e-32 1.89e-32 1.00e+00 8.46e-34 ...
receptor_score	double [321]	6.55e-01 3.20e-34 1.00e+00 3.51e-31 1.00e+00 4.75e-34 ...
lig_avg_expr	double [260]	0.0136 0.0527 0.0629 0.0673 3.6898 0.0350 ...
rec_avg_expr	double [321]	1.5900 0.0251 6.8804 0.0983 6.0916 0.0291 ...
CAF	list [4]	List of length 4
Endothelial_cell	list [4]	List of length 4
Macrophage	list [4]	List of length 4
Malignant_cell	list [4]	List of length 4
NK_cell	list [4]	List of length 4
T_cell	list [4]	List of length 4

Figura 3.6 lista contenente i punteggi S di ligandi e recettori e i loro livelli medi di espressione, suddivisi per cluster

Ovvero con una suddivisione per gruppi cellulari, ognuna delle quali contiene la lista dei quattro valori calcolati all'interno delle due funzioni.

3.1.3 Analisi dello score intercellulare

Una volta terminata l'esecuzione delle funzioni precedenti, la lista così ottenuta, insieme ad una serie di altre variabili, viene data in input alla funzione *compute_score_S_inter_all_cluster*, la quale si occupa di calcolare il punteggio intercellulare di interesse per la successiva trattazione. Nella figura 3.7 è possibile leggere il codice della funzione.

```
#number of cell clusters
N <- length(S_ligand_receptor_all_clusters)

#object that will contain, for each cluster couple, the respective ligands and receptors S score and the activation score of the ligand-receptor pair
LR_pairs_scores <- NULL

# each cell cluster as ligand-expressing cluster
for(cl_l in 1:N){
  # each cell cluster as receptor-expressing cluster
  for (cl_r in 1:N){

    #compute the ligands and receptors S score and the activation score of the LR pairs for a given cell cluster couple
    LR_pairs_scores <- rbind(LR_pairs_scores, compute_score_S_inter(scores_cluster_l = S_ligand_receptor_all_clusters[cl_l],
                                                                    scores_cluster_r = S_ligand_receptor_all_clusters[cl_r],
                                                                    LR_pairs = LR_pairs, cell_cluster= cell_cluster, exprMatr = exprMatr,
                                                                    LR_subunits = LR_subunits, method = method, exprMatr_type = exprMatr_type))

    gc(verbose = FALSE)
  }
}
```

Figura 3.7 funzione compute_score_S_inter_all_cluster

In essa si può notare la presenza di un doppio ciclo che itera sia sui cluster del ligando che sui cluster del recettore, andando a chiamare, per ogni possibile combinazione, la funzione *compute_score_S_inter*, i cui dati di output verranno poi uniti iterazione dopo iterazione a formare una macrotabella contenente tutti i dati necessari.
Per vedere quindi le linee di codice che effettivamente implementano lo score come visto nel paragrafo 2.1.2, occorre analizzare la funzione sopra menzionata, il cui codice è mostrato nella Figura 3.8.

```
#ligand and receptor S scores in the respective clusters
l_score <- scores_cluster_l[[1]]$ligand_score[LR_pairs$ligand]
r_score <- scores_cluster_r[[1]]$receptor_score[LR_pairs$receptor]

mean.count_L <- scores_cluster_l[[1]]$lig_avg_expr[LR_pairs$ligand]
mean.count_R <- scores_cluster_r[[1]]$rec_avg_expr[LR_pairs$receptor]
#save results

results <- rbind(results, data.frame(ligand = LR_pairs$ligand, receptor = LR_pairs$receptor, LR_pair = paste(LR_pairs$ligand,"-",LR_pairs$receptor),
                                     cluster_L = rep(names(scores_cluster_l),nrow(LR_pairs)), cluster_R = rep(names(scores_cluster_r),nrow(LR_pairs)),
                                     interaction = rep(paste(names(scores_cluster_l),"-->",names(scores_cluster_r) ),nrow(LR_pairs)),
                                     L_score_S_lr = l_score, R_score_S_lr = r_score, S_inter = apply(cbind(l_score,r_score),1,min),
                                     mean.count_L = mean.count_L, mean.count_R = mean.count_R, mean.product = mean.count_L * mean.count_R,
                                     stringsAsFactors = FALSE))
```

Figura 3.8 funzione compute_score_S_inter

In realtà il codice risulta essere concettualmente molto semplice, limitandosi ad estrarre i valori dalla lista data in output dalle precedenti funzioni, a calcolare (evidenziato) il punteggio intercellulare tramite la funzione min, e a salvare tutte le informazioni che possono risultare utili all'interno di un dataframe, poi dato in output.
Ciò che si otterrà quindi alla fine di tutto il processo, è la tabella osservabile nella Figura 3.9.

	ligand	receptor	LR_pair	cluster_L	cluster_R	interaction	L_score_S_lr	R_score_S_lr	S_inter	mean.count_L	mean.count_R
12	ADAM17	ITGB1	ADAM17 - ITGB1	B_cell	B_cell	B_cell --> B_cell	1.0000000	0.9999683	0.9999683	2.429410	2.039358
14	ADAM28	ITGA4	ADAM28 - ITGA4	B_cell	B_cell	B_cell --> B_cell	1.0000000	1.0000000	1.0000000	5.912817	3.164545
49	ARF1	CHRM3	ARF1 - CHRM3	B_cell	B_cell	B_cell --> B_cell	1.0000000	1.0000000	1.0000000	4.816584	4.322952
53	B2M	HLA-F	B2M - HLA-F	B_cell	B_cell	B_cell --> B_cell	1.0000000	1.0000000	1.0000000	13.099089	5.504355
54	B2M	LILRB1	B2M - LILRB1	B_cell	B_cell	B_cell --> B_cell	1.0000000	1.0000000	1.0000000	13.099089	2.886160
72	BTLA	CD79A	BTLA - CD79A	B_cell	B_cell	B_cell --> B_cell	1.0000000	1.0000000	1.0000000	2.874927	9.756175
73	BTLA	TNFRSF14	BTLA - TNFRSF14	B_cell	B_cell	B_cell --> B_cell	1.0000000	1.0000000	1.0000000	2.874927	3.210429
93	CALR	SCARF1	CALR - SCARF1	B_cell	B_cell	B_cell --> B_cell	1.0000000	0.9981116	0.9981116	5.897499	1.901612
99	CCL16	HRH4	CCL16 - HRH4	B_cell	B_cell	B_cell --> B_cell	1.0000000	0.9981682	0.9981682	2.547004	1.902802
121	CD55	CD97	CD55 - CD97	B_cell	B_cell	B_cell --> B_cell	1.0000000	0.8996980	0.8996980	4.779704	1.699899
122	CD55	CR1	CD55 - CR1	B_cell	B_cell	B_cell --> B_cell	1.0000000	1.0000000	1.0000000	4.779704	2.236276
240	FGF5	FGFR2	FGF5 - FGFR2	B_cell	B_cell	B_cell --> B_cell	1.0000000	0.8715091	0.8715091	3.702481	1.681644
279	GPI	AMFR	GPI - AMFR	B_cell	B_cell	B_cell --> B_cell	1.0000000	0.9933978	0.9933978	5.803083	1.849446
290	HLA-A	APLP2	HLA-A - APLP2	B_cell	B_cell	B_cell --> B_cell	1.0000000	1.0000000	1.0000000	10.012240	2.383282
292	HLA-A	LILRB1	HLA-A - LILRB1	B_cell	B_cell	B_cell --> B_cell	1.0000000	1.0000000	1.0000000	10.012240	2.886160

Figura 3.9 Tabella contenente i dati relativi alle coppie ligando recettore analizzate che costituisce l'output di scSeqComm. La figura riporta solo un piccolo sottoinsieme delle righe ottenute in output.

In tale tabella sono quindi riportati: il nome del ligando e del recettore, sia singolarmente che in coppia, il cluster di appartenenza di entrambi, anch'essi sia singolarmente che insieme, i punteggi singoli, il punteggio intercellulare e i livelli medi di espressione.
Da notare che in questo caso specifico tutti i punteggi sono molto alti poiché è stato applicato un filtro alla tabella per selezionare solo i casi con score sufficientemente elevati, come si può vedere dagli indici delle righe, i quali non sono più sequenziali; nelle analisi successive il filtro è però stato rimosso al fine di valutare gli score in tutte le combinazioni possibili, non solo quelle favorevoli per il punteggio di riferimento.
In aggiunta a ciò, occorre sottolineare che alla tabella è possibile aggiungere ulteriori colonne contenenti gli altri score implementati, in modo da poter effettuare un confronto immediato e visualizzare analogie e differenze tra i diversi punteggi.

3.2 Implementazione CellChat score

Per quanto riguarda il primo articolo analizzato, in cui è stato proposto il punteggio *CellChat*, rappresentato dalla formula:

$$P_{i,j}^{k} = \frac{L_i R_j}{K_h + L_i R_j} \times \left(1 + \frac{AG_i}{K_h + AG_i}\right) \cdot \left(1 + \frac{AG_j}{K_h + AG_j}\right) \times \frac{K_h}{K_h + AN_i} \cdot \frac{K_h}{K_h + AN_j} \times \frac{n_i n_j}{n^2}$$

al fine di poter confrontare tale punteggio con lo score intercellulare di riferimento, il quale tiene conto esclusivamente dei livelli di espressione media, non è stato considerato il contributo fornito dalle molecole agoniste e antagoniste; la formula della probabilità è stata pertanto semplificata nella forma seguente:

$$P_{i,j}^{k} = \frac{L_i R_j}{K_h + L_i R_j} \times \frac{n_i n_j}{n^2}$$

in cui si mantiene il contributo dato dai livelli di espressione media L_i e R_j, pesandolo per le dimensioni dei cluster i e j di riferimento.
Viene inoltre proposto un metodo alternativo a quello della semplice media per calcolare il livello medio di espressione, ovvero quello dell'espressione media di insieme, il quale sfrutta il primo, il secondo e il terzo quartile; pertanto, sono stati realizzati due differenti score

intercellulari: il primo, denotato come ‘semplificato’, utilizzando la media calcolata nella parte di codice già scritta; il secondo sfruttando il metodo proposto nell’articolo.

3.2.1 Prima versione (semplificata)

```r
compute_semplified_CellChat_score <- function(gene_expr, cell_group, S_inter_all_clusters, kh, gene_expr_type = "matrix"){

  # compute number of cells in each cluster
  cluster_length <-lapply(cell_group,length)

  # compute the semplified CellChat score
  n_cells = ncol(gene_expr)
  semplified_CellChat_score <- array(0, dim = nrow(S_inter_all_clusters))
  for (i in (1:nrow(S_inter_all_clusters))){
    semplified_CellChat_score[i] <- (S_inter_all_clusters$mean.count_L[i] * S_inter_all_clusters$mean.count_R[i]) /
      (Kh + (S_inter_all_clusters$mean.count_L[i] * S_inter_all_clusters$mean.count_R[i]))*
      (cluster_length[[S_inter_all_clusters$cluster_L[i]]]*cluster_length[[S_inter_all_clusters$cluster_R[i]]]/(n_cells^2))
  }

  return(semplified_CellChat_score)

}
```

Figura 3.10 funzione compute_semplified_CellChat_score

Questa prima versione dello score, mostrata nella Figura 3.10, non richiede di effettuare ulteriori calcoli, dato che quasi tutti i dati necessari da inserire all’interno della formula per il calcolo della probabilità $P_{i,j}$ vengono forniti in input alla funzione:

- *gene_expr* è la matrice di dati data in input dall’utente nel programma principale;
- *cell_group* è la variabile, sempre fornita dell’utente, contenente la lista delle cellule appartenenti ai vari cluster;
- *S_inter_all_clusters* è la tabella mostrata nel paragrafo 3.1.4;
- *kh* è il parametro costante, lasciato a discrezione dell’utente, in base al range dei valori in ingresso (se ad esempio il range va da 0 a 1, allora *kh* sarà posto a 0.5);
- *gene_expr_type* serve invece al codice per capire la forma con la quale sono stati forniti i dati, impostata di default in matrice.

