



Universidad
Internacional
de Valencia

Análisis Comparativo de Software para la Inferencia de la Comunicación Célula-Célula en Datos de Secuenciación de ARN de Célula Única

Titulación:
Máster en Bioinformática

Curso académico
2023 – 2024

Alumno/a: Korchi Meziani,
Elías
D.N.I: XXXXXXXXXXXX

Director/a de TFM: Sheila
Zúñiga Trejos

Convocatoria:
Segunda

De:
 Planeta Formación y Universidades

Índice

1. Introducción	7
1.1. Comunicación célula-célula	7
1.2. Tecnologías actuales para el estudio de las CCCs.	8
1.3. Secuenciación del ARN en bulk (ARN-seq).....	9
1.4. Secuenciación de ARN de célula única	10
1.5. Software de inferencia de CCC a partir de datos scARN-seq	11
1.5.1. Enfoque pseudo-bulk	13
1.5.2. Enfoque de resolución a nivel celular.....	14
1.6. Justificación	16
2. Objetivos.....	17
2.1. Objetivo principal	17
2.2. Objetivos específicos.....	17
3. Metodología	17
3.1. Datos	17
3.2. Softwares.....	18
3.3. Métricas de evaluación	18
3.3.1. Precisión.....	18
3.3.2. Concordancia	20
3.3.3. Consistencia.....	21
3.3.4. Tiempo de ejecución y uso de memoria	21
4. Resultados.....	22
4.1. Precisión.....	23
4.1.1. Evaluación de precisión basada en datos de CAGE	23
4.1.2. Evaluación de precisión basada en datos proteómicos.....	28
4.2. Concordancia	32
4.3. Consistencia	34
4.4. Requisitos computacionales	37
4.4.1. Tiempo	37

4.4.2. Memoria	39
5. Discusión	41
5.1. Precisión en la predicción de expresión génica y proteica	41
5.2. Diferencias metodológicas.....	42
5.3. Variabilidad.....	43
5.4. Eficiencia computacional	43
6. Conclusiones	44
7. Limitaciones y perspectivas futuras	45
8. Referencias bibliográficas	46

Lista de abreviaturas

- (CCC): Comunicación célula-célula
- (scARN-seq): Secuenciación de ARN de célula única
- (LR): Ligando-receptor
- (mARN): ARN mensajero
- (ARN-seq): Secuenciación del ARN en bulk
- (MEG): Matriz de expresión génica
- (MICC): Matriz de interacción célula-célula
- (FANTOM5): Functional Annotation of the Mammalian Genome
- (PBMC): Célula mononuclear de sangre periférica
- (CAGE): Cap Analysis of Gene Expression
- (TPM): Transcritos por millón
- (AUPRC): Área Bajo la Curva de la Precisión-Recuperación
- (RIC) Rango Intercuartílico

Índice de figuras

Figura 1: Flujo de trabajo general de un software que infiere CCC a partir de datos de scARN-seq y principales funciones de puntuación de las vías de comunicación.....	12
Figura 2: Ejemplo de un flujo de trabajo de un software de inferencia CCC a partir de datos scARN-seq con enfoque de resolución celular	15
Figura 3: Diagramas de cajas que muestran la distribución de los valores medios de AUPRC de CellChat y NICHES obtenidos tras comparar las predicciones de las interacciones LR de los conjuntos de datos PBMC con la base de datos CAGE de FANTOM5	24
Figura 4: Diagramas de cajas, separados por conjunto de datos, que muestran la distribución de los valores de AUPRC CAGE de CellChat y NICHES tras comparar las predicciones de las interacciones LR de los conjuntos de datos PBMC con la base de datos CAGE de FANTOM5.....	26
Figura 5: Diagramas de cajas que muestran la distribución de los valores medios de AUPRC proteómico de CellChat y NICHES tras comparar las predicciones de las interacciones LR de los conjuntos de datos PBMC con la base de datos proteómica	29
Figura 6: Diagramas de cajas, separados por conjunto de datos, que muestran la distribución de los valores de AUPRC proteómico de CellChat y NICHES tras comparar las predicciones de las interacciones LR de los conjuntos de datos PBMC con la base de datos proteómica	31
Figura 7: Diagramas de cajas, separados por la proporción de conjunto de dato, analizado que ilustran la distribución de los índices de Jaccard de CellChat y NICHES tras la comparación de las predicciones de interacciones LR entre el conjunto de datos original y el conjunto de datos reducido.....	35
Figura 8: Gráficas de dispersión del tiempo de ejecución de CellChat y NICHES al analizar las diferentes proporciones de los conjuntos de datos	38
Figura 9: Gráficas de dispersión que ilustran el consumo de memoria de CellChat y NICHES durante la construcción de la MICC y la evaluación de las interacciones	40

Índice de tablas

Tabla 1: Valores de AUPRC CAGE del tercer cuartil, primer cuartil y RIC de las predicciones de expresión génica obtenidos por CellChat y NICHES tras analizar los tres conjuntos de datos de PBMC 3k, 6k, y 8k, y compararlas con la base de datos CAGE de FANTOM5...	25
Tabla 2: Valores de AUPRC CAGE del tercer cuartil, primer cuartil y RIC de las predicciones de expresión génica obtenidos por CellChat tras analizar los tres conjuntos de datos de PBMC 3k, 6k, y 8k. RIC y compararlas con la base de datos CAGE de FANTOM5	27
Tabla 3: Valores de AUPRC CAGE del tercer cuartil, primer cuartil y RIC de las predicciones de expresión génica obtenidos por NICHES tras analizar los tres conjuntos de datos de PBMC 3k, 6k, y 8k. RIC y compararlas con la base de datos CAGE de FANTOM5	28
Tabla 4: Valores de AUPRC proteómico del tercer cuartil, primer cuartil y RIC de las predicciones de la presencia proteica obtenidos por CellChat y NICHES tras analizar los tres conjuntos de datos de PBMC 3k, 6k, y 8k y compararlas con la base de datos proteómica	30
Tabla 5: Valores de AUPRC proteómico del tercer cuartil, primer cuartil y RIC de las predicciones de la presencia proteica obtenidas por CellChat tras analizar los tres conjuntos de datos de PBMC 3k, 6k, y 8k y compararlas con la base de datos proteómica	32
Tabla 6: Valores de AUPRC proteómico del tercer cuartil, primer cuartil y RIC de las predicciones de la presencia proteica obtenidas por NICHES tras analizar los tres conjuntos de datos de PBMC 3k, 6k, y 8k y compararlas con la base de datos proteómica	32
Tabla 7: Comparación entre los ligandos y receptores únicos detectados por los softwares NICHES y CellChat, junto con el índice de similitud entre ambos softwares según el conjunto de dato.....	33
Tabla 8: Comparación entre los ligandos y receptores únicos detectados por los softwares NICHES y CellChat tras analizar el conjunto de dato	33
Tabla 9: Valores medios del índice de Jaccard de CellChat calculados tras las reducciones correspondientes en la cantidad de células de los conjuntos de datos.....	36
Tabla 10: Valores medios del índice de Jaccard de NICHES calculados tras las reducciones correspondientes en la cantidad de células de los conjuntos de datos.....	37

Resumen y palabras clave

La comunicación célula-célula es un proceso fundamental que involucra la transmisión de señales entre células. Este intercambio es clave porque mantiene el correcto funcionamiento en los organismos multicelulares. La comprensión de la comunicación célula-célula en condiciones normales resulta fundamental para desentrañar los mecanismos que regulan la homeostasis, mientras que su estudio en contextos alterados es crucial para profundizar en el conocimiento de la patogénesis y el desarrollo de diversas enfermedades.

La tecnología de secuenciación de ARN de célula única es una herramienta poderosa para investigar la biología celular, ya que mide la expresión génica a una resolución de célula individual, permitiendo así analizar la comunicación célula-célula con gran detalle.

El uso de los softwares para desentrañar las comunicaciones celulares con estos datos se realiza a nivel pseudo-bulk, es decir las células individuales se clasifican en tipos celulares. El problema de este enfoque es la pérdida de información, ya que la comunicación célula-célula opera a nivel celular. Recientemente se han desarrollado paquetes para el análisis de comunicación célula-célula con resolución a nivel de una sola célula. Un ejemplo destacado de estas herramientas es NICHES. Para demostrar el potencial de las nuevas generaciones de software aplicado al estudio de las comunicaciones celulares, se evalúa la herramienta NICHES frente a un software clásico y bien establecido en la comunidad como CellChat. El presente trabajo recoge los resultados de la evaluación aplicando diferentes métricas y conjuntos de datos de prueba para analizar los aspectos positivos y negativos.

Palabras clave: Comunicación célula-célula, Secuenciación de ARN de célula única, Interacción ligando-receptor, NICHES, CellChat.

1. Introducción

1.1. Comunicación célula-célula

La comunicación célula-célula (CCC) es el proceso mediante el cual las células detectan y responden las señales de su entorno. En este proceso, las células secretoras producen y liberan moléculas de señalización, conocidas como ligandos, que pueden incluir hormonas, neurotransmisores, factores de crecimiento y citoquinas. Estos ligandos viajan hasta alcanzar las células diana, que pueden estar situadas a diferentes distancias de la célula emisora. Según la distancia y el mecanismo de acción, los eventos de la señalización en la CCC se pueden clasificar en autocrina, yuxtacrina, paracrina y endocrina. La señalización autocrina es una interacción donde los ligandos se relacionan con los receptores de la propia célula, por lo que tanto el emisor y el receptor se encuentran en la misma célula. La yuxtacrina es un tipo comunicación que requiere un contacto físico directo entre las células, ya que los ligandos permanecen adheridos a la superficie de la célula emisora y se unen a los receptores en la célula adyacente, permitiendo la transmisión de señales sin la necesidad de que los ligandos se liberen en el espacio extracelular. En la señalización paracrina, los ligandos se liberan al espacio extracelular y actúa sobre las células receptoras que se hallan cercanas. Finalmente, la señalización endocrina ocurre cuando las células dianas se encuentran lejanas a las células receptoras, y estas deben comunicarse usando el torrente sanguíneo para transmitir el ligando (1).

Llegado el ligando al receptor, se produce una serie de cascadas de señales intracelulares, comúnmente mediadas por segundos mensajeros. Estos segundos mensajeros son responsables de transmitir la señal desde el receptor hacia el núcleo celular (2). Durante la cascada de señales, se producen mecanismos de regulación que modulan la intensidad y duración de la señal. Asegurando así que la respuesta celular de las células dianas sea adecuada al estímulo recibido. La llegada de las señales intracelulares culmina en cambios en la expresión génica y, en consecuencia, en la síntesis de proteínas específicas (1). Estos cambios permiten a la célula adaptarse a las condiciones externas.

La existencia de los organismos multicelulares depende de la CCC que se produce entre las células del organismo para llevar a cabo de las actividades celulares esenciales. Esta CCC entre los diferentes tipos celulares coordina los procesos vitales de los organismos

multicelulares, como, por ejemplo, el mantenimiento de las funciones fisiológicas, la respuesta inmunitaria, el crecimiento, la homeostasis, la diferenciación celular y tisular, la apoptosis (3), y todos los procesos vitales que requiere la supervivencia del organismo (4). Se ha demostrado que las alteraciones en la CCC están asociadas con una serie de enfermedades, dado que las alteraciones en este proceso pueden afectar la función normal de las células y los tejidos. Uno de los casos más destacados dónde la CCC juega un papel crucial es el cáncer. Un ejemplo de este contexto es la involucración de la CCC en la iniciación y progresión del cáncer mediante los fibroblastos asociados al cáncer que modulan la metástasis tumoral, influyendo sobre la angiogénesis, contribuyendo a la evasión de las células cancerosas del sistema inmune y la resistencia a terapias (5). Además, la alteración de la CCC también tiene una relevancia significativa en las enfermedades autoinmunes. Por ejemplo, en la pancreatitis autoinmune, la no expresión de TGF- β conduce a una pérdida de tolerancia inmune en las células dendríticas mieloides (6). Asimismo, en las enfermedades metabólicas, la sobreexpresión de TNF- α en los tejidos adiposo y muscular está asociada con la inflamación crónica, que a su vez está relacionada con la obesidad, la resistencia a la insulina y la diabetes tipo 2 (7).