Per poter implementare la formula semplificata è necessario ricavare sia le dimensioni dei vari cluster, sia il numero totale di cellule; per quanto riguarda le dimensioni, esse vengono ottenute attraverso la funzione *lapply*, la quale applica la funzione length alla lista *cell_group*; il numero totale delle cellule si ricava invece contando le colonne presenti nella matrice *gene_expr*.
A questo punto non resta che applicare, riga per riga della tabella *S_inter_all_clusters*, la formula della probabilità, creando così un vettore colonna, potenzialmente aggiungibile a tale

tabella, contenente i punteggi di tutte le possibili combinazioni; tale vettore è poi restituito in output tramite la variabile *semplified_CellChat_score*.

3.2.2 Seconda versione

La seconda versione del punteggio richiede numerosi passaggi per ottenere tutti i valori necessari da inserire all'interno della formula, pertanto l'analisi verrà suddivisa in più parti.

```
compute_CellChat_score <- function(ligands, receptors, gene_expr, cell_group, S_inter_all_clusters, kh = 0.5, gene_expr_type = "matrix"){

  # identify row indices of ligands and receptors
  lig_id <- which(row.names(gene_expr) %in% ligands)
  rec_id <- which(row.names(gene_expr) %in% receptors)
```

Figura 3.11 funzione compute_CellChat_score (prima parte)

Innanzitutto, come è possibile notare nella Figura 3.11, rispetto alla versione precedente, sono presenti due variabili aggiuntive date in input alla funzione:

- *ligands* è una variabile contenente i nomi dei ligandi;
- *receptors* è una variabile contenente i nomi dei recettori.

Esse, come si può vedere nelle righe successive di codice (Figura 3.12), sono funzionali ad ottenere gli indici delle righe della matrice dei dati in cui sono presenti rispettivamente i ligandi e i recettori contenuti nel database di scSeqComm.

```
# name of each cell cluster
cluster_names <- names(cell_group)

lig_rec_avg_expr <- list()

# each cell cluster
for(cl in cluster_names){

  cell_id <- which(colnames(gene_expr) %in% cell_group[[cl]])

  res <- list()

  # compute ligands and receptors ensemble average expression
  lig_quantiles <- t(apply(gene_expr[lig_id,cell_id],1,quantile, probs = c(0.25,0.50,0.75)))
  res$lig_avg_expr <- 0.5*lig_quantiles[,2] + 0.25*(lig_quantiles[,1] + lig_quantiles[,3])
  rec_quantiles <- t(apply(gene_expr[rec_id,cell_id],1,quantile, probs = c(0.25,0.50,0.75)))
  res$rec_avg_expr <- 0.5*rec_quantiles[,2] + 0.25*(rec_quantiles[,1] + rec_quantiles[,3])

  # save the results in a list with values divided by clusters
  lig_rec_avg_expr[[cl]] <- res

}
```

Figura 3.12 funzione compute_CellChat_score (seconda parte)

A questo punto, essendo necessario ricalcolare i livelli medi di espressione, vengono salvati in una variabile i nomi dei diversi gruppi cellulari e viene inizializzata una lista vuota che conterrà tali valori, suddivisi per cluster.
Viene quindi utilizzato un ciclo che itera sui vari nomi dei cluster precedentemente ottenuti e che per ogni cluster effettua le seguenti operazioni:

- ricavare gli indici delle colonne riferiti alle cellule appartenenti a quello specifico cluster;
- inizializzare una lista vuota temporanea per contenere i risultati temporanei;
- applicare alla sottomatrice (ricavata tramite gli indici appena estratti) la funzione *quantile*, che in questo caso specifico estrae i tre quartili di interesse per la formula;
- utilizzare la formula dell'espressione media di insieme per calcolare i livelli di espressione medi e salvarli nella lista temporanea;
- salvare i risultati nella lista inizializzata fuori dal ciclo.

```
cluster_length <-lapply(cell_group,length)

# calculate the CellChat score
n_cells = ncol(gene_expr)
CellChat_score <- array(0, dim = nrow(S_inter_all_clusters))
for (i in (1:nrow(S_inter_all_clusters))){

  # extract information from S_inter_all_clusters
  cluster_L <- S_inter_all_clusters$cluster_L[i]
  cluster_R <- S_inter_all_clusters$cluster_R[i]
  ligand <- S_inter_all_clusters$ligand[i]
  receptor <- S_inter_all_clusters$receptor[i]

  # extract the mean value of the specific ligand in the specific cluster
  mean_L <- lig_rec_avg_expr[[cluster_L]][[1]][ligand]

  # extract the mean value of the specific receptor in the specific cluster
  mean_R <- lig_rec_avg_expr[[cluster_R]][[2]][receptor]

  # compute CellChat_score
  CellChat_score[i] <- (mean_L * mean_R) / (Kh + (mean_L * mean_R))*
    (cluster_length[[S_inter_all_clusters$cluster_L[i]]]*cluster_length[[S_inter_all_clusters$cluster_R[i]]]/(n_cells^2))
}

return(CellChat_score)
```

Figura 3.13 funzione compute_CellChat_score (terza parte)

Una volta ottenuti i livelli medi di espressione, come nella prima versione, occorre ottenere la lunghezza dei cluster e il numero di cellule totali; una volta fatto ciò, nuovamente si utilizza un ciclo per calcolare il punteggio intercellulare.
Sia per necessità che per maggiore semplicità, gran parte delle informazioni contenute nelle variabili sono state estratte e salvate in delle variabili temporanee; in particolare sono stati

estratti i nomi dei cluster e dei ligandi, in modo da poter ottenere il valore specifico dalla variabile *lig_rec_avg_expr*, in cui erano stati salvati tutti i livelli medi di espressione.
A questo punto è stata nuovamente applicata la formula semplificata ricavata dall'articolo e il vettore dei risultati è stato mandato in output tramite la variabile *CellChat_score*.

3.3 Implementazione NATMI score

Per quanto riguarda i punteggi presentati nel paragrafo 2.2.2, essi, a differenza del caso precedente, non sono stati né modificati né semplificati in alcun modo; tuttavia, dato che anche l'articolo stesso in cui sono stati suggeriti, pur avendolo proposto, non ha utilizzato in alcun modo il 'total-expression weight', non risultando esso utile per le tipologie di analisi effettuate sugli score, non è stato implementato.

3.3.1 Peso medio di espressione

```
compute_NATMI_mean_expression_weight <- function(S_inter_all_clusters, gene_expr_type = "matrix"){

  NATMI_mean_expression_weight <- (S_inter_all_clusters$mean.count_L * S_inter_all_clusters$mean.count_R)

  return(NATMI_mean_expression_weight)

}
```

Figura 3.14 funzione compute_NATMI_mean_expression_weight

Il calcolo di questo score in realtà richiederebbe dei passaggi aggiuntivi rispetto a quelli riportati nel codice visibile nella Figura 3.14; tuttavia, essendo già stati calcolati i livelli di espressione media all'interno della funzione mostrata al paragrafo 3.1.4, è stato sufficiente dare in input alla funzione la tabella *S_inter_all_clusters*.
Il punteggio è infatti stato calcolato facendo il prodotto elemento per elemento tra i vettori colonna dei livelli medi di espressione di ligandi e recettori; esso è poi stato dato in output grazie alla variabile *NATMI_mean_expression_weight*.

3.3.2 Peso di specificità

```
compute_NATMI_specificity_weight <- function(S_inter_all_clusters, S_lig_rec_all_clusters, gene_expr_type = "matrix"){

  # sum mean expression in all clusters
  lig_rec_avg_expr_sum <- list()
  lig_rec_avg_expr_sum$lig_avg_expr <- S_lig_rec_all_clusters[[1]]$lig_avg_expr
  lig_rec_avg_expr_sum$rec_avg_expr <-S_lig_rec_all_clusters[[1]]$rec_avg_expr

  for (i in 2:length(S_lig_rec_all_clusters)){

    lig_rec_avg_expr_sum$lig_avg_expr <- lig_rec_avg_expr_sum$lig_avg_expr + S_lig_rec_all_clusters[[i]]$lig_avg_expr
    lig_rec_avg_expr_sum$rec_avg_expr <- lig_rec_avg_expr_sum$rec_avg_expr + S_lig_rec_all_clusters[[i]]$rec_avg_expr

  }
```

Figura 3.15 funzione compute_NATMI_specificity_weight (prima parte)

Il secondo punteggio proposto (Figura 3.15 e Figura 3.16) presenta delle difficoltà aggiuntive rispetto al primo; è infatti necessario calcolare la somma dei livelli di espressione in tutti i cluster; al fine di fare ciò, in input viene fornita la variabile *S_lig_rec_all_clusters*, dalla quale vengono estratti tutti i valori dei vari cluster e poi sommati all'interno della lista *lig_rec_avg_expr_sum* inizialmente vuota.

```
  # compute NATMI specificity weight
  NATMI_specificity_weight <- array(0, dim = nrow(S_inter_all_clusters))

  for (i in (1:nrow(S_inter_all_clusters))){
    ligand <- S_inter_all_clusters$ligand[i]
    receptor <- S_inter_all_clusters$receptor[i]
    NATMI_specificity_weight[i] <- (S_inter_all_clusters$mean.count_L[i]/lig_rec_avg_expr_sum$lig_avg_expr[ligand]) *
      (S_inter_all_clusters$mean.count_R[i]/lig_rec_avg_expr_sum$rec_avg_expr[receptor])

  }

  return(NATMI_specificity_weight)


}
```

Figura 3.16 funzione compute_NATMI_specificity_weight (seconda parte)

Una volta ottenuti tali valori, non resta che inizializzare una variabile che conterrà il punteggio e utilizzare un ciclo che calcoli, riga per riga della tabella *S_inter_all_clusters*, il peso di specificità così come indicato nell'articolo, ovvero:

$$SW = \frac{L_i}{\sum_{i=1}^{n} L_i} \times \frac{R_j}{\sum_{j=1}^{n} R_j}$$

Una volta terminato il ciclo il vettore colonna *NATMI_specificity_weight*, contenente lo specificity weight, viene dato come output della funzione.

3.4 Implementazione SoptSC score

Rispetto a quelli proposti nei precedenti articoli, questo è sicuramente il punteggio che risulta maggiormente alterato rispetto alla proposta originale degli autori; tuttavia, al fine di rendere omogenei e confrontabili tutti i punteggi, si è reso necessario andare ad effettuare due significative modifiche.
La prima modifica riguarda la formula stessa, che è stata ridotta al fine di eliminare i contributi delle molecole up-regolate e down-regolate all'interno della cellula, dato che gli altri punteggi tengono in considerazione esclusivamente i livelli medi di espressione e, solo in alcuni casi, le dimensioni dei cluster.
La formula semplificata ha quindi mantenuto esclusivamente il primo dei tre termini esponenziali, eliminando completamente il denominatore:

$$P_{i,j} = \exp\left(-\frac{1}{L_i R_j}\right)$$

La seconda modifica concerne invece il significato attribuito ai simboli L_i e R_j, i quali nella versione originale si riferiscono a livelli di espressione dei singoli geni all'interno di specifiche cellule, in particolare del ligando L nella cellula i e del recettore R nella cellula j; in questo caso tuttavia, al fine di allinearsi con le scelte fatte negli altri punteggi, tali simboli sono stati intesi come i livelli medi di espressione rispettivamente del ligando L nel cluster i e del recettore R nel cluster j, passando quindi da un livello particolare (cellule) ad uno più generale (gruppi cellulari).

3.4.1 Versione semplificata

```
compute_semplified_SoptSC_score <- function(S_inter_all_clusters, gene_expr_type = "matrix"){

  semplified_SoptSC_score <- exp(-1/(S_inter_all_clusters$mean.count_L * S_inter_all_clusters$mean.count_R))

  return(semplified_SoptSC_score)
}
```

Figura 3.17 funzione compute_semplified_SoptSC_score

In analogia con il peso medio di espressione trattato nel paragrafo 3.3.1, questo punteggio risulta essere di relativamente facile realizzazione grazie al fatto che la parte di calcoli necessaria per ottenere i livelli medi di espressione di ligandi e recettori nei rispettivi cluster viene già effettuata durante il calcolo del punteggio intercellulare di *scSeqComm*.
Grazie a ciò, come mostrato nella Figura 3.17, è sufficiente fornire in input la tabella *S_inter_all_clusters*, da cui è possibile estrarre i livelli medi di espressione ed inserirli all'interno della formula semplificata; il vettore dei risultati viene poi dato in output tramite la variabile *semplified_SoptSC_score*.

3.5 Dataset e database LR utilizzati

Una volta analizzato e compreso il codice del programma di base e le funzioni utilizzate per la realizzazione dello score intercellulare *scSeqComm*, e dopo aver implementato i cinque score selezionati come mostrato nei sottocapitoli precedenti, è necessario, al fine di verificare il funzionamento del codice e di effettuare dei confronti tra i diversi punteggi, utilizzare un set di dati scRNA-seq.
In questo caso è stato scelto come riferimento il dataset di *Tirosh et al.* [12], ovvero un set di dati scRNA-seq di melanoma metastatico umano disponibile pubblicamente online e già utilizzato per analisi in questo campo sia dall'articolo preso come riferimento [7], sia anche in un'altra analisi indipendente [13]. Le informazioni riguardanti sia il dataset che la suddivisione in gruppi cellulari (clustering) sono stati ottenuti dal GEO (GSE72056), database pubblico gestito dal National Center for Biotechnology Information (NCBI) degli Stati Uniti; in particolare le cellule sono state raggruppate in sette cluster: cellule endoteliali, cellule maligne, cellule T, cellule NK, cellule B, macrofagi e fibroblasti associati al cancro (CAFs).
Tale set di dati si presenta sotto forma di matrice, costituita da 4134 colonne, corrispondenti ad altrettante cellule appartenenti ai sette cluster appena citati, e da 11253 righe, corrispondenti invece ai geni sequenziati, per un totale di più di 46 milioni di valori.
Per quanto riguarda il database di coppie ligando – recettore utilizzato per questa analisi, sebbene il codice del pacchetto *scSeqComm* offra la possibilità di scegliere tra un'ampia gamma di database LR da poter fornire in input al codice principale assieme alla matrice dei dati, si è scelto di utilizzare il medesimo database LR utilizzato dai due studi sopra menzionati, al fine di mantenere una coerenza con le precedenti analisi e rendere più semplice un eventuale confronto con esse; tale database è costituito da 1901 coppie ligando – recettore [14].

Capitolo 4: Confronto tra gli score intercellulari continui

4.1 I punteggi continui

Una volta calcolati i punteggi grazie alle funzioni illustrate nel capitolo precedente, è stato possibile effettuare dei confronti tra i risultati forniti dai sei score a disposizione, ovvero quello già presente nel codice scSeqComm, e i cinque implementati successivamente.
Seppur i diversi score selezionati risultino essere differenti sia in termini di range di valori in cui esprimono l'intensità della comunicazione cellulare, sia per quanto riguarda gli specifici aspetti biologici che vanno a mettere in evidenza, è comunque opportuno porre a confronto i risultati ottenuti da essi, a fronte dell'utilizzo del medesimo set di dati. Tale confronto infatti permette di individuare sia eventuali similarità nelle capacità predittive dei punteggi, sia eventuali differenze dovute alle differenti scelte modellistiche adottate dagli score.

	scSeqComm	semplifiedCellChat	CellChat	NATMI_MEW	NATMI_SW	SoptSC
1	1.956163e-32	2.969780e-04	0.000000000	9.754657e-03	3.185486e-05	3.007750e-45
2	4.530162e-01	1.403466e-02	0.000000000	4.726479e+00	8.965958e-03	8.093094e-01
3	5.623827e-34	5.931465e-05	0.000000000	1.918322e-03	2.253102e-04	4.046608e-227
4	1.160093e-32	3.139973e-03	0.000000000	1.268228e-01	8.279841e-04	3.763392e-04
5	1.160093e-32	1.196300e-04	0.000000000	3.884163e-03	8.288885e-05	1.543087e-112
6	1.005949e-30	4.894681e-04	0.000000000	1.628317e-02	7.111690e-05	2.131194e-27
7	5.623827e-34	1.409179e-04	0.000000000	4.581674e-03	1.731660e-04	1.623760e-95
8	1.234990e-30	4.990530e-04	0.000000000	1.661262e-02	1.255550e-04	7.203767e-27
9	3.507841e-29	5.854795e-03	0.000000000	3.029004e-01	6.363612e-04	3.683102e-02
10	3.902935e-32	3.379390e-04	0.000000000	1.113003e-02	1.300933e-04	9.548176e-40
11	1.005949e-30	5.393694e-03	0.000000000	2.663380e-01	8.401908e-04	2.340917e-02
12	9.999683e-01	1.409671e-02	0.011640228	4.954435e+00	7.518113e-03	8.172262e-01
13	1.043873e-11	1.200909e-02	0.000000000	1.710572e+00	7.096538e-03	5.573288e-01
14	1.000000e+00	1.511544e-02	0.014784608	1.871138e+01	9.054227e-02	9.479596e-01
15	2.352399e-30	4.499842e-04	0.000000000	1.493043e-02	1.995004e-04	8.168212e-30

Figura 4.1 Tabella contenente i vettori colonna dei sei score presi in esame (prime 15 righe di 30968)

Al fine di confrontare visivamente i risultati dei diversi score, è stato creato un dataframe, avente come colonne i sei score intercellulari da analizzare e come righe le diverse coppie ligando recettore nei vari cluster; sono riportate nell'immagine le prime quindici righe, ma il totale di coppie analizzate è 30968.

A questo punto sono quindi stati effettuati dei confronti sfruttando la funzione *cor()*, la quale calcola la correlazione esistente tra due insiemi di dati, in questo caso specifico corrispondenti ai vettori colonna dei diversi score; in particolare la funzione è stata utilizzata nella sua versione di default, la quale calcola la correlazione lineare di Pearson; tale metodo è una misura di quanto due variabili tendono a variare insieme in modo lineare e può variare tra -1 e 1:

- se il valore è vicino a -1, vi è una forte correlazione negativa, ovvero quando una variabile aumenta, l'altra tende a diminuire;
- se il valore tende a 1, è presente una forte correlazione positiva, cioè quando una variabile aumenta, l'altra tende ad aumentare;
- se il valore si trova vicino a 0, ciò è indice di una correlazione debole o nulla, ovvero le due variabili non variano insieme nello stesso modo.

In particolare, nelle successive analisi i vari punteggi saranno considerati:

- debolmente correlati, quelli compresi tra 0 e 0.3;
- moderatamente correlati, quelli compresi tra 0.3 e 0.7;
- altamente correlati, quelli compresi tra 0.7 e 1.

Per le successive trattazioni i punteggi verranno indicati nel modo seguente:

- primo punteggio o *scSeqComm* → lo score intercellulare già implementato nel codice di partenza, presentato nel sottocapitolo 2.1 e mostrato nel sottocapitolo 3.1;
- secondo punteggio o *semplified_CellChat* → la prima versione del punteggio proposto nel tool *CellCha*t, ovvero sfruttando il livello medio di espressione; esso è stato presentato nel paragrafo 2.2.1 e la sua implementazione è mostrata nel paragrafo 3.2.1;
- terzo punteggio o *CellChat* → la seconda versione del punteggio proposto nel tool *CellChat*, ovvero quella che sfrutta l'espressione media di insieme; la formula di partenza è la medesima del precedente, mostrata nel paragrafo 2.2.1, mentre la sua implementazione è presente nel paragrafo 3.2.2;
- quarto punteggio o *NATMI_MEW* → il peso dell'espressione media proposto nel tool *NATMI*, facente riferimento alla formula del paragrafo 2.2.2, la cui implementazione è mostrata nel paragrafo 3.3.1;
- quinto punteggio o *NATMI_SW* → il peso di specificità proposto all'interno del tool *NATMI*; la formula è quella presentata nel paragrafo 2.2.2 e implementata nel paragrafo 3.3.2;
- sesto punteggio o *Sopt_SC* → il metodo di calcolo dell'intensità di interazione proposto nel tool SoptSC, mostrato nel paragrafo 2.2.3 ed implementato nel paragrafo 3.4.1.

4.2 Confronto con scSeqComm

4.2.1 Confronto tra scSeqComm e i punteggi CellChat

Dato che i confronti sono stati effettuati tra tutte le possibili combinazioni sensate di punteggi, al fine di procedere con ordine, si è partiti con i confronti tra ognuno dei cinque score implementati e lo score di riferimento, ovvero *scSeqComm*.
I primi ad essere posti a confronto con esso, sono state le due versioni del punteggio presentato nel tool *CellChat*; il confronto è utile non solo per vedere quindi se la versione semplificata adottata in questa analisi restituisce risultati in linea con lo score intercellulare di riferimento, ma anche per verificare come l'utilizzo dell'espressione media d'insieme ottenuta tramite i quartili differisca dalla versione che utilizza il semplice livello medio di espressione (utilizzato in tutti gli altri score).
Un altro fattore che viene tenuto in conto in queste valutazioni è il termine costante Kh, che viene suggerito di default a 0.5; tuttavia l'articolo, nel proporre tale valore, suppone che i dati della matrice di ingresso siano già normalizzati nel range che va da 0 a 1, cosa non vera però in questo caso; pertanto l'analisi ha coinvolto non due ma bensì quattro punteggi, dato che entrambe le versioni sono state calcolate ponendo Kh pari a 0.5 nel primo caso e a circa 1.894 nel secondo, dove tale valore è stato ricavato applicando la funzione *mean()* a tutta la matrice dei dati.
Per quanto riguarda il caso di Kh pari a 0.5, si può notare che, seppur in entrambi i casi la correlazione risulta essere poco rilevante, è sicuramente presente in misura maggiore nel terzo punteggio, ovvero nella versione con il peso medio d'insieme; la correlazione tra *scSeqComm* e *semplified_CellChat (Kh 0.5)*, infatti, è pari a circa 0.082, mentre tra *scSeqComm* e *CellChat (Kh 0.5)* è circa uguale a 0.234, quasi tre volte tanto.
Se si considerano le versioni con il termine Kh posto a 1.894, si osserva un aumento in entrambi i casi, anche se per quanto riguarda il *semplified_CellChat (Kh 1.894)* l'aumento risulta decisamente molto più significativo, con un punteggio di correlazione che, passando da 0.082 a quasi 0.148, risulta essere quasi raddoppiato. Nel caso dello score *CellChat (Kh 1.894)*, l'aumento c'è, ma risulta essere poco significativo, passando da 0.234 a 0.259. Tutti i confronti effettuati sono riassunti nella Tabella 4.2.

	semplified_CellChat (Kh 0.5)	***semplified_CellChat (Kh 1.894)***	***CellChat (Kh 0.5)***	***CellChat (Kh 1.894)***
scSeqComm	0.082	0.148	0.234	0.259

Tabella 4.2 Confronto tra scSeqComm e le quattro versioni dello score CellChat. La tabella contiene i valori di correlazione.

4.2.2 Confronto tra scSeqComm e i punteggi NATMI

Successivamente il confronto ha coinvolto i punteggi proposti nel tool *NATMI*, ovvero *NATMI_MEW* e *NATMI_SW*; a differenza degli altri punteggi implementati, in questo caso non vi sono eventuali problemi dovuti alla semplificazione delle formule, dato che i due score sono stati implementati come descritto nell'articolo; nonostante ciò, essi non sono stati pensati propriamente come evidenza dell'interazione tra ligando e recettore, bensì come indicatori rispettivamente della forza di connessione esistente tra due gruppi cellulari e della specificità di un gene per un dato cluster.
Come però suggerito dagli autori dell'articolo, essi possono anche essere utilizzati con lo scopo di segnalare interazione cellulare, pertanto risulta utile operare confronti anche con questi punteggi; a differenza di quelli proposti nel primo studio, entrambi gli score *NATMI* presentano una correlazione con *scSeqComm* decisamente maggiore; in particolare il *NATMI_MEW* presenta il livello di correlazione più elevato di tutta l'analisi, pari a circa 0.736, mentre *NATMI_SW*, con un valore di 0.473, si attesta comunque tra i più alti per quanto riguarda i punteggi continui. I confronti effettuati sono riassunti nella Tabella 4.3.

	NATMI_MEW	***NATMI_SW***
scSeqComm	0.736	0.473

Tabella 4.3 Confronto tra scSeqComm e i due score NATMI. La tabella contiene i valori di correlazione.