1.2. Tecnologías actuales para el estudio de las CCCs.

Actualmente, el estudio de las CCCs se puede enfocar desde diversos campos científicos, cada una aportando una perspectiva única sobre la estructura y la función de los complejos proteicos. En primer lugar, desde el ámbito de la biología estructural, se emplean técnicas como la cristalografía de rayos X y la criomicroscopía electrónica. La cristalografía de rayos X cristaliza las proteínas y las somete a rayos X para determinar su estructura tridimensional, mientras que la criomicroscopía electrónica congela rápidamente las muestras biológicas a temperaturas extremadamente bajas, preservando su estructura nativa (8). Estas técnicas permiten la visualización de complejos proteicos y estructuras celulares, facilitando el estudio de las interacciones ligando-receptor (LR).

Otro campo relevante para el estudio de las CCCs es la bioquímica. La espectrometría de masas es una técnica poderosa para identificar y caracterizar proteínas, así como sus

interacciones. Esta herramienta analiza la composición y la estructura de complejos proteicos, y permite crear redes de interacción. Proporciona información detallada sobre las interacciones físicas entre proteínas en células vivas, permitiendo la caracterización de complejos proteicos y la validación de interacciones observadas (8).

La proteómica, por su parte, ha sido tradicionalmente la disciplina de la biología moderna en la que más se han llevado a cabo experimentos para el estudio de las CCCs, debido a que las interacciones LR suelen ser interacciones proteína-proteína (1). Un ejemplo clásico es el sistema de dos híbridos en levadura, en el cual se fusionan proteínas sospechosas de interactuar con dos dominios proteicos: uno con un factor de unión al ADN y el otro con un factor de activación. Si las proteínas interactúan, los dominios se reconstituyen y activa la expresión de un gen reportero, indicando que la interacción ha ocurrido. Este sistema permite la detección y el análisis de redes de interacción proteica en un contexto biológico real. Otro ejemplo es la coimmunoprecipitación, una técnica que utiliza anticuerpos dirigidos a subunidades de un complejo proteico, si estas subunidades interactúan, su unión en solución provoca la precipitación del complejo (8).

Aunque estos métodos son preferibles debido a su capacidad para medir interacciones directas, pueden ser laboriosos y requieren condiciones experimentales específicas. En contraste, la secuenciación de ARN mensajero es capaz de proporcionar un enfoque más accesible, y con resultados fáciles de procesar (1). La técnica de la secuenciación de ARN mensajero permite investigar de manera indirecta estas interacciones a través de la medición de los niveles de los genes relacionados en miles de células. Además, esta tecnología permite obtener datos a partir de una amplia variedad de muestras, desde a granel, especímenes microdisecccionados hasta suspensiones de células individuales (1).

1.3. Secuenciación del ARN en bulk (ARN-seq)

Las moléculas de ARN mensajero (mARN) son componentes esenciales en todas las células vivas. Durante el proceso de transcripción, la información genética del ADN se transcribe en moléculas de mARN. Posteriormente, durante la traducción, estas moléculas de mARN son interpretadas por los ribosomas para sintetizar proteínas. El uso de la técnica

de ARN-seq para analizar la CCC se basa en la premisa de que, al identificar el número de transcritos se puede estimar la expresión génica de la muestra. La expresión génica, a su vez, refleja la abundancia de proteínas, siendo esta una relación no directa de las interacciones celulares. Esta información se considera suficiente para inferir la CCC a partir de mRNA (1). Para poder analizar el transcriptoma de la muestra, la técnica de ARN-seq aísla el mRNA, convierte el ARN en ADN complementario, prepara la biblioteca de secuenciación y secuencia en una plataforma de secuenciación (9). La limitación de este método es la obtención de un solo valor promedio para cada gen a partir de todas las células de la muestra. Esto significa que no se puede captar la heterogeneidad celular, ya que se obtiene una única medida generalizada de cada gen por muestra (10). Como resultado, esta tecnología no es capaz de captar la complejidad que hay en las interacciones celulares, mostrando un mapa incompleto (10). No obstante, la técnica de secuenciación de ARN de célula única (scARN-seq) sí que proporciona una información a una resolución de una sola célula.

1.4. Secuenciación de ARN de célula única

La tecnología de scARN-seq se distingue de la técnica de ARN-seq principalmente por el proceso de aislamiento y captura de células individuales, aunque el resto del procedimiento es similar. Con la información de mRNA a nivel unicelular, la tecnología de scARN-seq nos permite identificar patrones de expresión génica en células individuales, y con ello analizar con gran detalle la CCC. Esta capacidad es crucial para comprender la heterogeneidad celular, ya que, si el muestreo de células ha sido extenso revela todos los tipos de células presentes, incluidos los nuevos.

El análisis de la CCC a partir de datos scARN-seq se ha utilizado para comprender el proceso evolutivo de la formación de tumores, así como la adquisición de resistencia en los tratamientos de quimioterapia, una mejor comprensión del desarrollo de las células β y la patología en la diabetes (3) o descubrir células inmunes hiper-responsivas dentro de un grupo aparentemente homogéneo (11).

1.5. Software de inferencia de CCC a partir de datos scARN-seq

La mayoría de las inferencias de CCC se pueden conceptualizar como una forma de análisis de coexpresión, en el que el patrón de expresión de un ligando de una célula se empareja con un receptor de otra célula (1). Matemáticamente, la CCC se representa como una función de la expresión del ligando en la célula emisora y del receptor en la célula receptora, utilizando estos niveles de expresión para calcular una puntuación para cada par LR mediante una función de puntuación de comunicación. (Figura 1a). Los puntajes de comunicación pueden ser binarios o continuos. Los métodos de umbral de expresión y combinaciones diferenciales se aplican a puntajes binarios, mientras que los enfoques de producto de expresión y correlación de expresión cuantifican puntuaciones continuas. (Figura 1b). La medición de las puntuaciones individuales de comunicación expone los mecanismos de señalización específicos, sin embargo, a veces, no revela la imagen completa de la interacción entre las células. En consecuencia, para caracterizar a fondo la CCC, estas puntuaciones de comunicación a menudo deben agregarse para mostrar el mapa de la interacción celular.

Para la mayoría de los estudios de CCC con técnicas scARN-seq, el flujo de trabajo comienza con las muestras siendo analizadas para la cuantificación de los transcritos de las células individuales. El siguiente paso es el procesamiento de los datos obtenidos para crear una matriz de expresión génica (MEG), que registra los niveles de expresión de los genes de la muestra, en este caso, las células individuales. Cada fila de la matriz representa un gen, y cada columna representa una célula (1). Una vez obtenida la matriz, es necesario identificar los ligandos, receptores, y las interacciones LR, lo cual se consigue comparando con las bases de datos públicas de interacción proteína-proteína. Estas bases de datos abarcan una amplia gama de interacciones que participan en la CCC. Para cada tipo de células en una muestra, se evalúa cada candidato de interacción LR considerando el nivel de expresión del ligando y del receptor. Esto filtra la MEG para incluir solo ligandos y receptores (1). Generada la MEG filtrada, el análisis de la CCC finaliza al aplicar las puntuaciones de comunicación, obteniendo así una matriz de interacción celular-celular (MICC). La matriz es una representación que muestra cómo diferentes tipos de células se

comunican entre sí. Cada fila y columna de la matriz corresponden a un tipo celular, y los valores indican la fuerza o probabilidad de interacción entre ellas interpretada por la puntuación de comunicación.

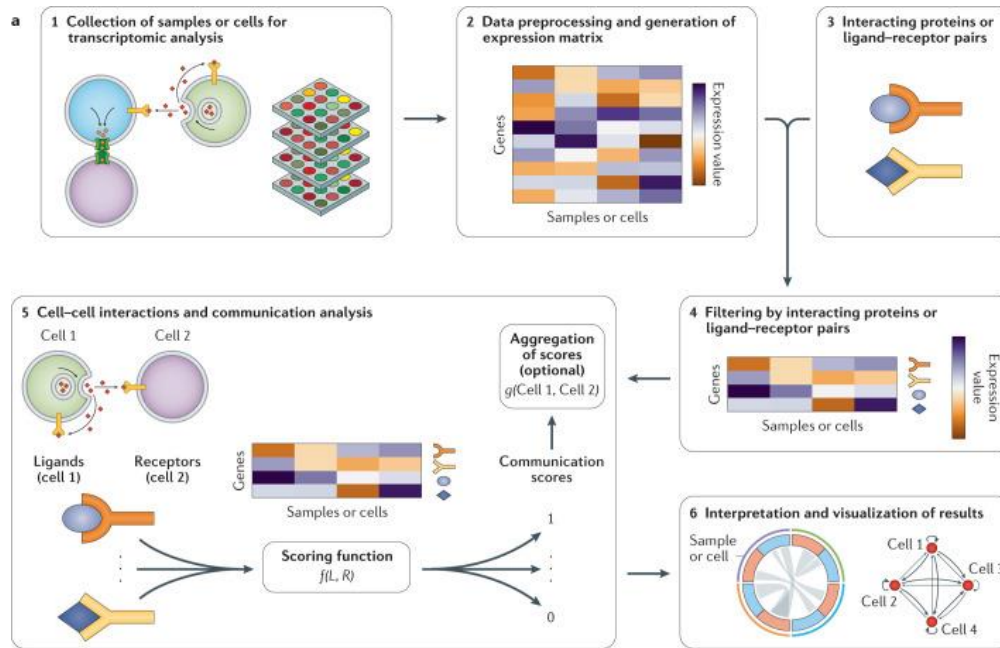


Figura 1: Flujo de trabajo general de un software que infiere CCC a partir de datos de scARN-seq y principales funciones de puntuación de las vías de comunicación (1). A) El primer paso del flujo es la recolección de muestras para el análisis mediante la transcriptómica para obtener el número de transcritos de la muestra. El segundo paso es el preprocesamiento para corregir errores y eliminar artefactos, y la generación de la MEG que relaciona el número de transcritos con la célula o muestra.

El flujo del trabajo prosigue con el tercer paso, la asociación de los transcritos con los genes de los ligandos y receptores en la matriz a partir de las bases de datos externas de interacción LR. A continuación, en el cuarto paso se filtra para conseguir una MEG con sólo ligandos y receptores. El penúltimo paso del flujo de trabajo es la relación matemática de la coexpresión génica de los ligandos y receptores con las funciones de comunicación para calcular la puntuación de comunicación. Esta puntuación se puede agregar para ilustrar el estado general de la interacción. Finalmente, el último paso del flujo es la visualización mediante gráficos, y la interpretación de la CCC. B) Recomendación de uso de los diferentes métodos de puntaje de comunicación para la puntuación de comunicación en función del tipo de secuenciación de mARN.

Existen dos tipos principales de softwares para la inferencia de CCC: aquellos que enfocan las células desde una perspectiva pseudo-bulk, surgidos primero y que operan a nivel de poblaciones celulares para describir estas interacciones, y, por otro lado, aquellos con el enfoque de resolución a nivel celular, surgidos recientemente, que permiten una mayor precisión en la identificación de interacciones específicas al operar en célula individual. CellChat es un ejemplo destacado de los softwares de enfoque pseudo-bulk (13), conocido por su fiabilidad tras comparaciones con otras herramientas similares (14–16). En contraste, la herramienta NICHES representa el enfoque de resolución a nivel celular (17).

1.5.1. Enfoque pseudo-bulk

El enfoque pseudo-bulk en el contexto de las CCCs es una estrategia utilizada para analizar las interacciones LR entre grupos celulares a partir de datos de expresión génica obtenidos en estudios de scARN-seq (13). El proceso comienza con la clasificación de las células, que puede ser definida por el usuario, agrupando previamente las células bajo etiquetas como tipos celulares o estados de diferenciación, o bien es realizada automáticamente por el software, mediante análisis computacionales y estadísticos exhaustivos que identifican patrones de expresión y los clasifica en diversos clústeres (14).

Una vez etiquetada la MEG, ya sea manualmente o automáticamente, el siguiente paso es agrupar las células según estas etiquetas, promediando el perfil de expresión génica dentro de cada grupo. Esto genera un perfil de expresión media para cada tipo celular, lo que permite convertir los datos de una escala de células individuales a una escala poblacional (13). Este procedimiento ayuda a reducir el ruido técnico presente en los datos unicelulares, facilitando el análisis de las tendencias generales de expresión génica.