4.2.3 Confronto tra scSeqComm e Sopt_SC

Il punteggio proposto nel tool *SoptSC* è sicuramente quello la cui formula è stata maggiormente semplificata rispetto all'originale; è possibile pertanto che esso, non tenendo conto dei molteplici fattori con cui era stata concepita la formula, dia risultati non il linea con la versione

originale; nonostante ciò però, quello che si ottiene è un punteggio che, seppur con valori su una scala completamente diversa rispetto a *scSeqComm*, è abbastanza in linea con lo score di riferimento; esso infatti si attesta relativamente in alto per quanto concerne il livello di correlazione, secondo soltanto al *NATMI_MEW*.
Il livello di correlazione tra *scSeqComm* e *Sopt_SC* è infatti pari a 0.666, pari a poco meno del triplo del livello più alto ottenuto dai vari punteggi di *CellChat*, dimostrandosi quindi molto più affine con lo score di riferimento.

4.2.4 Considerazioni generali

Unendo le considerazioni fatte nei tre paragrafi precedenti è possibile notare come in linea generale il tool che propone punteggi che danno risultati maggiormente in linea con quelli forniti dallo score *scSeqComm* è *NATMI*, con il suo punteggio *NATMI_MEW* che si presenta con una correlazione di livello alto e il punteggio *NATMI_SW* che ne manifesta una di tipo moderato; anche il tool *SoptSC*, in realtà, propone un punteggio che ha una correlazione con lo score di riferimento abbastanza elevata, collocandosi in una via di mezzo tra il livello moderato e quello alto; al contrario di questi invece, tutte e quattro le versioni della formula di CellChat sono debolmente correlate con *scSeqComm*.
Dunque, in conclusione, per quanto riguarda le loro versioni continue, si può affermare che gli score *NATMI* e *SoptSC* siano relativamente affini a *scSeqComm* nella segnalazione dell'evidenza della comunicazione intercellulare, mentre *CellChat* si discosti abbastanza da tali previsioni.

4.3 Confronto tra gli altri punteggi

Oltre ad operare confronti tra i diversi punteggi e lo score principale di riferimento, risultano essere di interesse anche analisi di correlazione tra i punteggi stessi implementati; in particolare le osservazioni di maggiore rilevanza sono sicuramente quelle relative al confronto tra i diversi punteggi *CellChat* realizzati, in quanto, seppur utilizzando dati diversi, essi sfruttano la medesima formula per il calcolo dello score.
I punteggi a disposizione sono quattro, pertanto le possibili combinazioni per cui calcolare la correlazione sarebbero pari a $\binom{4}{2}$, ovvero 6; risultano tuttavia di utilità per questa analisi solo

quattro di esse, cioè quelle atte a verificare le variazioni nel punteggio dovute al cambiamento o del parametro *Kh* oppure del metodo di calcolo dei livelli medi di espressione.

	CellChat (Kh 1.894)	***semplified_CellChat (Kh 0.5)***
CellChat (Kh 0.5)	0.981	0.672
semplified_CellChat (Kh 1.894)	0.780	0.961

Tabella 4.4 Confronto tra le quattro versioni dello score CellChat. La tabella contiene i valori di correlazione.

Nella Tabella 4.4 sono riportate le quattro combinazioni appena menzionate, dove la scelta del valore del parametro *Kh* è indicata tra parentesi, mentre, come detto nel sottocapitolo 4.1, la scelta dei semplici livelli medi di espressione rispetto all'espressione media di insieme è indicata dalla dicitura *semplified*.

È possibile notare come il passaggio del parametro *Kh* da 0.5 a 1.984 (vedi diagonale principale della tabella 4.4) risulta essere molto meno significativo rispetto all'utilizzo dell'espressione media d'insieme al posto del semplice livello medio d'espressione (vedi diagonale secondaria della tabella 4.4); la correlazione tra le due versioni del punteggio *semplified_CellChat* è infatti pari a 0.961, mentre per quanto riguarda il punteggio *CellChat*, essa è pari addirittura a 0.981.

La variazione del metodo di calcolo dei livelli medi di espressione di ligandi e recettori porta invece i punteggi ad avere un andamento abbastanza diverso; nel caso delle versioni con *Kh* pari a 1.894, infatti, il livello di correlazione si attesta intorno a 0.780, mentre per quanto riguarda le versioni originali, esso scende addirittura sotto la soglia di 'alta correlazione', arrivando a circa 0.672.

Da tali osservazioni si può concludere che la variazione del parametro *Kh* risulta essere meno significativa nel caso del punteggio *CellChat*, avendo i due score una correlazione molto prossima ad 1; inoltre ponendo il parametro *Kh* pari a 0.5 i punteggi risultano maggiormente suscettibili al cambiamento di metodo per il calcolo di L_i e R_j.

Per quanto riguarda la correlazione tra i punteggi *CellChat* e quelli proposti da *NATMI* e *SoptSC* invece, i risultati sono i medesimi che si possono osservare con lo score di riferimento; essi,

infatti, in tutte le loro versioni, risultano essere debolmente correlati con gli altri punteggi intercellulari, non superando mai la soglia di 0.3 come mostrato nella tabella 4.5.

	NATMI_MEW	***NATMI_SW***	***Sopt_SC***
semplified_CellChat (Kh 0.5)	0.072	0.017	0.188
semplified_CellChat (Kh 1.894)	0.129	0.048	0.251
CellChat (Kh 0.5)	0.179	0.080	0.251
CellChat (Kh 1.894)	0.208	0.081	0.250

Tabella 4. 5 Confronto tra le quattro versioni CellChat e gli altri score. La tabella contiene i valori di correlazione.

Gli ultimi confronti che è possibile fare sono quelli tra i tre punteggi rimanenti, ovvero *NATMI_MEW*, *NATMI_SW* e *Sopt_SC*; diversamente da quanto si potrebbe pensare, tra le tre possibili combinazioni, quella che risulta dare una minore correlazione è quella dei due punteggi proposti nel secondo articolo, ovvero il peso di espressione medio e il peso di specificità; essi, infatti, hanno un grado di correlazione pari a 0.401, inferiore a tutte le altre correlazioni, fatta eccezione per quelle riguardanti i punteggi *CellChat*.
Per quanto riguarda la correlazione con il punteggio *Sopt_SC*, tra i due punteggi *NATMI* quello che risulta essere maggiormente in accordo con il suo andamento è il *NATMI_MEW*, così come nel caso del punteggio di riferimento *scSeqComm*; il livello di correlazione tra *Sopt_SC* e *NATMI_MEW*, è infatti pari a 0.614, mentre tra *Sopt_SC* e *NATMI_SW*, non supera il valore di 0.546.

Capitolo 5: Confronto tra gli score intercellulari binari

5.1 Valutazione dei livelli di soglia

Seppur al fine di avere un'indicazione generale del grado di concordanza tra due punteggi risulti utile il confronto tra le loro versioni continue, spesso gli score sono usati per fornire una informazione binaria, ovvero se la comunicazione cellulare è da considerarsi in corso (indicato con il numero 1) oppure no (indicato con il numero 0). Pertanto, risulta di interesse anche un'analisi del grado di concordanza tra le versioni binarie dei punteggi implementati, in modo tale da capire, a prescindere dai valori assoluti forniti dai diversi score, se essi segnalino o meno le medesime coppie ligando – recettore come attive.

Dato che tutti i punteggi presi in esame in questo studio sono di tipo continuo e che in linea generale non sono stati suggeriti all'interno degli articoli dei veri e propri valori di soglia (o di 'threshold') che permettessero di compiere facilmente la conversione, si è reso necessario testare diversi livelli di soglia e analizzare i risultati al variare di tale soglia.

Per ottenere tali valori di soglia è possibile utilizzare diversi metodi a seconda dell'obiettivo che si vuole raggiungere; in questa analisi l'interesse principale è osservare come varia il grado di correlazione tra lo score di riferimento *scSeqComm*, per il quale si è scelta la soglia, proposta anche all'interno dell'articolo stesso, di 0.5, e i diversi score implementati passando dalle versioni continue a quelle binarie.

Per questo motivo nella scelta della soglia si sono tenuti in conto due principali fattori, ovvero il grado di correlazione con *scSeqComm* e il numero di coppie ligando – recettore segnalate come attive; sono infatti stati selezionati i livelli di threshold che da un lato garantissero un maggior grado di correlazione, dall'altro che fornissero un numero di coppie attive simile a quello dello score di riferimento, pari a 3519 su un totale di 30968; in particolare si sono ritenute adatte a questo scopo esclusivamente le soglie che fornivano un valore di coppie attive compreso tra le 2000 e le 5000 unità e sono state preferite quelle che, in presenza di livelli di correlazione abbastanza simili, restituivano un numero di coppie più vicino a 3519.

Al fine di valutare la concordanza tra uno score reso binario attraverso un determinato livello di soglia, e lo score binario ottenuto da *scSeqComm*, è stata sviluppata la funzione *compute_threshold_value*, riportata, suddivisa in più parti, nelle Figure 5.1, 5.2 e 5.3.

```r
compute_threshold_value <- function(rate, score_bin_ref, score){

  # create the x vector based on rate value given in input
  x <- seq(min(score), max(score), rate)

  # create the y vector which will contain the value of correlation
  y <- rep(0,length(x))

  # calculate the correlation with different threshold values
  for (i in (1:length(x))){

    score_bin <- score
    score_bin[score_bin <= x[i]] <- 0
    score_bin[score_bin > x[i]] <- 1
    y[i] <- mcc(score_bin, score_bin_ref)

  }
```

Figura 5.1 funzione compute_threshold_value (prima parte)

La funzione si occupa di calcolare la correlazione presente tra due punteggi binari, di cui uno preso come riferimento e con soglia fissa, al variare nel secondo del valore di threshold utilizzato per la conversione.

La funzione necessita di tre valori in input:

- *rate* → valore numerico che indica gli intervalli di variazione della soglia, ovvero di quanto aumenta il suo valore ad ogni ciclo;
- *score_bin_ref* → punteggio già convertito in binario sfruttando una soglia fissata che verrà preso come riferimento per il calcolo della correlazione (in generale è stato utilizzato *scSeqComm*);
- *score* → punteggio preso in analisi all'interno della funzione.

Per prima cosa viene generato il vettore contenente tutti i possibili valori di soglia, basandosi sul range dello score in ingresso e sul tasso di incremento fornito in input; successivamente viene inizializzato il vettore vuoto che conterrà i differenti valori di correlazioni corrispondenti ai relativi livelli di soglia; a questo punto attraverso un ciclo viene fatta ad ogni iterazione la conversione in binario del punteggio in esame e calcolato il grado di correlazione sfruttando la funzione *mcc()*, ovvero una funzione che implementa il coefficiente di correlazione di Matthews, noto in statistica con il nome di *phi coefficient*; esso è un coefficiente che fornisce, proprio come il coefficiente lineare di Pearson utilizzato precedentemente nel quarto capitolo tramite la funzione *cor()* con gli score continui, una misura del livello di correlazione tra due insiemi di dati; il coefficiente di correlazione di Matthews risulta tuttavia essere più appropriato per descrivere l'associazione tra due variabili binarie, essendo stato concepito specificamente per esse.

	y = 1	y = 0	totale
x = 1	n_{11}	n_{10}	$n_{1\bullet}$
x = 0	n_{01}	n_{00}	$n_{0\bullet}$
totale	$n_{\bullet 1}$	$n_{\bullet 0}$	n

Tabella 5.2 Schema tabella per il confronto tra score binari

Basandosi sulla Tabella 5.2 il coefficiente di correlazione di Matthews considera due variabili binarie come positivamente correlate se la maggior parte dei dati ricade nella diagonale principale, ovvero o in n_{11} o in n_{00}; viceversa considera due variabili binarie come negativamente correlate se la maggior parte dei dati sono contenuti nella diagonale secondaria, ovvero in n_{10} o in n_{01}.
La formula esatta per calcolare il *phi coefficient* è la seguente:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 0}n_{\bullet 1}}}$$

```
# plot
title <- deparse(substitute(score))
plot(x,y,
     type = "l",                          # Tipo di grafico (linee e punti)
     col = "red",                         # Colore delle linee e dei punti
     pch = 16,                            # Tipo di punto
     xlab = "threshold",                  # Etichetta dell'asse x
     ylab = "correlation",                # Etichetta dell'asse y
     main = title,                        # Titolo principale
     sub = "correlazione con ScSeqComm",  # Sottotitolo
     col.main = "black",                  # Colore del titolo principale
     col.sub = "black",                   # Colore del sottotitolo
     col.lab = "black",                   # Colore dei titoli degli assi
     cex.lab = 1.2,                       # Dimensione del testo delle etichette degli assi
     cex.axis = 0.8,                      # Dimensione del testo degli assi
     cex.main = 1.3,                      # Dimensione del testo del titolo principale
     cex.sub = 0.8,                       # Dimensione del testo del sottotitolo
     lwd = 2)                             # Larghezza delle linee
```

Figura 5.3 funzione compute_threshold_value (seconda parte)

A questo punto, come è possibile osservare nella Figura 5.3, viene effettuato un plot che permette di visualizzare in un grafico l'andamento della correlazione al variare del valore di threshold.

```
  # find the peaks of correlation
  peaks <- findpeaks(y)
  correlations <- peaks[,1]
  thresholds <- x[peaks[,2]]
  counter <- rep(0, length(thresholds))

  for (j in (1:length(thresholds))){

    counter[j] <- length(score[score > thresholds[j]])

  }

  results <- data.frame(correlations = correlations, thresholds = thresholds, counter = counter)

  return(results)

}
```

Figura 5.4 funzione compute_threshold_value (terza parte)

Infine, dato che ad essere di interesse per le analisi sono i valori in cui è massima la correlazione, è stata sfruttata, come mostrato nella Figura 5.4, la funzione *findpeaks()* per trovare i picchi del vettore correlazione; essa ha permesso di ottenere sia i valori massimi di correlazione, sia i livelli di soglia corrispondenti a tali picchi, che una volta estratti, sono stati restituiti in output. Assieme ad essi in output è stato fornito un altro dato molto importante, ovvero il numero di coppie ritenute attive dallo score con quel particolare livello di soglia; infatti, come già precedentemente accennato, uno dei criteri di selezione del livello di soglia è proprio il numero di coppie segnalate come attive dallo score utilizzando tale livello.

5.1.1 Soglia CellChat score

Come visto nel paragrafo 4.2.1 i punteggi CellChat presi in considerazione in queste analisi sono risultati essere quattro, *semplified_CellChat (Kh 0.5)*, *semplified_CellChat (Kh 1.894)*, *CellChat (Kh 0.5)* e *CellChat (Kh 1.894)*.
Procedendo con ordine, il primo punteggio da convertire è *semplified_CellChat*, per il quale è stata utilizzato come punto di partenza la funzione presentata nell'introduzione del sottocapitolo 5.1. Tramite essa è stato possibile ottenere il grafico mostrato nella Figura 5.5 e la Tabella 5.6.

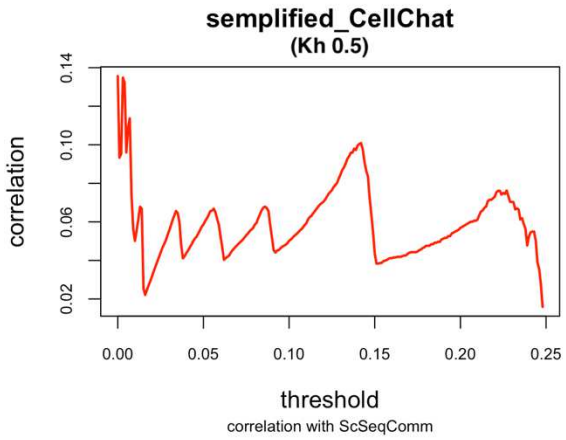


Figura 5.5 grafico dell'andamento della correlazione (MCC) tra semplified_CellChat (Kh 0.5) e scSeqComm in funzione del livello di soglia

	correlations	thresholds	counter
1	0.13487098	0.003	9131
2	0.11374079	0.007	4773
3	0.06782607	0.013	3168
4	0.06564709	0.034	1836
5	0.06682730	0.056	1242
6	0.06789564	0.086	740
7	0.09800670	0.138	291
8	0.10088267	0.142	253
9	0.07501373	0.225	62
10	0.07626862	0.227	58
11	0.06698237	0.233	42
12	0.06186798	0.236	34

Tabella 5.6 tabella con i valori di correlazione (MCC) tra semplified_CellChat (Kh 0.5) e scSeqComm e il numero di coppie segnalate come attive in funzione del livello di soglia

Si può facilmente notare, sia visivamente, sia osservando la prima colonna della Tabella 5.6, che i maggiori valori di correlazione si hanno per valori molto bassi di soglia e questo sottolinea ancora una volta come l'affinità tra il punteggio di riferimento e il *semplified_CellChat*, risulta essere molto scarsa; infatti questo andamento è dovuto al fatto che man mano che il punteggio aumenta il numero di coppie segnalate come attive, esse non coincidono con quelle proposte da *scSeqComm*, andando quindi a ridurre la correlazione tra i due score. Si può notare, osservando la Tabella 5.6, che gli unici due valori di soglia che, in base ai due criteri precedentemente fissati, risultano essere validi, sono 0.007 e 0.013; sebbene la soglia 0.007 dia un grado di correlazione pari a quasi il doppio di quello fornito dalla soglia di 0.013, quest'ultima fornisce un numero di coppie attive molto più vicino a quello di riferimento e risulta essere inoltre decisamente più conservativa nella selezione delle coppie, fattore che come già affermato nel paragrafo 1.2.2, è importante nella valutazione di uno score.

Fatte queste considerazioni, è stata ritenuta maggiormente in linea con i criteri stabiliti nel sottocapitolo 5.1 la soglia 0.013, che è quindi stata scelta come livello di riferimento per la conversione in binario di questo score, anche se ciò è andato a discapito del grado di correlazione, già relativamente basso.

Seguendo il medesimo schema di analisi del punteggio precedente, sono riportate, nella Figura 5.7 e nella Tabella 5.8, rispettivamente il grafico e la tabella relativi al punteggio *semplified_CellChat (Kh 1.894)*, dove per la seconda sono riportati solo i primi 12 valori di 30, dato che i successivi risultano in ogni caso poco significativi.

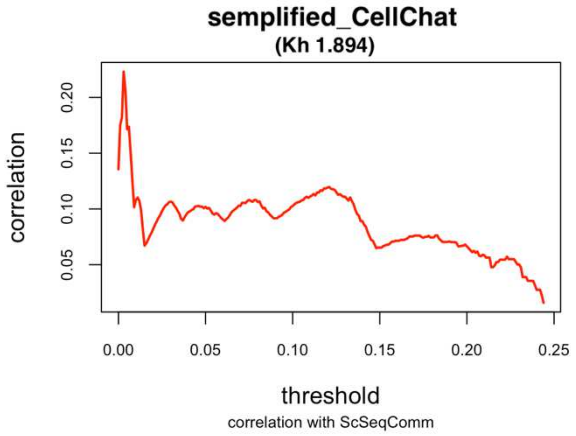


Figura 5.7 grafico dell'andamento della correlazione (MCC) tra semplified_CellChat (Kh 1.894) e ScSeqComm in funzione del livello di soglia

	correlations	thresholds	counter
1	0.22312452	0.003	6284
2	0.17367353	0.006	3599
3	0.11021435	0.011	2333
4	0.10655751	0.030	1244
5	0.10272826	0.046	857
6	0.10212480	0.048	828
7	0.10183811	0.050	792
8	0.10093582	0.052	756
9	0.09618891	0.056	695
10	0.10296210	0.069	528
11	0.10553985	0.071	506
12	0.10806318	0.075	473

Tabella 5.8 tabella con i valori di correlazione (MCC) tra semplified_CellChat (Kh 1.894) e ScSeqComm e il numero di coppie segnalate come attive in funzione del livello di soglia

Nella versione alternativa del secondo punteggio, ovvero quella con il parametro Kh posto a 1.894, si nota un andamento del livello di correlazione leggermente diverso; se da un lato come nel caso precedente il picco di maggiore correlazione è posto in prossimità di una soglia nulla, anche qui indicativo del fatto che tale score sia poco affine con quello di riferimento, dall'altro i picchi presentano mediamente gradi di correlazione più elevati e si manifestano in corrispondenza di livelli di soglia di minore entità. Analizzando in particolare la Tabella 5.8 si può osservare anche in questo caso la presenza di due livelli di threshold rientranti nel range di accettabilità, ovvero 0.006 e 0.011, tuttavia risulta evidente che il primo valore sia di gran lunga preferibile al secondo in base ai criteri esposti nel sottocapitolo 5.1; infatti esso non solo si avvicina maggiormente al numero di coppie preso come riferimento (3519), ma presenta anche un grado di correlazione decisamente più elevato, pur rimanendo comunque inferiore a 0.3 e venendo ancora quindi classificato come debole; pertanto, il livello di soglia scelto come riferimento per questo score è 0.006.

Il terzo punteggio analizzato è *CellChat (Kh 0.5)*, per il quale, come nelle due versioni del secondo punteggio, si è utilizzata la funzione *compute_threshold_value*, i cui output sono riportati nella Figura 5.9 e nella Tabella 5.10.

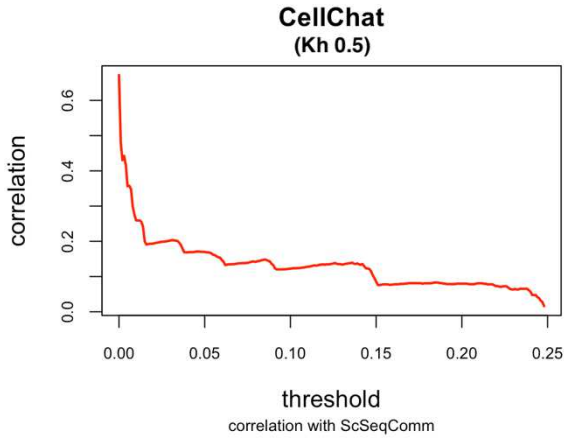


Figura 5.9 grafico dell'andamento della correlazione (MCC) tra CellChat (Kh 0.5) e scSeqComm in funzione del livello di soglia

	correlations	thresholds	counter
1	0.44229984	0.003	2939
2	0.35774888	0.006	1685
3	0.25935819	0.012	922
4	0.20330358	0.031	625
5	0.16949006	0.041	507
6	0.17036074	0.049	485
7	0.14293632	0.078	343
8	0.14832428	0.085	309
9	0.13165057	0.114	199
10	0.13444066	0.119	188
11	0.13860572	0.126	175
12	0.13915864	0.136	150

Tabella 5.10 tabella con i valori di correlazione (MCC) tra CellChat (Kh 0.5) e scSeqComm e il numero di coppie segnalate come attive in funzione del livello di soglia

Ponendo a confronto il grafico della Figura 5.9 con quelli dei punteggi precedenti, seppur visivamente possa sembrare il contrario, si nota subito una significativa differenza nel livello di correlazione medio con il punteggio *scSeqComm*, in quanto sono presenti valori che superano addirittura la soglia di 0.3, indice quindi della presenza di una moderata correlazione tra i due score; si osserva inoltre che i livelli di threshold si abbassano significativamente in termini di valori assoluti, a causa probabilmente di un valore medio dei punteggi più basso rispetto a *semplified_CellChat*.

Osservando la Tabella 5.10 è evidente come vi sia un unico livello di soglia rientrante nei parametri stabiliti nel sottocapitolo 5.1, ovvero il primo picco della lista; l'assenza di picchi precedenti è probabilmente un altro indice del fatto che la versione dello score che sfrutta l'espressione media di insieme al posto dei semplici livelli medi di espressione risulta essere maggiormente correlata della sua controparte con lo score di riferimento.

In conclusione, quindi la soglia scelta per effettuare la conversione in binario del punteggio *CellChat (Kh 0.5)* è 0.003.

Ultimo punteggio tra quelli implementati basandosi sul primo articolo è *CellChat (Kh 1.894)*, per il quale il procedimento utilizzato è il medesimo dei tre precedenti; la funzione restituisce in output il grafico mostrato nella Figura 5.11 e la Tabella 5.12.