Tras la creación de la matriz pseudo-bulk con los perfiles de expresión agregados, las herramientas utilizan bases de datos o modelos predictivos para identificar pares de genes que codifican ligandos y receptores (14). Estos genes están diferencialmente expresados o sobreexpresados en cada grupo celular. Las bases de datos utilizadas contienen interacciones previamente conocidas entre ligandos y receptores, las cuales han sido curadas, es decir, seleccionadas y verificadas manualmente por expertos. Ejemplos de estas bases de datos incluyen OmniPath (18), Functional Annotation of the Mammalian Genome (FANTOM5) (4), CellChatDB (19) y ConnectomeDB (20).

Posteriormente, se cuantifican las posibles interacciones LR entre los diferentes grupos celulares. Este análisis implica evaluar la coexpresión de ligandos en un grupo celular y receptores en otro grupo. El nivel de expresión de los genes codificantes de ligandos y receptores se utiliza para estimar inferir las posibles interacciones de señalización entre los distintos tipos celulares y aplicar las puntuaciones de comunicación (13)

Como el enfoque pseudo-bulk agrupa en clústeres o en tipos celulares la expresión génica de las células individuales, presenta un problema a la hora de inferir de las CCCs a partir de los datos scARN-seq puesto que la agrupación de las células se basan en valores medios de expresión (12). La expresión media no aprovecha completamente la resolución a nivel de célula única, lo que oculta la rica variedad de patrones de señalización entre las células (17,21). Por lo que no representa biológicamente la CCC ya que esta no opera a nivel de grupo, si no que las interacciones ocurren entre células individuales (22).

1.5.2. Enfoque de resolución a nivel celular

Los métodos computacionales para el análisis de scARN-seq han avanzado considerablemente, lo que ha permitido el desarrollo de softwares capaces de inferir la CCC a partir de datos de scARN-seq con resolución a nivel de célula individual. El enfoque de resolución celular en el análisis de interacciones LR se enfoca en estudiar cómo interactúan y se comunican las células individuales, evitando promediar datos a nivel de grupo (13,20). Este enfoque proporciona una visión más precisa y detallada de las interacciones moleculares entre células.

El proceso comienza tomando como entrada una MEG. Los softwares identifican los pares de genes que codifican ligandos y receptores, así como aquellos genes diferencialmente expresados que codifican ligandos y receptores en cada célula individual, siguiendo un enfoque similar al del análisis pseudo-bulk.

Luego, el software permite al usuario definir cuál de los clústeres o tipos celulares son los emisores y receptores. Dado que el objetivo final de estos programas es construir la MICC, se implementan procedimientos de optimización para manejar la alta dimensionalidad y el gran volumen de datos que implica el análisis de las interacciones celulares (13).

A diferencia del enfoque pseudo-bulk, que agrupa datos y genera un promedio de expresión para cada tipo celular, el enfoque de resolución celular conserva la variabilidad de la expresión génica a nivel individual, permitiendo identificar interacciones celulares específicas. Esto revela patrones de señalización que podrían quedar ocultos en los análisis basados en promedios, lo que ofrece una comprensión más precisa de las dinámicas de comunicación entre células en su contexto biológico.

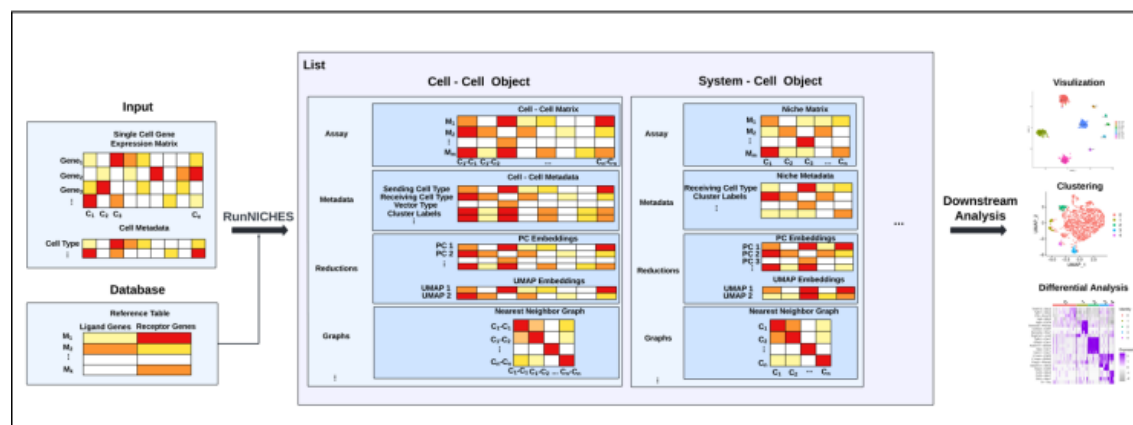


Figura 2: Ejemplo de un flujo de trabajo de un software de inferencia CCC a partir de datos scARN-seq con enfoque de resolución celular (23). El software toma como entrada una MEG de scARN-seq, y la metadata asociada de las células. Vía la función de RunNICHES, la herramienta referencia los mecanismos de interacción LR a través de bases de datos externas para crear la MICC, donde las columnas representan pares de células y las filas los mecanismos de interacción. Esta matriz se genera como salida, junto con otra información útil, y que pueden procesarse para análisis posteriores como la visualización de datos, agrupamiento y análisis diferencial.

1.6. Justificación

La CCC es un proceso fundamental en los organismos multicelulares. A través de la CCC, las células intercambian señales que regulan una amplia gama de procesos biológicos, desde el desarrollo embrionario hasta la respuesta inmune y la homeostasis tisular. Las alteraciones en la CCC están implicadas en diversas patologías, incluyendo el cáncer, enfermedades cardiovasculares, y trastornos inmunológicos. Por tanto, comprender en detalle cómo las células se comunican entre sí es crucial para el desarrollo de nuevas estrategias terapéuticas y diagnósticas.

La tecnología de scARN-seq permite analizar la CCC al identificar ligandos y receptores a nivel de célula individual. La mayoría de las herramientas desarrolladas que infieren la CCC tienen un enfoque pseudo-bulk, es decir hay una pérdida de la resolución en este proceso, ya que se usa la expresión media de los ligandos y receptores para detectar las interacciones.

Recientemente, se han desarrollado softwares con un enfoque de resolución a nivel celular que permiten inferir interacciones LR directamente entre células individuales, lo que es crucial para capturar la heterogeneidad celular y las interacciones únicas que pueden perderse en los enfoques que promedian la expresión génica. Estos métodos ofrecen una representación más precisa de las CCCs, especialmente en entornos biológicos complejos.

A pesar de los avances en este enfoque, es crucial realizar comparaciones rigurosas entre el enfoque de resolución a nivel celular y el enfoque clásico pseudo-bulk para evaluar sus capacidades. Esto permitirá determinar si la inferencia de CCC a nivel de célula individual supera al enfoque pseudo-bulk en términos de precisión, eficiencia y relevancia biológica.

2. Objetivos

2.1. Objetivo principal

Establecer un pipeline de análisis para el estudio de las CCCs en datos de scARN-seq.

2.2. Objetivos específicos

- Evaluar y comparar la precisión de las herramientas NICHES y CellChat en la predicción de interacciones LR utilizando bases de datos validadas como FANTOM5 y proteómica, y medir la consistencia de sus predicciones frente a variaciones en el tamaño de las muestras, empleando índices como el de Jaccard y el de Similitud.
- Analizar el tiempo de ejecución y el consumo de memoria de NICHES y CellChat al procesar distintos conjuntos de datos scARN-seq.

3. Metodología

3.1. Datos

A pesar de que todos los softwares comparten dos componentes principales el método para puntuar la comunicación, y la base de datos para estimar las interacciones. Los softwares pueden utilizar diferentes métodos para descifrar las interacciones según los conjuntos de datos de entrada. Estos, a su vez, dan lugar a sistemas de puntuación diversos, difíciles de comparar y evaluar. Las dificultades para evaluar el rendimiento de los softwares de CCC se ven aún más agravadas por la falta de un estándar de referencia adecuado (10,24). Para abordar esta problemática, Cabello-Aguilar y colaboradores (25) desarrollaron un pseudoestándar de oro basado en la comparación de los datos de células mononucleares de sangre periférica (PBMC) con las bases de datos del proyecto FANTOM5 (4) y una base de datos proteómica de interacciones LR (26).

En este estudio de *benchmark* se emplearon varios conjuntos de datos scARN-seq. En primer lugar, se utilizaron tres conjuntos de PBMC provenientes de la empresa 10xGenomics, con 3000, 6000 y 8000 células respectivamente. Los datos PBMC se obtuvieron a partir de muestras de sangre de tres donantes humanos sanos, uno para cada

conjunto. El conjunto de datos PBMC 3k fue adquirido del paquete SeuratData (27), mientras que los dos restantes se descargaron de la página web de 10xGenomics (28,29).

Para evaluar NICHES y CellChat en diferentes contextos biológicos, se seleccionaron tejidos del corazón y del hígado debido a la relevancia de la CCC en estos tejidos. En el corazón, la interacción entre diferentes tipos celulares es crucial para la contracción sincronizada y la función cardiovascular (30). En el tejido hepático, esta comunicación es esencial para la función metabólica, la respuesta inmunológica (31) y la regeneración (32).

Los datos del tejido cardíaco fueron extraídos de una muestra de catorce corazones humanos, que consistía en 7,495 células después de ser filtradas por los autores del estudio. Estos datos están disponibles a través del enlace de descarga proporcionado en la publicación (33). Por otro lado, el tercer conjunto de datos proviene de una muestra del lóbulo hepático derecho de un donante humano sano. Este conjunto consiste en 8884 células filtradas. Estos datos se obtuvieron del paquete de R asociado al mismo estudio (34).

3.2. Softwares

Las herramientas analizadas, NICHES y CellChat, son softwares de inferencia de interacciones celulares a partir de datos de scARN-seq que emplean enfoques diferentes para procesar la información. NICHES infiere las CCCs a nivel de célula única, proporcionando una alta resolución en la identificación de interacciones específicas. En contraste, CellChat agrupa las células individuales en clústeres utilizando la expresión génica media para inferir la CCC.

3.3. Métricas de evaluación

3.3.1. Precisión

El primer paso de este estudio *benchmark* entre NICHES y CellChat es el cálculo de la métrica de la precisión. Para ello, este *benchmark* utilizó las bases de datos del proyecto FANTOM5, que se basa en la tecnología Cap Analysis of Gene Expression (CAGE) (4) y

una base de datos proteómica de interacciones LR obtenida mediante espectrometría de masas (26).

La razón de la combinación de estos recursos es que proporciona un pseudoestándar de oro, ya que la tecnología CAGE ofrece una mayor precisión en la expresión génica que la tecnología de scARN-seq, mientras que la espectrometría de masas, al ser una tecnología proteómica, valida la presencia proteica de los ligandos y receptores (15). La tecnología utilizada en FANTOM5, CAGE, proporciona una cuantificación directa de la abundancia de transcritos y la detección de variantes de transcripción, lo que resulta especialmente valioso en el contexto de datos de célula única. Además, CAGE ofrece una mayor profundidad de secuenciación y una menor pérdida de datos durante el proceso, en comparación con la tecnología de scARN-seq (15). Esto se debe a que CAGE se basa en la captura de los extremos 5' de los mARN, donde se encuentran los sitios de inicio de la transcripción, aprovechando la estructura única de la caperuza 5', una guanina metilada que estabiliza el mARN y está presente en la mayoría de los transcritos maduros (35). Esta captura de los extremos 5' de los mARN proporciona información precisa sobre los sitios de inicio de la transcripción, lo que ayuda a identificar qué transcritos están presentes y de qué gen provienen (36). Este enfoque específico permite una detección más clara y precisa de los transcritos, incluyendo aquellos con baja abundancia, y asegura una preparación eficiente de la biblioteca.

Los autores establecieron que un umbral de expresión superior a 10 transcritos por millón (TPM) para los genes que codifican ligandos y receptores es un umbral adecuado para garantizar una buena tasa de verdaderos positivos en la inferencia de pares LR (4). Este umbral se definió para clasificar las interacciones LR. Así, si la expresión tanto del ligando como del receptor en la base de datos CAGE de FANTOM5 es superior a 10 TPM, la interacción se considera positiva y verdaderamente existente; de lo contrario, se considera falsa o inexistente (15).