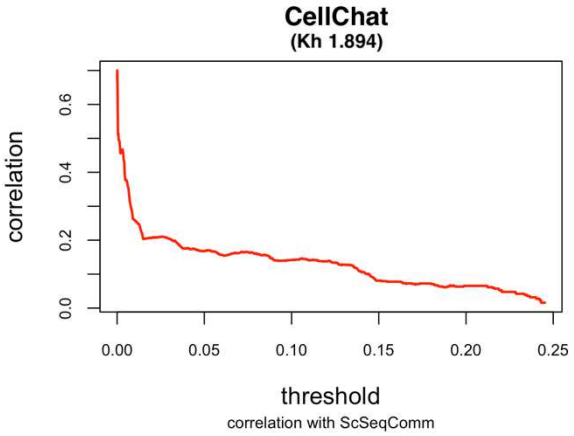


Figura 5.11 grafico dell'andamento della correlazione (MCC) tra CellChat (Kh 1.894) e ScSeqComm in funzione del livello di soglia

	correlations	thresholds	counter
1	0.6999793	0.0001	6195
2	0.5153000	0.0007	3667
3	0.4947554	0.0010	3408
4	0.4608937	0.0023	2704
5	0.4638355	0.0026	2595
6	0.4660172	0.0029	2486
7	0.4672384	0.0031	2407
8	0.3783625	0.0048	1539
9	0.3782333	0.0050	1515
10	0.3757823	0.0054	1450
11	0.2636310	0.0091	870
12	0.2613736	0.0095	852

Tabella 5.12 tabella con i valori di correlazione (MCC) tra CellChat (Kh 1.894) e ScSeqComm e il numero di coppie segnalate come attive in funzione del livello di soglia

Ponendo attenzione al grafico della Figura 5.11 e mettendolo a confronto con quello della Figura 5.9, si osserva un andamento molto simile, con la presenza però di qualche piccola diversità; il grafico di *CellChat (Kh 1.894)* risulta infatti sia leggermente spostato verso l'alto, con livelli di correlazione mediamente più alti a parità di soglia, sia più regolare nella sua decrescita, mostrando di avere quindi una distribuzione più omogenea nel range di valori rispetto alla sua controparte.

In questo caso, inoltre, a differenza dei precedenti, si è reso necessario andare a cambiare un parametro dato in input alla funzione, ovvero il tasso di aumento del valore di threshold; esso, infatti, è stato cambiato da 0.001 a 0.0001 a causa dei valori significativamente ridotti dello score intercellulare, i quali non permettevano, con il precedente rate, di individuare correttamente i picchi di correlazione. Questo ha inevitabilmente comportato un aumento delle coppie soglia – correlazione comprese nel range di accettabilità definito in partenza, come si può vedere osservando la Tabella 5.12; tuttavia, sono stati scartati tutti i valori di soglia dal quarto al settimo, poiché essi, oltre ad allontanarsi dal numero di coppie attive dello score di riferimento (3519), hanno anche livelli di correlazione via via decrescenti.

Le uniche due soglie rimaste che soddisfano entrambi i criteri scelti per l'individuazione della soglia sono 0.0007 e 0.001, entrambe molto vicine al numero di riferimento e con gradi di correlazione abbastanza simili; per motivi sia di concordanza con le precedenti soglie individuate, definite su tre cifre dopo la virgola, sia di conservatività dello score, è stato scelto come threshold di riferimento 0.001.

5.1.2 Soglia NATMI score

I punteggi proposti all'interno del tool *NATMI* sono sia differenti rispetto a quelli di *CellChat*, sia differenti tra di loro, soprattutto per quanto riguarda il range di valori; è stato pertanto necessario adeguare di conseguenza il valore di *rate* dato in input alla funzione *compute_threshold_value*, per evitare sia di avere troppi picchi di correlazione con differenze poco significative, sia di allungare eccessivamente i tempi macchina necessari per la loro computazione.

Il primo score *NATMI* preso in considerazione per la valutazione della soglia di riferimento è *NATMI_MEW*, ovvero il quarto score tra quelli esaminati in questa analisi; in questo caso la funzione, utilizzata inserendo in input un valore di *rate* pari a 0.1, restituisce in output il grafico mostrato nella Figura 5.13 e la relativa Tabella 5.14.

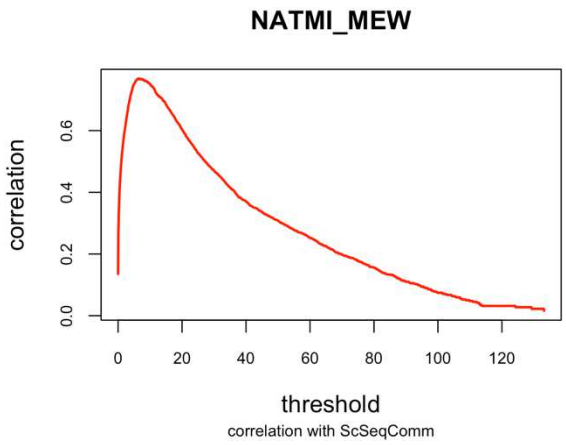


Figura 5.13 grafico dell'andamento della correlazione (MCC) tra NATMI_MEW e scSeqComm in funzione del livello di soglia

	correlations	thresholds	counter
1	0.7685120	6.2	4714
2	0.7683706	6.7	4487
3	0.7674135	7.1	4309
4	0.7658936	7.8	4026
5	0.7627338	8.3	3817
6	0.7598621	8.9	3618
7	0.7590928	9.1	3557
8	0.7573997	9.3	3479
9	0.7538766	9.7	3350
10	0.7440336	10.6	3072
11	0.7158107	12.4	2599
12	0.7110116	12.9	2466

Tabella 5.14 tabella con i valori di correlazione (MCC) tra NATMI_MEW e scSeqComm e il numero di coppie segnalate come attive in funzione del livello di soglia

Il grafico della Figura 5.13 appare notevolmente differente rispetto a quelli precedenti, andando ad evidenziare come il grado di correlazione presente tra *NATMI_MEW* e *scSeqComm* sia molto elevato, soprattutto se comparato con i gradi di correlazione molto bassi dei punteggi *CellChat*; si nota infatti un andamento che vede valori di correlazione molto bassi per livelli di soglia tendenti a zero, i quali poi aumentano velocemente e in modo significativo fino a raggiungere un picco massimo di quasi 0.8, per poi decrescere lentamente fino a tendere nuovamente a zero.

Tutti i valori all'interno della Tabella 5.14 soddisfano i criteri adottati in questa analisi e ciò implica la necessità di compiere ulteriori valutazioni; il picco effettivo del grafico, corrispondente al massimo grado di correlazione, potrebbe essere considerato come il migliore seguendo esclusivamente il criterio del grado di correlazione, se non fosse che i picchi successivi variano in modo relativamente ridotto il valore di correlazione, rendendo quindi le considerazioni riguardanti tale criterio poco rilevanti; risulta quindi più utile basarsi, per la scelta del valore di soglia, sul secondo criterio, ovvero sul numero di coppie ligando – recettore segnalate come attive. Dato che tale criterio suggerisce di scegliere livelli di soglia che restituiscono un numero di coppie attive prossimo il più possibile a 3519, la scelta del valore di threshold ricade su due possibili opzioni, più o meno equidistanti dal valore di riferimento, ovvero 9.1 e 9.3; tale scelta non risulta essere eccessivamente rilevante, dato che le differenze in relazione ai valori in gioco sono minime; pertanto, al solo fine di rimanere coerenti con le decisioni prese in precedenza, si è presa come soglia di riferimento per il quarto punteggio quella che rende il punteggio maggiormente conservativo, ovvero 9.3.

Il secondo score *NATMI* preso in considerazione è il quinto punteggio, ovvero *NATMI_SW*; a differenza del precedente score questo si mantiene maggiormente in linea con i punteggi *CellChat*, in quanto, dopo diverse valutazioni, il tasso di crescita del threshold è stato settato a 0.001, valore che si è rivelato essere un buon compromesso tra precisione e velocità di computazione; dall'utilizzo della funzione *compute_threshold_value* con il parametro *rate* così impostato, si sono ottenuti in output il grafico della Figura 5.15 e l Tabella 5.16.

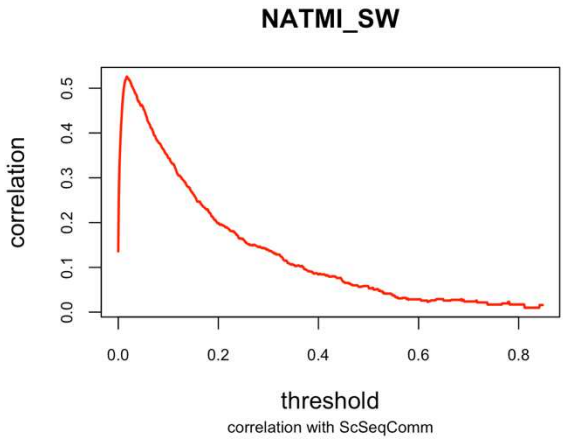


Figura 5.15 grafico dell'andamento della correlazione (MCC) tra NATMI_SW e scSeqComm in funzione del livello di soglia

	correlations	thresholds	counter
1	0.52581321	0.017	8156
2	0.52408034	0.019	7604
3	0.47117209	0.041	4077
4	0.46278485	0.047	3583
5	0.33021943	0.111	1273
6	0.30475866	0.123	1071
7	0.27930074	0.141	839
8	0.24702275	0.161	661
9	0.24701771	0.163	653
10	0.23011532	0.177	560
11	0.22936945	0.179	552
12	0.22204163	0.183	524

Tabella 5.16 tabella con i valori di correlazione (MCC) tra NATMI_SW e scSeqComm e il numero di coppie segnalate come attive in funzione del livello di soglia

Si può notare, osservando la Figura 5.15, come il grafico, se si esclude il fatto che il grado di correlazione sia mediamente inferiore, risulti essere nella forma abbastanza sovrapponibile a quello precedente, sintomo anche in questo caso di una elevata affinità con lo score *scSeqComm*; sono però presenti alcune differenze, che si ripercuotono anche sui dati presenti all'interno delle tabelle, come ad esempio il fatto che il picco principale sia molto più sottile e poco arrotondato, cosa che si traduce in un minor numero di soglie nel range di accettabilità.
Analizzando la tabella, infatti, si osserva la presenza di solo due livelli di threshold che rientrano nei parametri stabiliti in principio per quanto riguarda il numero di coppie ligando – recettore segnalate come attive, ovvero 0.047 e 0.041; come nelle casistiche precedenti anche in questo caso, data la debole differenza nel grado di correlazione tra le due soglie, si è preferito privilegiare il valore di threshold che si avvicina maggiormente al numero di coppie attive di riferimento e che in questo caso risulta anche essere quello maggiormente conservativo.
La soglia di riferimento per la conversione in binario del quinto punteggio è quindi stata fissata al valore di 0.047.

5.1.3 Soglia Sopt_SC score

Il sesto e ultimo punteggio preso in analisi, anch'esso continuo e quindi con necessità di individuare una soglia adeguata per la conversione in binario, è *Sopt_SC*; esso, data la formula con cui è stato implementato, risulta essere uno score i cui valori spaziano nel range che va da 0 a 1; il tasso di incremento del livello di soglia più adeguato è pertanto risultato essere, come nella maggior parte degli altri punteggi, 0.001; fissato il *rate* a tale valore è quindi stata utilizzata la funzione *compute_threshold_value* in modo da ottenere il grafico visibile nella Figura 5.17 e la Tabella 5.18; a differenza dei casi precedenti, a causa della particolare natura dello score, i picchi che maggiormente aderiscono ai criteri scelti nel sottocapitolo 5.1 e che quindi sono stati riportati nella Tabella 5.18 non sono i primi dodici, bensì gli ultimi dodici dei ventidue totali individuati dalla funzione *findpeaks()*.

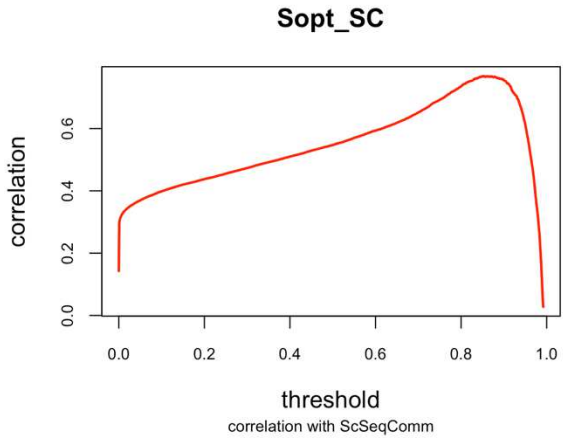


Figura 5.17 grafico dell'andamento della correlazione (MCC) tra Sopt_SC e scSeqComm in funzione del livello di soglia

	correlations	thresholds	counter
11	0.7679933	0.863	4449
12	0.7669678	0.866	4387
13	0.7675519	0.869	4305
14	0.7671155	0.873	4209
15	0.7655252	0.876	4135
16	0.7664671	0.878	4071
17	0.7626941	0.887	3802
18	0.7607212	0.890	3714
19	0.7601753	0.893	3628
20	0.7575902	0.898	3480
21	0.7455769	0.909	3108
22	0.7153467	0.923	2582

Tabella 5.18 tabella con i valori di correlazione (MCC) tra Sopt_SC e scSeqComm e il numero di coppie segnalate come attive in funzione del livello di soglia

Il grafico del sesto punteggio (Figura 5.17) risulta essere abbastanza differente rispetto a quelli degli score proposti nei primi due articoli; se infatti da un lato, come gli score *NATMI*, esso presenta una relativamente bassa correlazione in corrispondenza di livelli di threshold prossimi a zero, dall'altro in questo grafico il picco principale risulta essere decisamente più spostato a destra dei precedenti, probabilmente a causa della diversa distribuzione dei valori all'interno del range dello score; nonostante questa differenza, il picco principale di correlazione è in realtà molto simile a quello del punteggio *NATMI_MEW*, si presenta cioè molto largo e raggiunge un grado di correlazione massimo molto elevato.

In particolare, la prima di queste due somiglianze si ripercuote sui valori della tabella (Tabella 5.18), i quali risultano tutti all'interno dell'intervallo di valori di riferimento; ad eccezione dell'ultimo picco, in cui il livello di correlazione è inferiore di qualche centesimo, in tutti gli altri picchi il grado di correlazione con *scSeqComm* è abbastanza omogeneo; il criterio di scelta del livello di soglia è quindi ricaduto sul numero comunicazioni cellulari segnalate come 'in corso'.

I due threshold che maggiormente si avvicinano al valore di riferimento, pari a 3519, sono 0.893 e 0.898 e sono da considerare entrambi livelli di soglia molto validi; si è pertanto optato per la soglia di 0.898 esclusivamente per seguire il medesimo principio di conservatività adottato in precedenza.

5.2 Confronto tra i punteggi binari

Una volta selezionate le soglie che meglio soddisfavano i criteri proposti nel sottocapitolo 5.1, è stato possibile effettuare la conversione dei punteggi continui nelle loro versioni binarie; al fine di effettuare questa operazione è stata implementata una semplice funzione, denominata *compute_binary_conversion*, la quale permette di automatizzare il processo di conversione degli score.

```
compute_binary_conversion <- function(score,threshold){

  score_bin <- score
  score_bin[score_bin <= threshold] <- 0
  score_bin[score_bin > threshold] <- 1


  return(score_bin)
}
```

Figura 5.19 funzione compute_binary_conversion

La funzione (Figura 5.19) riceve in input lo score nella sua versione continua e la soglia da utilizzare come riferimento per la conversione in binario; sfruttando questi due parametri converte tutti i numeri minori o uguali alla soglia in 0 e tutti i valori maggiori di essa in 1. Successivamente restituisce in output la versione binaria dello score così creata.
Dopo aver ottenuto in tal modo i punteggi binari, è necessario effettuare dei confronti tra gli stessi al fine di valutarne la validità e l'affinità reciproca; la funzione più adatta a questo scopo è sicuramente la funzione *table()*, la quale restituisce in output una tabella dello stesso tipo di quella mostrata nella Figura 5.20.

0	0	25793
1	0	1695
0	1	1592
1	1	1888

Figura 5.20 esempio tabella data in output dalla funzione table()

In tale tabella viene riportato, per ogni possibile combinazione di 0 e 1 dei due rispettivi score, il numero di associazioni ligando – recettore che la rispecchiano; in questo modo per ogni coppia di punteggi posso conoscere la percentuale di concordanza nelle loro predizioni e stabilire così un valore di affinità.

Come però è possibile notare da questo esempio, il numero di coppie ligando – recettore segnalate come non in comunicazione da entrambi, ovvero la combinazione 00, è in generale decisamente più grande delle altre tre combinazioni, in linea con quanto atteso dalla biologia.

Nelle successive analisi comparative tra punteggi binari, oltre a considerare i risultati globali sopra citati, si è anche eseguita una analisi sui soli risultati in cui almeno uno score era in disaccordo con gli altri. Questa seconda analisi, in cui di fatto non vengono considerate le coppie ligando-recettore in cui i punteggi di tutti gli otto score sono in accordo (22208 coppie rilevate non attive, e 146 rilevate attive, da tutti gli score, per un totale del 72% delle coppie totali), ha lo scopo di evidenziare i risultati in cui vi è ambiguità per almeno uno score.

Al fine di valutare il livello di affinità presente tra i diversi score, sono stati scelti due parametri che forniscono, a partire dalla tabella di riferimento, un unico valore numerico, grazie al quale è quindi possibile fare confronti più facilmente; i due parametri scelti sono i seguenti:

- la percentuale di accordo, ovvero il rapporto tra il numero di casi in cui le due previsioni risultano concordi e il numero totale di previsioni, espresso in percentuale;
- il coefficiente di similarità di Jaccard, definito nel modo seguente:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

con A e B insiemi dei valori positivi dei due insiemi di dati, che si traduce in questo caso specifico nella formula:

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

dove con M si indica l'insieme delle ricorrenze all'interno della tabella della combinazione posta come pedice.

Per quanto riguarda la percentuale di accordo, indicata d'ora in avanti con il simbolo P, essa sarà calcolata sulla tabella modificata, dato che in questo modo le percentuali ottenute nei diversi casi risultano più in linea con l'effettiva affinità tra gli score; il coefficiente di similarità di Jaccard sarà invece applicato sulla tabella originale, dato che in ogni caso le 22208 coppie

segnalate da tutti gli score come non attive, non sono coinvolte nella formula (non è presente M_{00}).

5.2.1 Affinità dei punteggi CellChat con scSeqComm

In linea con le analisi precedenti i primi punteggi ad essere posti a confronto con lo score di riferimento sono le quattro versioni del punteggio *CellChat* proposto nel primo articolo; in particolare il primo ad essere analizzato è *semplified_CellChat (Kh 0.5)*, la cui tabella di confronto con *scSeqComm*, generata grazie alla funzione *table(),* è riportata di seguito nelle sue due versioni: la prima (Tabella 5.21) comprendente tutte e 30968 le coppie ligando – recettore, la seconda (Tabella 5.22) privata delle 22354 coppie identificate allo stesso modo da tutti gli score.

semplified_CellChat (Kh 0.5)

scSeqComm	*1*	*0*
1	562	2957
0	2606	24843

Tabella 5.21 tabella affinità semplified_CellChat (Kh 0.5) e scSeqComm (versione originale)

semplified_CellChat (Kh 0.5)

scSeqComm	*1*	*0*
1	416	2957
0	2606	2635

Tabella 5.22 tabella affinità semplified_CellChat (Kh 0.5) e scSeqComm (versione filtrata)

Come è ben evidente sia dalla Tabella 5.21 che dalla Tabella 5.22, il numero di coppie segnalate in modo discorde da *scSeqComm* e *semplified_CellChat (Kh 0.5*), tenendo conto che entrambi i punteggi segnalano all'incirca lo stesso numero di associazioni ligando – recettore come attive, è decisamente molto alto; tale discordanza risulta maggiormente enfatizzata nella Tabella 5.22, in cui sono state rimosse le coppie segnalate allo stesso modo da tutti gli otto score; infatti, se si va a calcolare il numero di associazioni concordi sulle 8614 rimaste, otteniamo una percentuale di accordo P pari a circa il 35.4%.
Se invece si prende in considerazione l'indice di Jaccard, considerando che il totale di coppie segnalate come attive da almeno uno dei due score è pari a 6125, mentre il numero associazioni ligando – recettore identificate come attive da entrambi è uguale a solo 562, si ottiene un valore molto basso, pari a circa 0.092.
Per quanto riguarda invece la versione del secondo score con il parametro *Kh* posto a 1.894, il confronto con *scSeqComm* è mostrato nelle Tabelle 5.23 e 5.24.

		semplified_CellChat (Kh 1.894)	
		1	*0*
scSeqComm	*1*	956	2563
	0	2643	24806

Tabella 5.23 tabella affinità semplified_CellChat (Kh 1.894) e scSeqComm (versione originale)

		semplified_CellChat (Kh 1.894)	
		1	*0*
scSeqComm	*1*	810	2563
	0	2643	2598

Tabella 5.24 tabella affinità semplified_CellChat (Kh 1.894) e scSeqComm (versione filtrata)

Nuovamente, come nel caso precedente, le Tabelle 5.23 e 5.24 confermano la presenza di una scarsa affinità tra i punteggi *scSeqComm* e *semplified_CellChat (Kh 1.894)*, dovuta al fatto che i due score tendono a segnalare entrambi come attive solo una piccola parte delle coppie considerate; se infatti si calcola la percentuale di accordo P tra i due score, si ottiene circa il 39.6%, che, pur essendo leggermente più alta rispetto alla versione del punteggio con *Kh* posto a 0.5, resta comunque molto bassa.

Anche in questo caso è utile calcolare anche il coefficiente di similarità di Jaccard, il quale è pari a 0.155; come si può notare, esso, seppur si mantenga su valori decisamente molto bassi, risulta essere abbastanza più alto rispetto all'altra versione del punteggio *semplified_CellChat*, andando a confermare che la versione con *Kh* posto a 1.894 risulta essere più affine con lo score di riferimento.

Passando invece ai punteggi, sempre tratti dal tool *CellChat*, ma che sfruttano l'espressione media di insieme, come prima cosa è stata analizzata la versione con il parametro *Kh* posto a 0.5; confrontando tale punteggio con *scSeqComm* si sono ottenute le Tabelle 5.25 e 5.26.

		CellChat (Kh 0.5)	
		1	*0*
scSeqComm	*1*	1608	1911
	0	1331	26118

Tabella 5.25 tabella affinità CellChat (Kh 0.5) e scSeqComm (versione originale)

		CellChat (Kh 0.5)	
		1	*0*
scSeqComm	*1*	1492	1991
	0	1331	3910

Tabella 5.26 tabella affinità CellChat (Kh 0.5) e scSeqComm (versione filtrata)

Ponendo a confronto le tabelle 5.25 e 5.26 con le tabelle della versione semplificata dello score analizzate in precedenza (Tabelle 5.21, 5.22, 5.23, 5.24), è subito evidente che esse sono molto differenti, soprattutto per quanto riguarda il numero di coppie ligando – recettore segnalate

come ‘non in comunicazione’, decisamente presenti in numero inferiore in questa versione del punteggio. Infatti, seppur non si raggiungano livelli di affinità elevatissimi, sicuramente vi è una notevole differenza rispetto alle percentuali di accordo precedenti, passando da valori del 35-40% ad una percentuale di accordo di quasi il 62.4%.
Per quanto riguarda il coefficiente di similarità di Jaccard, si nota anche in esso un andamento che va ad evidenziare una maggiore affinità di *CellChat (Kh 0.5)* con *scSeqComm* rispetto alle due versioni semplificate; tale coefficiente J è infatti pari in questo caso a 0.332, più del doppio di *semplified_CellChat (Kh 1.894)* e quasi quattro volte rispetto a *semplified_CellChat (Kh 0.5)*.
L’ultimo dei quattro punteggi di questo tool è *CellChat (Kh 1.894)*; anche con esso è stata utilizzata la funzione *table()* e i dati ottenuti sono raccolti nelle Tabelle 5.27 e 5.28.

		CellChat (Kh 1.894)	
		1	*0*
scSeqComm	*1*	1909	1610
	0	1499	25950

Tabella 5.27 tabella affinità CellChat (Kh 1.894) e scSeqComm (versione originale)

		CellChat (Kh 1.894)	
		1	*0*
scSeqComm	*1*	1763	1610
	0	1499	3742

Tabella 5.28 tabella affinità CellChat (Kh 1.894) e scSeqComm (versione filtrata)

Questo score si dimostra abbastanza in linea con l’andamento della sua controparte con diverso parametro *Kh*; infatti anch’esso presenta una maggiore affinità rispetto alle versioni semplificate ed è abbastanza simile a *CellChat (Kh 0.5)*, con una percentuale di accordo pari al 63.9%, superiore di appena un 1.5% rispetto al precedente; se infatti da un lato il numero di associazioni indicate come attive da entrambi sono aumentate di 301 unità, allo stesso tempo si sono ridotte di 198 quelle segnalate da entrambi come inattive, compensandosi parzialmente.
La differenza del coefficiente di similarità di Jaccard è invece relativamente più rilevante rispetto a quella della percentuale di accordo; essendoci infatti un totale di coppie ligando – recettore segnalate come attive da almeno uno dei due score pari a 5018 e un numero di associazioni indicate da entrambi i punteggi come in comunicazione pari a 1909, si ottiene un coefficiente di Jaccard di 0.380.