Por otro lado, la base de datos proteómica, basada en espectrometría de masas, se utiliza para verificar la presencia de interacciones LR (15). Para que una interacción LR se considere presente, tanto el ligando como el receptor deben estar registrados en la base de datos proteómica y tener un conteo espectral promedio superior a 2 (4).

Para la evaluación de la precisión se empleó el Área Bajo la Curva de la Precisión-Recuperación (AUPRC). Los softwares de inferencia de interacciones LR analizaron las tres bases de datos PBMC, generando una MICC. Durante el análisis, se verificó que cada interacción LR de la MICC superase los umbrales establecidos tanto en los datos de CAGE como en la base de datos proteómica. Posteriormente, se filtraron todas aquellas interacciones LR que alcanzaran los umbrales establecidos por los autores.

Con la matriz filtrada, se procedió a calcular dos métricas clave, la precisión y la recuperación. La precisión se define como la fracción de predicciones positivas que son verdaderos positivos, es decir, la proporción de interacciones LR predichos que realmente están presentes en las bases de datos. La recuperación mide la fracción de verdaderos positivos identificados sobre el total de verdaderos positivos reales, indicando cuántas interacciones verdaderas fueron detectadas correctamente por el modelo.

A partir de estas métricas, se construyó la curva de precisión-recall, que muestra la relación entre precisión y recuperación. Además, para una mejor visualización y evaluación del desempeño del modelo en la inferencia de interacciones LR se calculó el AUPRC (17).

3.3.2. Concordancia

El segundo paso en la comparación es evaluar la concordancia, que mide cómo coinciden las predicciones de interacciones entre células realizadas por diferentes herramientas. Para ello, se precisó usar el índice de similitud, una modificación del índice de Jaccard para medir el solapamiento en la inferencia de interacciones LR entre los softwares. Este índice se calcula determinando la unión de la intersección de los resultados sobre el mínimo de los tamaños de los conjuntos (15).

Dado que CellChat es un software que, por defecto, emplea su propia base de datos, CellChatDB (19), se utilizó el paquete CCCbank (15) para asegurar una comparación justa. Este paquete integra múltiples bases de datos en CellChat, incluyendo la base de datos FANTOM5 (4), que es la base de datos por defecto utilizada por NICHES para inferir interacciones LR.

3.3.3. Consistencia

El tercer paso del *benchmark* consistió en medir la consistencia de los softwares. En el contexto de los softwares para inferir CCC a partir de datos de scARN-seq, la consistencia de un software se refiere a su capacidad para mantener un rendimiento consistente frente a variaciones en las condiciones de los datos de entrada. Para ello se muestreó aleatoriamente diferentes proporciones de células de los cinco conjuntos de datos y se calculó el índice de similitud en la inferencia de interacciones (15).

El índice de similitud utilizado en este estudio es una variante del índice de Jaccard (15). En este caso, el numerador corresponde a la intersección de las interacciones predichas por los softwares, mientras que el denominador es el valor mínimo entre los tamaños de los conjuntos A y B, es decir, $\min(|A|, |B|)$ (15). Este índice se calcula como la proporción del tamaño de la intersección de dos conjuntos de resultados sobre el tamaño de su unión, y toma valores entre 0 y 1, donde 1 indica máxima similitud, los conjuntos son idénticos, y 0 indica ninguna similitud.

3.3.4. Tiempo de ejecución y uso de memoria

El último paso del *benchmark* ha sido la evaluación el tiempo de ejecución y el uso de memoria. Los resultados se obtuvieron después de que los softwares los diferentes conjuntos de datos reducidos en el apartado de métricas de consistencia para obtener información sobre el tiempo de ejecución y el consumo de memoria.

El código de este *benchmark* se encuentra en la página GitHub: [elioe300/Benchmark-de-Software-para-la-Inferencia-de-la-Comunicaci-n-C-lula-C-lula-en-datos-scARN-seq](https://github.com/elioe300/Benchmark-de-Software-para-la-Inferencia-de-la-Comunicaci-n-C-lula-C-lula-en-datos-scARN-seq) (github.com)

4. Resultados

Las herramientas representativas para el enfoque pseudo-bulk y resolución a nivel celular utilizadas para la comparación fueron CellChat y NICHES, respectivamente. CellChat es un software que ha sido validado en diversos *benchmarks* como una opción fiable (14–16), mientras que NICHES es el software de enfoque a resolución celular más usado y citado (37–43).

CellChat inicia la inferencia de la CCC detectando primero los genes que codifican ligandos y receptores mediante su propia base de datos curada, CellChatDB (19), respaldada por la literatura. En el contexto de este benchmark, se utilizará la base de datos FANTOM5 (4). Posteriormente, identifica los ligandos y receptores que están diferencialmente sobreexpresados en cada tipo celular utilizando la prueba de Wilcoxon. Para cuantificar las CCCs entre dos grupos celulares mediadas por estos genes señalizadores, CellChat asigna a cada interacción un valor de probabilidad que indica la certeza de que dicha interacción realmente ocurra. Este valor se modela utilizando la ley de acción de masas, basada en la media geométrica de los valores promedio de expresión de un ligando y sus subunidades en un grupo celular, y de un receptor y sus subunidades en otro grupo celular, así como en sus cofactores. La expresión cero de cualquier subunidad implica que el ligando es inactivo. Las interacciones significativas predichas se determinan mediante una prueba estadística que permuta aleatoriamente las etiquetas de los grupos celulares y recalcula la probabilidad de interacción, lo que permite evaluar la significancia de las interacciones observadas.

Por otro lado, NICHES toma como entrada la MEG, y utiliza bases de datos externas como FANTOM5 (4) para referenciar los mecanismos de interacción LR para crear la MICC. En esta matriz, las filas corresponden a los mecanismos de interacción, mientras que las columnas representan las interacciones LR, y la puntuación de cada interacción se obtiene multiplicando la expresión del ligando por la expresión del receptor para cada mecanismo, siendo este un método cero preservante. Si uno de los componentes, ya sea una subunidad del ligando o receptor, tuviera una expresión nula, la interacción se asigna automáticamente como inexistente, siendo el valor de la interacción de cero.

4.1. Precisión

Para realizar la evaluación y comparación de la precisión se combinaron la base de datos de 144 tipos celulares primarias del proyecto FANTOM5, que usa la tecnología CAGE (4) y una base de datos de interacciones LR a nivel proteómico para 28 poblaciones primarias de células inmunes humanas (26). La combinación de estos recursos proporciona un pseudoestándar de oro (15).

4.1.1. Evaluación de precisión basada en datos de CAGE

Tras analizar los conjuntos de datos PBMC de 3k, 6k, y 8k células utilizando NICHES y CellChat, se obtuvieron las interacciones LR predichas por cada software según el procedimiento descrito anteriormente. Con las interacciones LR resultantes, se compararon con los datos CAGE para obtener el AUPRC CAGE en cada conjunto de datos, así como el valor promedio de AUPRC de CAGE, obtenido a partir de las diferentes relaciones entre tipos celulares de cada software. La siguiente figura muestra el valor promedio de AUPRC CAGE de cada software:

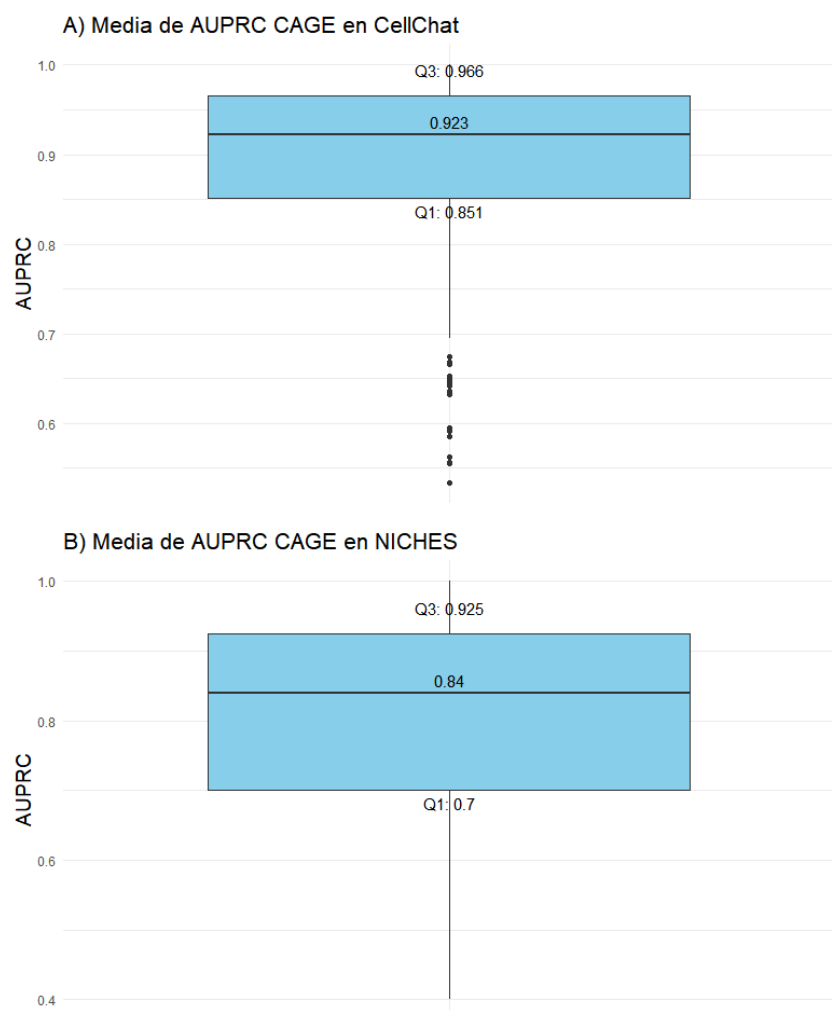


Figura 3: Diagramas de cajas que muestran la distribución de los valores medios de AUPRC de CellChat y NICHES obtenidos tras comparar las predicciones de las interacciones LR de los conjuntos de datos PBMC con la base de datos CAGE de FANTOM5 (4). A) El diagrama de caja ilustra el primer y el tercer cuartil, y la mediana de los valores AUPRC CAGE de CellChat, así como los valores atípicos. B) El diagrama de caja muestra el primer y el tercer cuartil, y la mediana de los valores AUPRC CAGE de NICHES.

Los resultados en el valor de la mediana ilustrados tanto en la Figura 3a como la Figura 3b del AUPRC CAGE indicaron una ligera diferencia en la precisión a la hora de predecir la expresión génica entre las dos herramientas. La mediana que presentó CellChat fue de 0.923, superior a la mediana obtenida por NICHES que fue de 0.840, lo que sugiere que CellChat tiene una ventaja en la precisión al verificar expresión génica de las interacciones LR.

Un aspecto importante a mencionar fue el rango intercuartílico (RIC), que se calculó restando el tercer cuartil del primer cuartil. El RIC mide la dispersión de los valores en un diagrama de caja. Un RIC más amplio indica una mayor variabilidad en torno a la mediana. En este caso, un RIC más amplio indica que hay más variabilidad en las predicciones de la expresión de las interacciones LR. La tabla 1 ilustra los valores del tercer y primer cuartil de los valores AUPRC CAGE de los softwares NICHES y CellChat ilustrados en la Figura 3a y 3b, además del cálculo del RIC.

Software	Tercer cuartil	Primer cuartil	RIC
CellChat	0.966	0.851	0.115
NICHES	0.925	0.700	0.225

Tabla 1: Valores de AUPRC CAGE del tercer cuartil, primer cuartil y RIC de las predicciones de expresión génica obtenidos por CellChat y NICHES tras analizar los tres conjuntos de datos de PBMC 3k, 6k, y 8k, y compararlos con la base de datos CAGE de FANTOM5 (4).

El RIC fue mayor en NICHES, lo que señaló una mayor dispersión en la distribución de estos datos en comparación con CellChat. Esto indica que las predicciones de NICHES fueron más variables en la predicción de la expresión génica dentro del RIC.

Otro rasgo a destacar en el diagrama de cajas fueron los valores atípicos. Son aquellos puntos que se encontraron fuera de los bigotes, y se extienden hasta 1.5 veces del RIC desde el primer y tercer cuartil. Los valores atípicos son cruciales para detectar observaciones que se desviaron considerablemente del patrón general de los datos. En el contexto de predicciones de interacciones LR, la existencia de estos valores en el diagrama de caja indica que la herramienta estaba sobreestimando o subestimando la presencia de estas interacciones.

Para mejorar la visualización de los valores atípicos y facilitar una interpretación más detallada de los resultados, se ha optado por separar los valores de AUPRC CAGE de NICHES y CellChat según el conjunto de datos. Esta agrupación permitió una resolución más alta en el análisis de los datos.

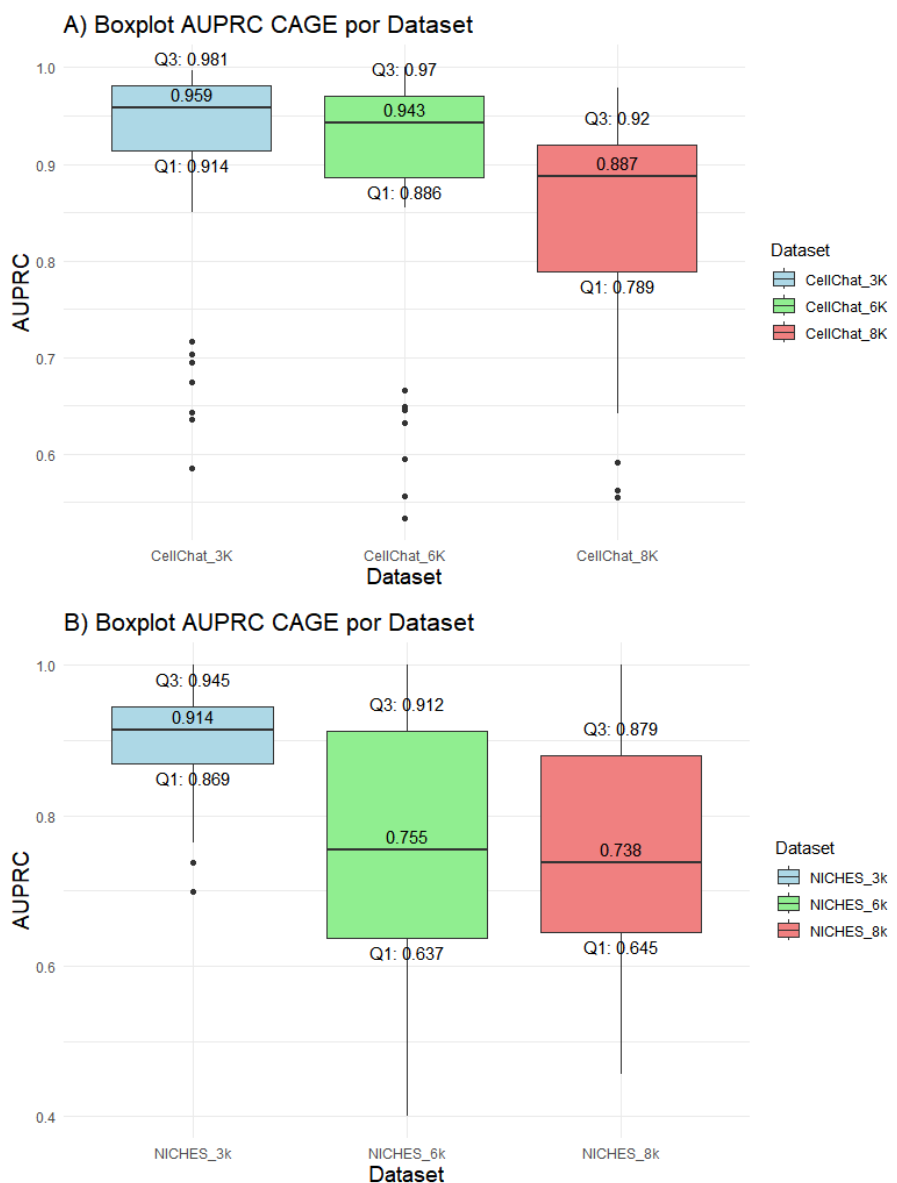


Figura 4: Diagramas de cajas, separados por conjunto de datos, que muestran la distribución de los valores de AUPRC CAGE de CellChat y NICHES tras comparar las predicciones de las interacciones LR de los conjuntos de datos PBMC con la base de datos CAGE de FANTOM5 (4). A) Los diagramas de cajas presentan el primer y el tercer cuartil, y la mediana de los valores AUPRC CAGE de CellChat, así como los valores atípicos, de cada conjunto de datos diferenciado por colores. B) Los diagramas de cajas muestran el primer y el tercer cuartil, y la mediana de los valores AUPRC CAGE de NICHES, así como los valores atípicos, de cada conjunto de datos diferenciado por colores.

Los resultados divididos por conjunto de datos siguieron resaltando las diferencias en la precisión entre NICHES y CellChat. La Figura 4b muestra las medianas obtenidas por NICHES tras analizar los tres conjuntos de datos de PBMC, estos fueron de 0.914, 0.755,

y de 0.738. Estos valores son menores en comparación a los conseguidos por CellChat, cuyas medianas fueron más altas 0.959, 0.943, y de 0.887, tal y como se muestra en la Figura 4a. Esta diferencia fue especialmente notable en el segundo conjunto de datos, PBMC 6k, donde la mediana de NICHES fue 0.755, mientras que la de CellChat fue 0.943.

Un aspecto a destacar fue la diferencia en los valores atípicos entre las medidas de AUPRC CAGE obtenidas con NICHES y CellChat. CellChat presentó una mayor cantidad de valores atípicos en comparación con NICHES, con un total de 17 valores atípicos frente a 2 de NICHES. Los valores atípicos por debajo del RIC, significa que la precisión fue mucho más baja de lo esperado, aunque este hecho ocurriese en ambos softwares, el elevado número registrado por CellChat destacó significativamente. Esto sugiere que CellChat fue más susceptible a la variabilidad extrema en sus predicciones, lo que podría llevar a que, en ciertas ocasiones, subestimase significativamente la expresión génica.

Los RICs reflejan la distribución de los valores en los diagramas de cajas. Las tablas 2 y 3 recogen los resultados ilustrados en las Figuras 4a y 4b del tercer y primer cuartil, junto a la diferencia de estos para obtener el RIC:

CellChat			
Conjunto de dato	Tercer cuartil	Primer cuartil	RIC
PBMC 3k	0.981	0.914	0.067
PBMC 6k	0.970	0.886	0.085
PBMC 8k	0.920	0.789	0.131

Tabla 2: Valores de AUPRC CAGE del tercer cuartil, primer cuartil y RIC de las predicciones de expresión génica obtenidos por CellChat tras analizar los tres conjuntos de datos de PBMC 3k, 6k, y 8k. RIC y compararlos con la base de datos CAGE de FANTOM5 (4).

NICHES			
Conjunto de dato	Tercer cuartil	Primer cuartil	RIC
PBMC 3k	0.945	0.869	0.076
PBMC 6k	0.912	0.637	0.276
PBMC 8k	0.879	0.645	0.234

Tabla 3: Valores de AUPRC CAGE del tercer cuartil, primer cuartil y RIC de las predicciones de expresión génica obtenidos por NICHES tras analizar los tres conjuntos de datos de PBMC 3k, 6k, y 8k. RIC y compararlos con la base de datos CAGE de FANTOM5 (4).

Los valores demostraron un mayor RIC en NICHES, lo que indica que los valores de AUPRC CAGE están más dispersos. Esto sugiere que las predicciones de NICHES están más distribuidas alrededor de la mediana, en contraste con CellChat que mostró un RIC mucho más estrecho, lo que indica que la mayoría de sus valores de AUPRC CAGE están más concentrados alrededor de la mediana.

4.1.2. Evaluación de precisión basada en datos proteómicos

Siguiendo la metodología mencionada para el cálculo de AUPRC proteómico, se presentan los resultados obtenidos, que reflejan la capacidad de los modelos para predecir la presencia proteica, en la siguiente figura:

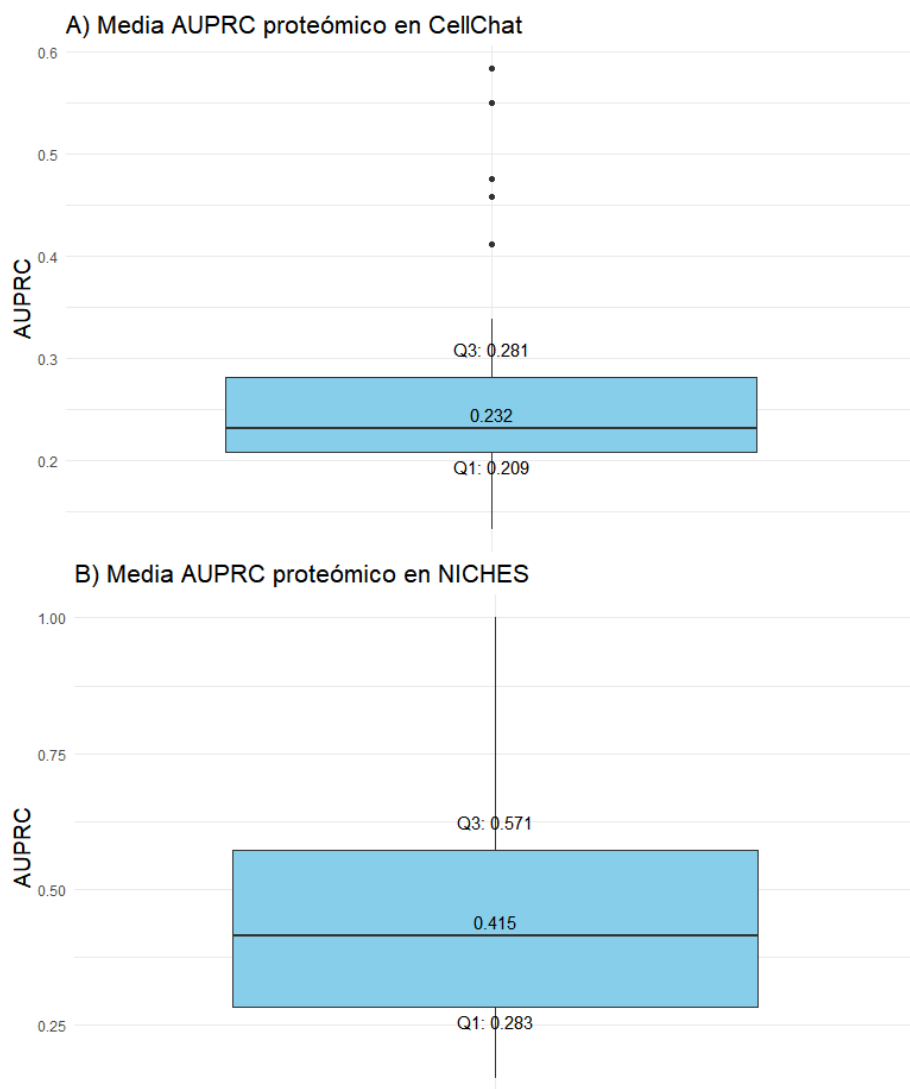


Figura 5: Diagramas de cajas que muestran la distribución de los valores medios de AUPRC proteómico de CellChat y NICHES tras comparar las predicciones de las interacciones LR de los conjuntos de datos PBMC con la base de datos proteómica (26). A) El diagrama de caja presenta el primer y el tercer cuartil, y la mediana de los valores AUPRC proteómico de CellChat, así como los valores atípicos. B) El diagrama de caja ilustra el primer y el tercer cuartil, y la mediana de los valores AUPRC proteómico de NICHES.

Se puede ver en las Figura 5a y 5b que de AUPRC proteómico de NICHES presentó una mediana superior de 0.415 en comparación de la mediana obtenida por CellChat, que fue de 0.231. Este resultado sugiere que NICHES presenta una ventaja en precisión al verificar la presencia de proteica de ligandos y receptores, indicando una mayor sensibilidad en la detección de interacciones proteicas a nivel celular en comparación con CellChat.

Obtenidos los valores AUPRC proteómicos de NICHES y CellChat, se calcularon los RICs mediante la diferencia del tercer y primer cuartil para analizar la distribución de los resultados de ambos softwares, tal y como se muestra en la Tabla 4:

Software	Tercer cuartil	Primer cuartil	RIC
CellChat	0.281	0.209	0.072
NICHES	0.571	0.283	0.288

Tabla 4: Valores de AUPRC proteómico del tercer cuartil, primer cuartil y RIC de las predicciones de la presencia proteica obtenidos por CellChat y NICHES tras analizar los tres conjuntos de datos de PBMC 3k, 6k, y 8k y compararlas con la base de datos proteómica (26).