5.2.2 Affinità dei punteggi NATMI con scSeqComm

I due punteggi appartenenti al tool *NATMI* sono risultati essere, nelle loro versioni continue, decisamente più affini con *scSeqComm* rispetto alle quattro versioni dello score *CellChat*; non è tuttavia scontato che tale risultato si ripeta una volta convertiti i diversi punteggi in forma binaria; pertanto, sono di seguito riportate le analisi di *NATMI_MEW* e *NATMI_SW*, nell'ordine già seguito nelle versioni continue.
Le Tabelle 5.29 e 5.30 si riferiscono quindi al quarto punteggio, ovvero *NATMI_MEW*.

		NATMI_MEW	
		1	*0*
scSeqComm	*1*	2746	773
	0	733	26716

Tabella 5.29 tabella affinità NATMI_MEW e scSeqComm (versione originale)

		NATMI_MEW	
		1	*0*
scSeqComm	*1*	2600	773
	0	733	4508

Tabella 5.30 tabella affinità NATMI_MEW e scSeqComm (versione filtrata)

In linea con le osservazioni precedenti il punteggio presenta un'alta affinità con lo score di riferimento, avendo un numero più che dimezzato di previsioni discordanti con *scSeqComm* rispetto al migliore dei punteggi *CellChat*; osservando la tabella di destra si possono infatti contare ben 7108 occorrenze delle combinazioni 00 e 11, contro le sole 1506 di 01 e 10, dati dai quali si ottiene facilmente la percentuale di accordo tra i due punteggi, pari a poco più dell'82.5%.
Da notare come se si fosse effettuata l'analisi di questo primo parametro facendo riferimento alle tabelle nella loro versione originale, in questo caso si sarebbe ottenuta una percentuale pari al 95%, la quale si sarebbe comunque mantenuta maggiore rispetto all'82% di grado di concordanza di *semplified_CellChat (Kh 0.5)*, ma non avrebbe reso allo stesso modo la differenza effettivamente esistente tra i due score (82.5% contro appena il 35.4%).
Anche il coefficiente di similarità di Jaccard contribuisce a confermare lo score *NATMI_MEW* come il più affine con *scSeqComm*, attestandosi infatti intorno al valore di 0.646, tale valore è di gran lunga maggiore rispetto a tutti quelli finora riscontrati e suggerisce una similarità elevata tra i due score.
Per quanto riguarda invece i dati relativi al quinto punteggio, ovvero *NATMI_SW*, essi sono contenuti all'interno delle Tabelle 5.31 e 5.32.

		NATMI_SW	
		1	*0*
scSeqComm	*1*	1862	1657
	0	1721	25728

Tabella 5.31 tabella affinità NATMI_SW e scSeqComm (versione originale)

		NATMI_SW	
		1	*0*
scSeqComm	*1*	1716	1657
	0	1721	3520

Tabella 5.32 tabella affinità NATMI_SW e scSeqComm (versione filtrata)

In questo caso è possibile notare una discrepanza con quanto osservato per i punteggi continui; *NATMI_SW*, infatti, era risultato essere significativamente più affine rispetto alle diverse versioni dello score *CellChat* con il punteggio di riferimento, cosa non più vera dopo la trasformazione in binario; anche se non risulta evidente dalle tabelle, molto simili ad esempio ai due score *CellChat (Kh 0.5)* e *CellChat (Kh 1.894)*, se si effettua un calcolo della percentuale di accordo con *scSeqComm* si ottiene un valore pari a 60.8%, leggermente inferiore ai due punteggi appena citati.
La situazione è leggermente diversa per quanto riguarda il coefficiente di similarità di Jaccard; esso, infatti, in questo caso è pari a 0.337, ancora inferiore di qualche centesimo rispetto a *CellChat (Kh 1.894)*, il cui coefficiente vale 0.380, ma assume un valore leggermente superiore rispetto al coefficiente di *CellChat (Kh 0.5)*, pari a 0.332.

5.2.3 Affinità del punteggio SoptSC con scSeqComm

L'ultimo score dell'analisi, ovvero *Sopt_SC*, era risultato, nelle valutazioni effettuate con le versioni continue dei punteggi, molto affine con *scSeqComm*, secondo in graduatoria solo al primo dei punteggi proposti da *NATMI*, ovvero *NATMI_MW*; al fine di verificare la presenza di questa tendenza anche nella versione binaria, sono riportati nelle Tabelle 5.33 e 5.34 i dati relativi a tale score.

		Sopt_SC	
		1	*0*
scSeqComm	*1*	2747	772
	0	733	26716

Tabella 5.33 tabella affinità Sopt_SC e scSeqComm (versione originale)

		Sopt_SC	
		1	*0*
scSeqComm	*1*	2601	772
	0	733	4508

Tabella 5.34 tabella affinità Sopt_SC e scSeqComm (versione filtrata)

Osservando le Tabelle 5.33 e 5.34 è evidente, anche senza il calcolo della percentuale di accordo, che quest'ultima sarà abbastanza alta rispetto alla media degli altri score, in linea con i risultati ottenuti con la sua versione continua; la tabella di destra mostra come vi siano 7109 occorrenze delle combinazioni 00 e 11 a fronte di un totale di 8614 casi; da questi valori è possibile ottenere una percentuale di accordo uguale all'incirca all'82.5%, che è in realtà decisamente più alta rispetto a quella della sua controparte continua, arrivando a pareggiare la percentuale dello score più affine in assoluto con *scSeqComm*, ovvero *NATMI_MEW*.
L'alta similarità tra *Sopt_SC* e *scSeqComm* è confermata anche dal calcolo del coefficiente di Jaccard, il quale è pari a 0.646, anche in questo caso uguale al coefficiente di *NATMI_MEW*.

Conclusioni

Negli ultimi anni, grazie all'introduzione della nuova tecnica di sequenziamento dell'RNA a singola cellula (sc-RNA seq), sono stati fatti numerosi progressi nell'analisi delle comunicazioni intercellulari; in particolare, grazie all'impiego di metodi bioinformatici in grado di utilizzare i livelli di espressione genica di particolari molecole, in genere coinvolte nella comunicazione cellulare, per ottenere degli score che stimino l'evidenza di tale comunicazione, è possibile andare ad approfondire, a seconda delle caratteristiche del metodo, diversi aspetti biologici di interesse.

Il principale obiettivo del lavoro svolto all'interno di questa tesi è quello di fornire degli strumenti aggiuntivi per l'analisi delle comunicazioni intercellulari, unendo al già esistente tool informatico *scSeqComm*, una serie di altri score intercellulari, così da garantire all'utente che lo utilizza di avere a disposizione un'ampia gamma di possibilità tra cui scegliere, in base al principale aspetto della comunicazione sul quale vuole focalizzarsi.

Al fine di raggiungere tale obiettivo è stato in primis necessario trovare degli score adatti a tale scopo dalla letteratura scientifica, che in questo caso sono stati selezionati da tre diversi tool, ovvero *CellChat*, *NATMI* e *SoptSC*; da essi sono stati ricavati un totale di cinque score, che, dopo essere stati opportunamente modificati, sono stati implementati in R e integrati nel programma principale di *scSeqComm*; per effettuare tale operazione si è però reso necessario non solo apprendere un nuovo linguaggio di programmazione, ma anche comprendere ed analizzare il codice già scritto, con particolare riguardo per lo score già implementato.

Una volta realizzate le funzioni in grado di calcolare i diversi score è stato scelto il set di dati Tirosh, disponibile online, sia per verificarne il corretto funzionamento che per effettuare dei confronti tra gli score implementati e lo score di riferimento, già presente all'interno del codice; tali confronti avevano come obiettivo principale andare a studiare il livello di affinità presente tra i diversi score ed analizzarne le capacità predittive e sono stati operati sia tra le versioni originali continue, sia tra le versioni binarie ottenute mediante l'utilizzo di opportuni livelli di soglia, selezionati secondo criteri consoni con lo scopo del confronto.

Dai confronti effettuati lo score di riferimento *scSeqComm* è risultato essere scarsamente affine con entrambi gli score ricavati a partire dal tool *CellChat*, sia nelle loro versioni continue che in quelle binarie; i due score ottenuti da *NATMI* invece, seppur entrambi risultino essere in generale maggiormente correlati con *scSeqComm* rispetto ai precedenti, hanno un comportamento leggermente diverso; il primo, ovvero *NATMI_MEW*, presenta un alto grado di affinità con lo score di riferimento in entrambe le versioni, mentre il secondo, *NATMI_SW*,

presenta un livello di affinità medio nella sua versione continua e addirittura minore dello score *CellChat* in quella binaria. L'ultimo score, facente in questo caso riferimento al tool *SoptSC*, manifesta infine una correlazione medio – alta nella sua versione continua e alta nella sua versione binaria.

Grazie all'elaborazione condotta all'interno di questa tesi, è stato possibile ottimizzare e ampliare il tool scSeqComm, al fine di garantire agli utilizzatori una selezione più ampia di opzioni. Tuttavia, è opportuno notare che vi è spazio per ulteriori miglioramenti futuri; ad esempio, si potrebbe considerare l'inclusione di ulteriori metriche di valutazione all'interno del tool o l'impiego di dataset diversi per condurre valutazioni comparative con quelli già impiegati. Inoltre, si potrebbe esaminare l'opportunità di integrare i risultati ottenuti tramite gli score intercellulari con le informazioni prodotte da altre analisi, al fine di conferire maggiore robustezza alle conclusioni raggiunte e di arricchire la comprensione dei processi biologici sottostanti alla comunicazione intercellulare.

Bibliografia

[1] Silverthorn Dee. U., *'Fisiologia umana. Un approccio integrato'*, Ottava Edizione, Pearson, 158-173, 2020.

[2] Hwang, B., Lee, J.H. & Bang, D., '*Single-cell RNA sequencing technologies and bioinformatics pipelines'*, Exp Mol Med 50, 1-14, 2018

[3] Armingol, E., Officer, A., Harismendy, O. et al., 'Deciphering *cell–cell interactions and communication from gene expression*', Nat Rev Genet 22, 71–88, 2021

[4] Browaeys R. et al., '*NicheNet: modeling intercellular communication by linking ligands to target genes'*, Nat. Methods, 17, 159–162, 2020.

[5] Cheng J. et al., '*Inferring microenvironmental regulation of gene expression from single-cell RNA sequencing data using scMLnet with an application to COVID-19*', Brief Bioinform., 22, 988–1005, 2021.

[6] Hu Y. et al., '*CytoTalk: de novo construction of signal transduction networks using single-cell transcriptomic data*', Sci. Adv., 7, eabf1356, 2021.

[7] Baruzzo G., Cesaro G., Di Camillo B., '*Identify, quantify and characterize cellular communication from single-cell RNA sequencing data with scSeqComm*', Bioinformatics, Volume 38, Issue 7, 1920–1929, 2022.

[8] Almet A.A. et al., '*The landscape of cell–cell communication through single-cell transcriptomics'*. Curr. Opin. Syst. Biol., 26, 12–23, 2021.

[9] Jin, S., Guerrero-Juarez, C.F., Zhang, L. et al., '*Inference and analysis of cell-cell communication using CellChat*'. Nat Commun 12, 1088, 2021.

[10] Hou, R., Denisenko, E., Ong, H.T. et al., '*Predicting cell-to-cell communication networks using NATMI'*, Nat Commun 11, 5011, 2020.

[11] Shuxiong Wang, Matthew Karikomi, Adam L MacLean, Qing Nie, '*Cell lineage and communication network inference via optimization for single-cell transcriptomics'*, Nucleic Acids Research, 47, 11, 2019.

[12] Tirosh I. et al., '*Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-Seq*'. Science, 352, 189–196, 2016.

[13] Kumar M.P. et al., '*Analysis of single-Cell RNA-Seq identifies cell-cell communication associated with tumor characteristics*', Cell Rep., 25, 1458–1468.e4, 2018.

[14] Ramilowski, Jordan A., et al. "*A draft network of ligand–receptor-mediated multicellular signalling in human.*" Nature communications, 6.1,7866, 2015.