El RIC de NICHES indicó que sus predicciones estuvieron más distribuidas sobre la mediana, mientras que los resultados de CellChat tuvieron una menor distribución. Esto sugiere que CellChat presenta una menor variabilidad a la hora de presentar las predicciones en términos de presencia proteica.

Para una mayor resolución de los resultados se procedió a separar el AUPRC proteómico de CellChat y NICHES por conjunto de datos, los resultados se muestran en la siguiente figura:

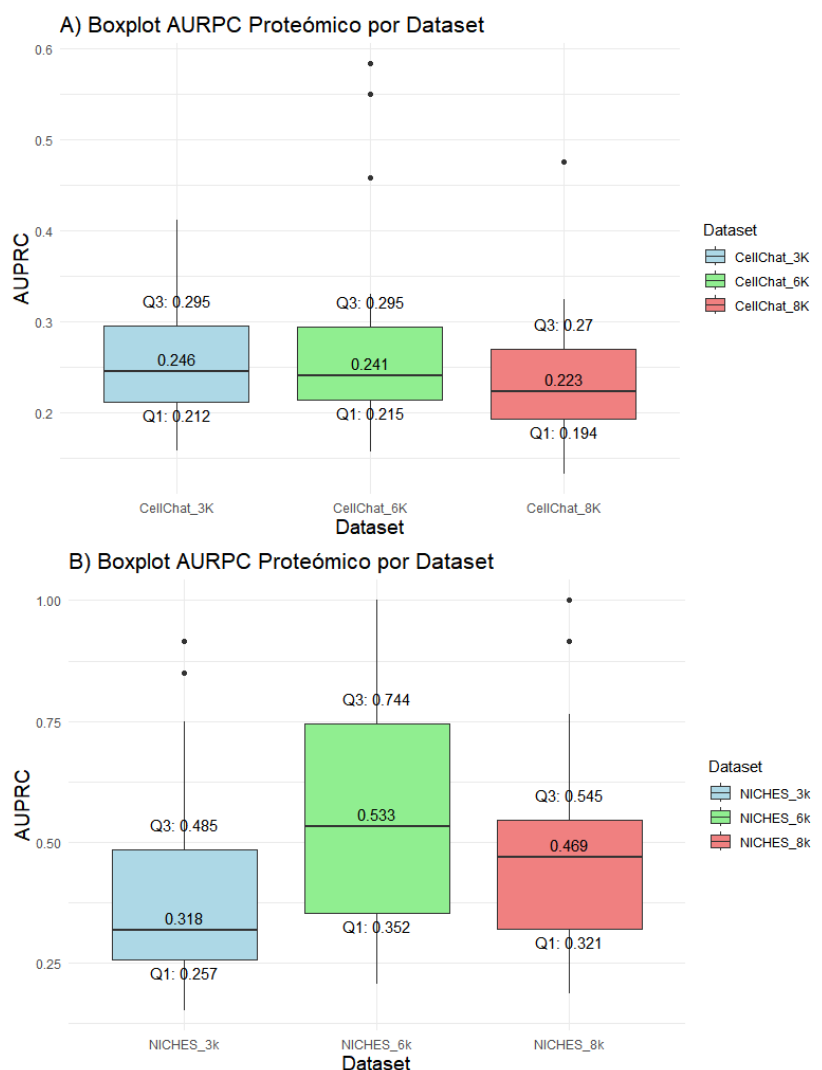


Figura 6: Diagramas de cajas, separados por conjunto de datos, que muestran la distribución de los valores de AUPRC proteómico de CellChat y NICHES tras comparar las predicciones de las interacciones LR de los conjuntos de datos PBMC con la base de datos proteómica (26). A) Los diagramas de cajas presentan el primer y el tercer cuartil, y la mediana de los valores AUPRC proteómico de CellChat, así como los valores atípicos, de cada conjunto de datos diferenciado por colores. B) Los diagramas de cajas muestran el primer y el tercer cuartil, y la mediana de los valores AUPRC proteómico de NICHES, así como los valores atípicos, de cada conjunto de datos diferenciado por colores.

Los resultados observados en la Figura 6b demostraron que NICHES alcanzó mejores resultados en términos de AUPRC proteómico, ya que presentó medianas superiores de

0.318, 0.533 y de 0.469 en comparación con las medianas presentadas por CellChat de 0.246, 0.241 y de 0.223, observables en la Figura 6a.

Es relevante destacar que en tanto PBMC 6k como PBMC 8k NICHES alcanzara el máximo de 1, demostrando que, en al menos una de las mediciones, fue capaz de predecir la presencia proteica todas las interacciones LR presentes en el conjunto de datos.

Un aspecto a valorar fueron las distribuciones de los diagramas de caja, en este caso, los RICs de NICHES fueron mayores que las de CellChat, tal como se muestra en las tablas 5 y 6:

CellChat			
Conjunto de dato	Tercer cuartil	Primer cuartil	RIC
PBMC 3k	0.295	0.212	0.083
PBMC 6k	0.295	0.215	0.081
PBMC 8k	0.270	0.194	0.076

Tabla 5: Valores de AUPRC proteómico del tercer cuartil, primer cuartil y RIC de las predicciones de la presencia proteica obtenidas por CellChat tras analizar los tres conjuntos de datos de PBMC 3k, 6k, y 8k y compararlas con la base de datos proteómica (26).

NICHES			
Conjunto de dato	Tercer cuartil	Primer cuartil	RIC
PBMC 3k	0.485	0.257	0.228
PBMC 6k	0.744	0.352	0.392
PBMC 8k	0.545	0.321	0.224

Tabla 6: Valores de AUPRC proteómico del tercer cuartil, primer cuartil y RIC de las predicciones de la presencia proteica obtenidas por NICHES tras analizar los tres conjuntos de datos de PBMC 3k, 6k, y 8k y compararlas con la base de datos proteómica (26).

4.2. Concordancia

Las tablas 7 y 8 muestran las interacciones predichas por CellChat y NICHES tras analizar las interacciones de los conjuntos de datos, además de la evaluación mediante el índice de similitud para evaluar la concordancia:

Conjunto de dato	Interacciones predichas por NICHES	Interacciones predichas por CellChat	Índice de similitud
PBMC 3k	4447	5445	0.73
PBMC 6k	2214	4897	0.60
PBMC 8k	4857	7953	0.60
Aurícula izquierda	1067	803	0.28
Ventrículo izquierda	806	445	0.26
Hígado	8781	10656	0.66

Tabla 7: Comparación entre los ligandos y receptores únicos detectados por los softwares NICHES y CellChat, junto con el índice de similitud entre ambos softwares según el conjunto de dato.

Conjunto de dato	Ligandos y receptores detectados por NICHES	Ligandos y receptores detectados por CellChat
PBMC 3k	442	335
PBMC 6k	428	277
PBMC 8k	613	390
Aurícula izquierda	464	113
Ventrículo izquierda	449	98
Hígado	414	295

Tabla 8: Comparación entre los ligandos y receptores únicos detectados por los softwares NICHES y CellChat tras analizar el conjunto de dato

Los resultados indicaron una diferencia tanto en el número de interacciones predichas como en el número de ligandos y receptores detectados por los softwares NICHES y CellChat. En cuanto al número de ligandos y receptores, los valores indicaron que NICHES identificó un mayor número que el detectado por CellChat. Sin embargo, en términos de interacciones predichas, NICHES infirió menos interacciones que CellChat, a excepción en los datos de la aurícula y ventrículo izquierdo.

El índice de similitud reveló que, en la mayoría de los conjuntos de datos, hay una concordancia positiva entre los dos softwares, ya que la mayoría de los valores alcanzó o superó el 60% de similitud. El valor más alto del índice de similitud se obtuvo en el caso del conjunto de datos PBMC 3k con un índice de 0.73. En los conjuntos del hígado, PBMC 6k, y PBMC 8k los índices de similitud alcanzaron el 60%, lo que indica que aproximadamente el 60% de las interacciones LR fueron detectadas por ambos softwares.

Sin embargo, en los tejidos cardíacos, la similitud entre las predicciones fue baja en comparación con el resto con un índice de 0.28 para la aurícula izquierda y 0.26 para el ventrículo izquierdo, lo que indica una reducida superposición en las interacciones predichas, sugiriendo que cada software podría estar capturando distintos aspectos de las interacciones LR en esos contextos.

En resumen, NICHES detectó más ligandos y receptores que CellChat, pero predijo menos interacciones, a excepción en el conjunto de datos de la aurícula y el ventrículo izquierdo. Esta diferencia sugiere que NICHES fue más conservador en la predicción de interacciones LR, mientras que CellChat tendió a predecir un mayor número de interacciones LR potenciales, aunque con menos ligandos y receptores involucrados. Los resultados indicaron que NICHES podría estar centrado en interacciones más específicas y limitadas, mientras que CellChat parece adoptar un enfoque más amplio y exhaustivo en cuanto a la conectividad. En general, la concordancia entre ambos softwares fue positiva, pero la baja concordancia en los tejidos cardíacos sugiere que los softwares capturaron interacciones distintas o que uno de ellos fuese más sensible a las particularidades de estos tejidos.

4.3. Consistencia

Al analizar cómo varió el rendimiento en la predicción de las interacciones LR del software al reducir la cantidad de células al 90%, 80%, 70%, 60% y 50% del total original, se cuantificaron los resultados mediante el índice de Jaccard. Los resultados obtenidos se encuentran divididos por la proporción de reducción aplicado al conjunto de datos original, y por software, tal como ilustran las Figuras 7a y 7b:

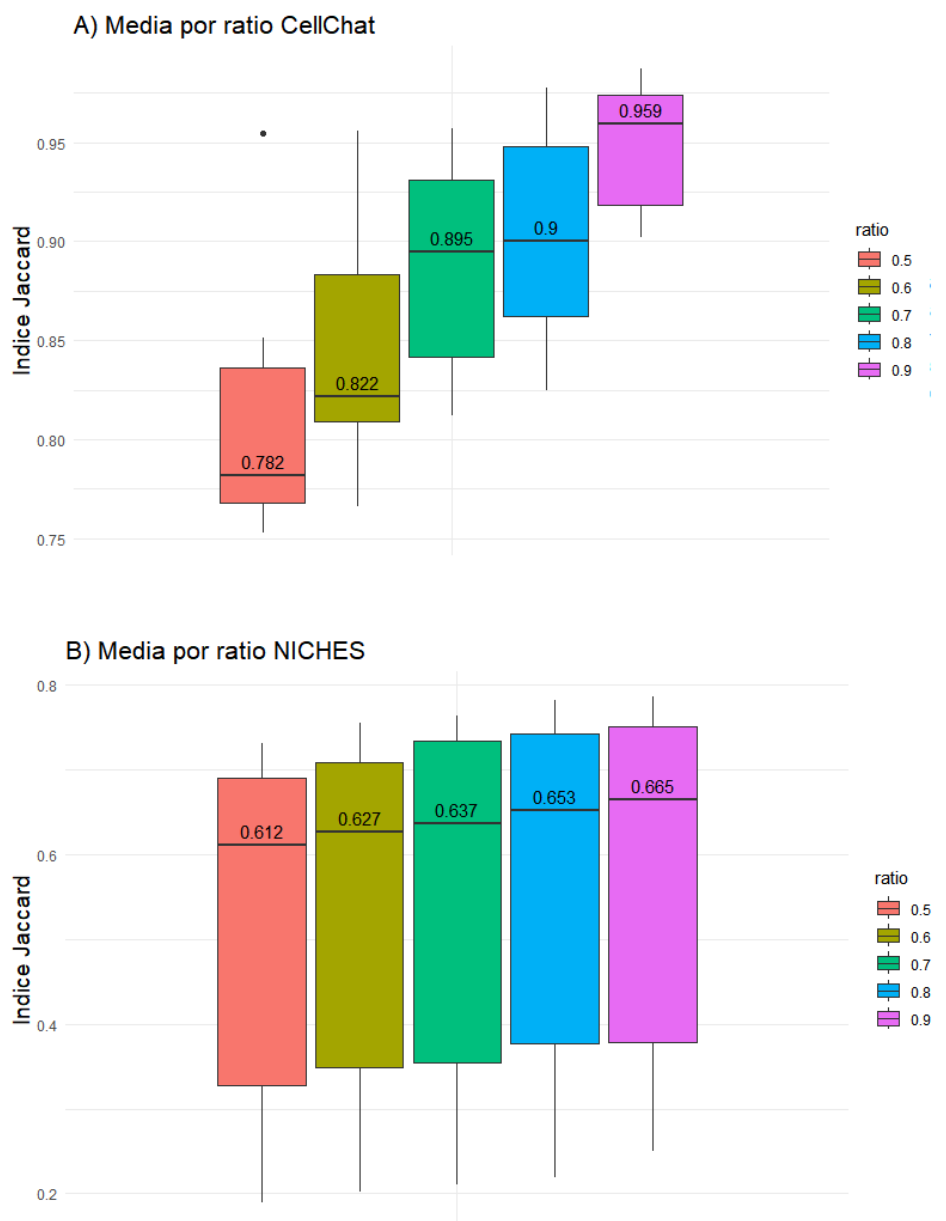


Figura 7: Diagramas de cajas, separados por la proporción de conjunto de dato analizado, que ilustran la distribución de los índices de Jaccard de CellChat y NICHES tras la comparación de las predicciones de interacciones LR entre el conjunto de datos original y el conjunto de datos reducido. A) Los diagramas de caja representan las medianas de los índices de Jaccard obtenidos con CellChat. Los índices están categorizados por color según las diferentes proporciones de reducción aplicadas a los conjuntos de datos. B) Los diagramas de caja muestran las medianas de los índices de Jaccard correspondientes a NICHES. Los índices están categorizados según las diferentes proporciones de reducción aplicadas a los conjuntos de datos.

Los resultados de los valores medios de los índices de Jaccard indicaron que ambos softwares, disminuyeron su rendimiento a medida que se fue reduciendo el número de células en el análisis. La tabla 9 ilustra la tendencia de CellChat:

CellChat	
Reducción	Índice de Jaccard
90%	0.959
80%	0.900
70%	0.895
60%	0.822
50%	0.782

Tabla 9: Valores medios del índice de Jaccard de CellChat calculados tras las reducciones correspondientes en la cantidad de células de los conjuntos de datos.

La mediana del índice de Jaccard para la reducción al 90% mostró un valor de 0.959, lo que indica que el 96% de las predicciones realizadas por CellChat tras la reducción se encontraron en las predicciones obtenidas con el conjunto completo de datos. A medida que la reducción en el número de células aumenta, el índice de Jaccard disminuyó progresivamente. Para la reducción al 50% de los datos originales, la mediana de 0.782 sugiere que CellChat seguía detectando el 78% de las interacciones originales. En general, los valores del índice de Jaccard se mantuvieron razonablemente altos.

En contraste, para NICHES, los resultados mostrados en la tabla 10 revelaron una tendencia diferente:

NICHES	
Reducción	Índice de Jaccard
90%	0.665
80%	0.653
70%	0.637

60%	0.627
50%	0.611

Tabla 10: Valores medios del índice de Jaccard de NICHES calculados tras las reducciones correspondientes en la cantidad de células de los conjuntos de datos.

La mediana del índice de Jaccard para la reducción al 90% en NICHES mostró un valor de 0.665, lo que indica que el 66% de las predicciones realizadas tras la reducción estuvieron presentes en las predicciones obtenidas con el conjunto completo de datos. A medida que la reducción en el número de células aumenta, el índice de Jaccard se redujo ligeramente. Para la reducción al 50% de los datos originales, la mediana de 0.611 sugiere que NICHES seguía detectando aproximadamente el 61% de las interacciones originales. Aunque los valores del índice de Jaccard fueron más bajos en comparación con otros softwares, NICHES se mantuvo consistente.

CellChat mostró una mayor consistencia que NICHES, ya que sus valores del índice de Jaccard se mantuvieron más altos a medida que se redujo el número de células. Esto significa que CellChat conservó mejor la capacidad de detectar las interacciones originales, incluso con una disminución considerable en el tamaño del conjunto de datos. Aunque es lógico que ambos softwares perdieran precisión al reducirse la cantidad de células, CellChat siguió identificando el 78% de las interacciones cuando el conjunto se redujo al 50%, mientras que NICHES solo alcanzó el 61%. La disminución más pronunciada del índice de Jaccard en NICHES sugiere que fue más sensible a la reducción del tamaño de los datos.

4.4. Requisitos computacionales

4.4.1. Tiempo

La computadora utilizada para medir el tiempo de ejecución en este *benchmark* posee el sistema operativo Windows 10 de 64 bits y está equipada con un procesador Intel Core i7-4500U a 1.80 GHz (hasta 2.40 GHz), 8 GB de memoria RAM y un disco SSD de 240 GB. El tiempo de ejecución se obtuvo tras el análisis de los datos reducidos de todos los conjuntos de datos a la hora de medir la consistencia. Los resultados de los numerosos análisis se ilustran en la figura 8:

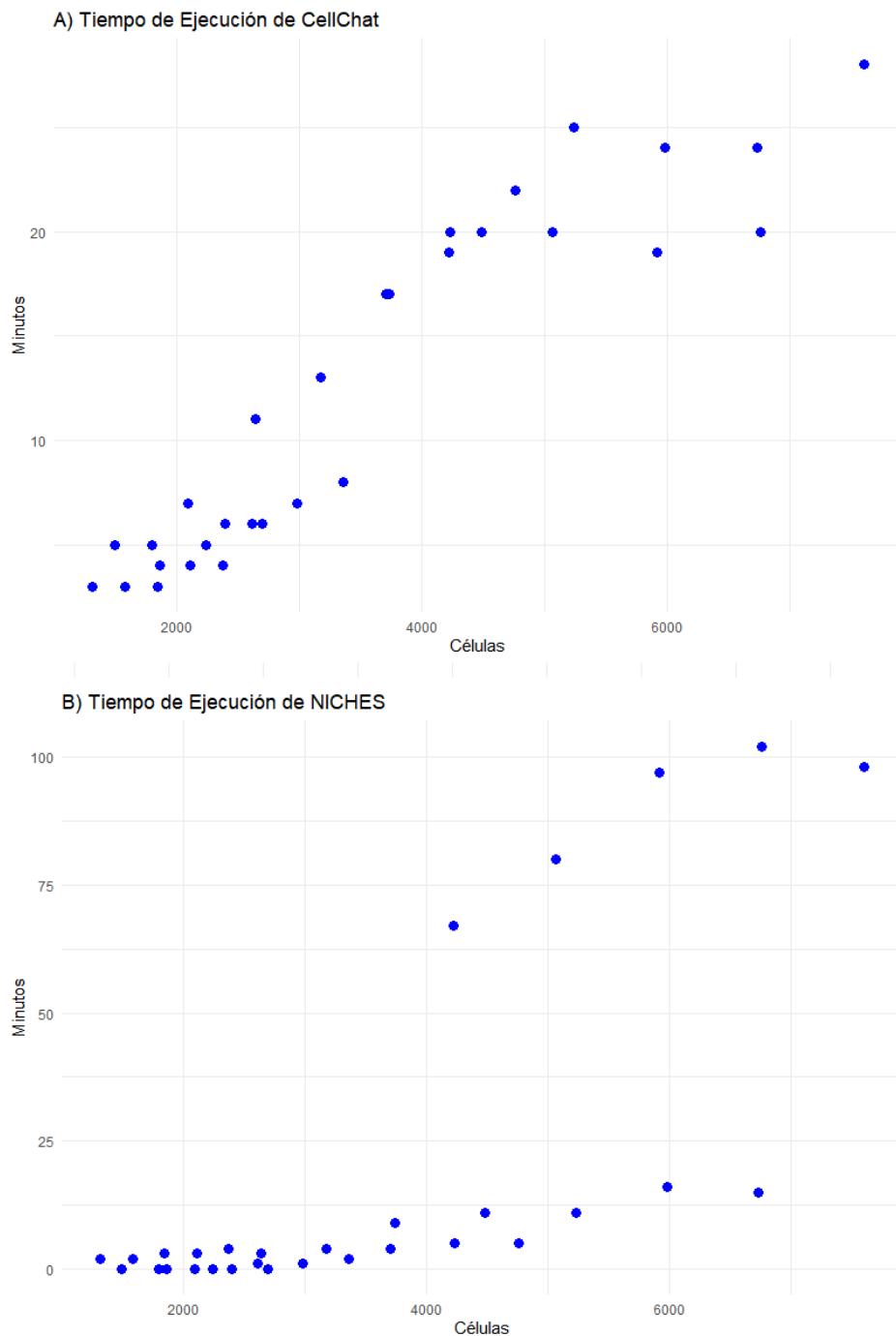


Figura 8: Gráficas de dispersión del tiempo de ejecución de CellChat y NICHES al analizar las diferentes proporciones de los conjuntos de datos. A) La gráfica de dispersión muestra el tiempo de ejecución de CellChat. B) La gráfica de dispersión representa el tiempo de ejecución de NICHES. En ambas gráficas, el eje X indica el número de células analizadas, y el eje Y muestra el tiempo total de ejecución en minutos requerido para procesar los datos.

Como se puede comprobar en la Figura 8b, NICHES realizó su función en un rango de 0 a 4 minutos en conjuntos de datos con menos de 3000 células, realizó su función en un rango de 0 a 4 minutos. En contraste, la Figura 8a muestra que CellChat presentó tiempos de ejecución más altos en los mismos conjuntos de datos, con un rango de 3 a 11 minutos. Esto sugiere que CellChat, debido a su metodología para calcular las interacciones, podría tener una mayor complejidad en el procesamiento.

Como se puede observar en las Figuras 8a y 8b, los tiempos de ejecución aumentan a medida que el número de células a analizar, y que NICHES se sigue manteniendo una ventaja en el tiempo de ejecución sobre CellChat. Sin embargo, NICHES experimentó tiempos de ejecución notablemente largos en el análisis de los datos del hígado, con más de una hora para la reducción al 50% y valores cercanos a las dos horas para la reducción al 90%. Este incremento en el tiempo de ejecución no se debe al número de células per se, sino al número de tipos celulares. El conjunto de dato del tejido hepático se encuentra agrupado en 20 tipos diferentes. NICHES, al utilizar la función `FindAllMarkers()` de Seurat (27) para identificar ligandos y receptores diferenciados, requirió de más tiempo de procesamiento a medida que aumentaba el número de clústeres.

Los resultados concluyeron que ambos softwares mostraron un incremento en el tiempo de ejecución a medida que aumentaron el número de células. NICHES, en particular, exhibió un patrón en el que el tiempo de ejecución es bajo para conjuntos de datos pequeños, pero que aumenta significativamente con el número de clústeres a analizar.

4.4.2. Memoria

Diferentes herramientas usan diferentes algoritmos y estrategias de optimización, lo que puede llevar a variaciones en el uso de memoria (15). Para evaluar la eficiencia en el uso de memoria, se midió la cantidad de RAM requerida por cada software tanto para crear la MICC como para predecir las interacciones LR. Esta comparación permitió determinar cuál de los softwares fue más eficiente en términos de manejo de memoria.

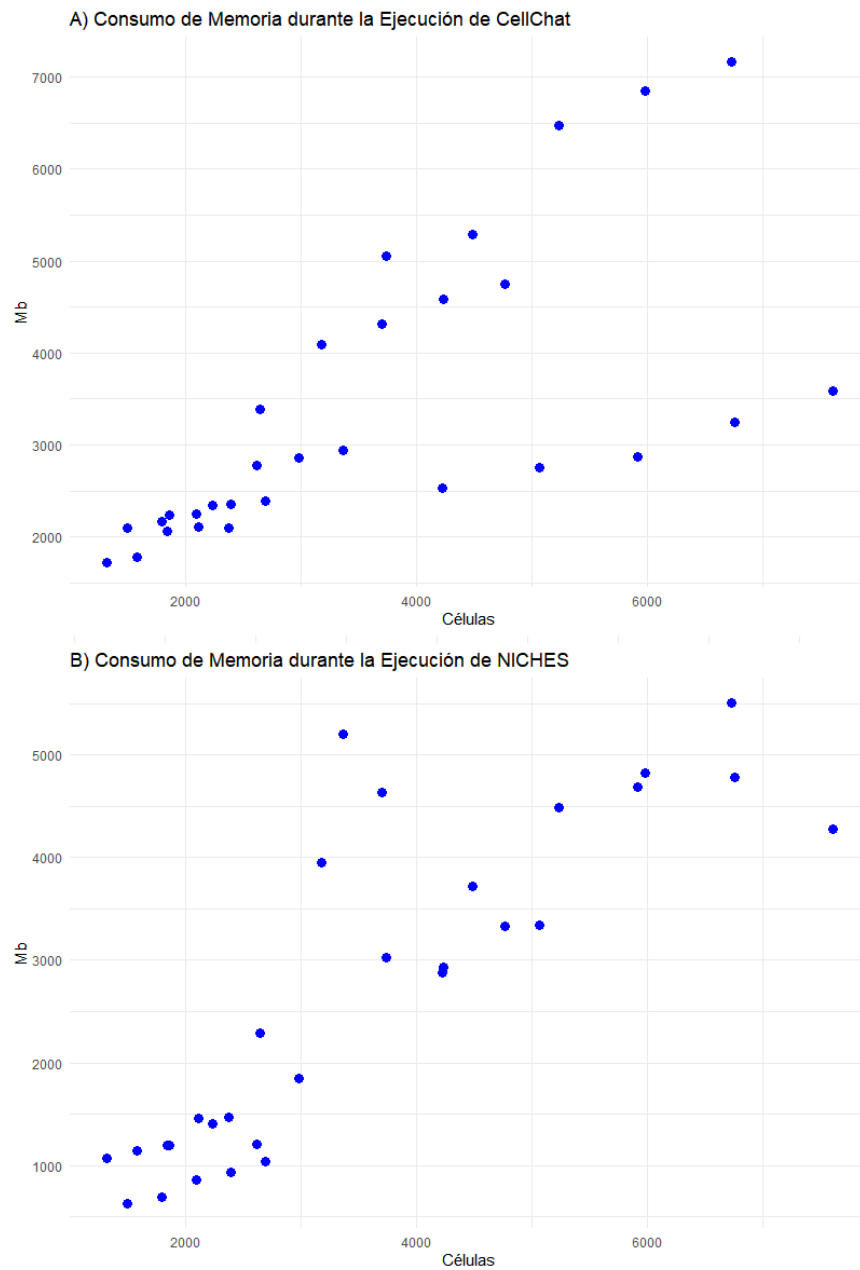


Figura 9: Gráficas de dispersión que ilustran el consumo de memoria de CellChat y NICHES, durante la construcción de la MICC y la evaluación de las interacciones. A) La gráfica de dispersión muestra el consumo de memoria del software CellChat. B) La gráfica de dispersión representa la memoria usada tras ejecutar NICHES. En ambas gráficas, el eje X indica el número de células analizadas en los cinco conjuntos de datos, y el eje Y representa la cantidad de memoria RAM utilizada, en megabytes, por el ordenador durante la ejecución de los códigos.

Los resultados ilustrados en las Figuras 9a y 9b muestran como ambos softwares mantuvieron una relación ascendente entre el número de células y el consumo de memoria. Se puede ver en la Figura 9a que el consumo de memoria de CellChat varió entre 2,000 y 3,000 MB para muestras de alrededor de 2,000 células, y cuando el número de células aumentó a aproximadamente 6,000, el consumo de memoria llegó hasta alrededor de 7,000 MB.

En el caso de NICHES, la Figura 9b ilustra una tendencia menos pronunciada en comparación con CellChat. Para aproximadamente 2,000 células, el consumo de memoria osciló entre 1,000 y 2,000 MB, mientras que, para alrededor de 6,000 células, alcanzó los 5,000 MB.

En conclusión, CellChat presentó un incremento más rápido y lineal en el consumo de memoria, alcanzando hasta 7,000 MB para aproximadamente 6,000 células, mientras que NICHES fue más eficiente en cuanto a consumo de memoria, con un máximo de 5,000 MB.

5. Discusión

Los resultados de este estudio comparativo entre NICHES y CellChat sugieren que ambos softwares presentan fortalezas y debilidades según el contexto y los objetivos del análisis.

5.1. Precisión en la predicción de expresión génica y proteica

En términos de precisión para predecir la expresión génica, CellChat demuestra una mayor exactitud en comparación con NICHES, con medianas de AUPRC más altas en los datos de CAGE. Los valores del RIC también sugieren que CellChat exhibe menos variabilidad en sus predicciones en comparación con NICHES. Sin embargo, la mayor cantidad de valores atípicos en CellChat señala una mayor susceptibilidad a la variabilidad extrema, lo que implica que sus predicciones para ciertos tipos celulares, tejidos o condiciones experimentales pueden no ser completamente precisas. En contraste, NICHES muestra

menos valores atípicos, lo que demuestra que sus predicciones son menos propensas a la variabilidad extrema.

Para la predicción de la presencia proteica, NICHES muestra una mayor precisión, evidenciado al presentar medianas de AUPRC proteómicas más altas en comparación con CellChat. Sin embargo, los valores del RIC de AUPRC proteómico respaldan los de los RIC de AUPRC CAGE de CellChat, ya que exhibe menos variabilidad en la mayoría de sus predicciones en comparación con NICHES por presentar un menor RIC que su contraparte. CellChat presentó resultados similares en AURPC CAGE y proteómica en un *benchmark* anterior, donde los valores AURPC CAGE son mucho menores que los valores AUPRC proteómicos (15).

5.2. Diferencias metodológicas

Estos resultados pueden ser explicados por la diferencia metodológica en los enfoques de estos dos softwares para identificar una interacción LR. CellChat emplea un enfoque probabilístico basado en la ley de acción de masas, lo que le permite asignar una probabilidad a cada interacción LR, calculada a partir de los promedios de expresión génica de las células individuales. Este método puede llevar a la identificación de más interacciones al promediar la expresión entre las células de los grupos, maximizando las probabilidades de detectar ligandos y receptores en los clústeres. Además, la consideración de cofactores y la cooperatividad entre las subunidades proteicas, permite a CellChat capturar un espectro más amplio de posibles interacciones, incluido si algunas subunidades o cofactores están expresados a niveles bajos. Estos enfoques proporcionan a CellChat una mayor capacidad en la precisión al predecir la expresión génica (19). Este hecho se ve evidenciado por los mayores valores de AUPRC CAGE. Sin embargo, este enfoque no garantiza que estos genes predichos se traduzcan en proteínas, ni que las proteínas estén activas o en las concentraciones necesarias para que la interacción ocurra (10). Esto puede explicar por qué CellChat no es tan preciso a la hora de predecir la presencia proteica.

Por otro lado, NICHES al igual que CellChat, emplea un método que preserva los ceros al calcular la expresión de ligandos y receptores. Es decir, si alguno de los componentes, ya sea una subunidad del ligando o del receptor, tiene una expresión nula, la interacción se

considera automáticamente inexistente, asignándole un valor de cero (17). NICHES tiene un enfoque más conservador que CellChat ya que no tiene en cuenta la consideración de complejos multiméricos. Este enfoque a su vez, proporciona NICHES una mayor precisión en la predicción de presencia proteica, ya que las interacciones solo se predicen cuando tanto el ligando como el receptor están expresados a niveles significativos.

Estas diferencias metodológicas y de enfoque también se reflejan en el número de interacciones LR predichas en los diferentes contextos biológicos. CellChat, debido a su enfoque pseudo-bulk y a la aplicación en la ley de masas, realiza un análisis más amplio, manifestando una mayor cantidad de interacciones LR predichas en comparación con NICHES. Además, CellChat mostró una mayor consistencia frente a la variabilidad en los datos de entrada, como se reportó en un *benchmark* anterior (15) donde resultó ser una de las herramientas más consistentes. En el caso del número de ligandos y receptores detectados, NICHES al analizar cada célula de manera individual es más sensible a detectar genes en subpoblaciones específicas de células, como se ha evidenciado en el apartado de concordancia.

5.3. Variabilidad

En términos de variabilidad, los resultados muestran que NICHES presenta una mayor distribución alrededor de la mediana que CellChat. Esto es atribuible a la ausencia de un enfoque de promediado en NICHES, que ayuda a reflejar a la variabilidad inherente en la expresión génica entre diferentes tejidos, mientras que CellChat al promediar la expresión génica suaviza la heterogeneidad celular y tiende a generar un rango más uniforme de interacciones predichas. Aunque NICHES es más conservador y puede subestimar el número de interacciones posibles, mantiene una estabilidad en diversos contextos biológicos. CellChat, a pesar de su consistencia en los datos de entrada muestra una mayor variabilidad en ciertos contextos biológicos, como el tejido cardíaco.

5.4. Eficiencia computacional

NICHES y CellChat presentan también diferencias en la optimización de los cálculos para predecir las interacciones. NICHES es más eficiente en términos de ejecución y memoria

que CellChat. Esto es debido a la forma que tiene NICHES de calcular la MICC a partir de datos de scARN-seq, NICHES comienza construyendo una matriz de adyacencia. Debido a la alta complejidad computacional de esta matriz, NICHES utiliza un esquema de muestreo aleatorio para reducir la cantidad de conexiones consideradas, seleccionando una muestra representativa de las conexiones en lugar de considerar todas las posibles conexiones entre las células. Este enfoque facilita el cálculo y reduce la demanda de recursos computacionales (17). Por otro lado, aunque CellChat presenta optimizaciones en la interpretación de datos mediante técnicas de reducción de dimensionalidad y agrupamiento, no reduce el número de células analizadas (19). En conclusión, NICHES presenta un uso más eficiente de la memoria a la hora de construir e inferir las interacciones LR, gracias a la manera de gestionar los datos, mientras que CellChat al no usar ninguna técnica de optimización tiene un peor rendimiento en cuanto a eficiencia al usar memoria.

6. Conclusiones

A continuación, se enumeran las conclusiones alcanzadas en este estudio:

- NICHES y CellChat muestran diferencias significativas en la precisión de sus predicciones. CellChat es más preciso en la predicción de la expresión génica, presentando valores de AUPRC más altos en los análisis basados en datos de CAGE. Por otro lado, NICHES es más preciso en la detección de la presencia proteica, presentando valores más altos de AUPRC en los datos proteómicos,
- NICHES tiende a ser más conservador y más estable en la predicción de interacciones específicas entre ligandos y receptores, pero a costa de detectar un menor número de interacciones en comparación con CellChat.
- CellChat demostró una mayor consistencia frente a la reducción del tamaño de los conjuntos de datos, manteniendo una mayor proporción de interacciones predichas incluso con reducciones significativas en el número de células.
- NICHES es más eficiente al analizar conjuntos de datos con un número reducido de células, mostrando tiempos de ejecución más cortos y un menor uso de memoria. Sin embargo, su eficiencia disminuye cuando aumenta el número de tipos celulares en los conjuntos de datos, lo que incrementa significativamente su tiempo de ejecución.

7. Limitaciones y perspectivas futuras

- Aunque se evaluaron conjuntos de datos de diferentes contextos biológicos, sería ideal aumentar el número de conjuntos de datos para abarcar una gama más amplia de contextos. Esto permitiría una evaluación más completa de los softwares, dado que pueden existir grandes diferencias en el número de ligandos y receptores, así como en las interacciones LR predichas.
- Debido a las capacidades computacionales de la máquina utilizada, el *benchmark* se realizó con conjuntos de datos relativamente pequeños.
- Una limitación del estudio es el uso de un pseudoestándar de oro y enfoques de evaluación indirectos, debido a la dificultad de establecer un marco experimental exhaustivo que genere datos de alta calidad que sirvan como un estándar de oro fiable para la evaluación y el refinamiento de estos softwares. Esto puede limitar la validez de las conclusiones, ya que estos enfoques comparativos no siempre reflejan de manera completa la realidad experimental.
- Una posible dirección futura incluye la integración de nuevas bases de datos y algoritmos más refinados que permitan analizar interacciones celulares más complejas en diferentes contextos biológicos

8. Referencias bibliográficas

1. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell–cell interactions and communication from gene expression. *Nat Rev Genet.* 2021;22(2):71-88.
2. Lizaveta B. Secondary Messengers – A Role in The Regulation of Cell Function. *J Nephrol Ren Manag.* 5 de marzo de 2022;1:1-8.
3. Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin Transl Med.* 29 de marzo de 2022;12(3):e694.
4. Ramilowski JA, Goldberg T, Harshbarger J, Kloppmann E, Lizio M, Satagopam VP, et al. A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun.* 22 de julio de 2015;6:7866.
5. Naito Y, Yoshioka Y, Ochiya T. Intercellular crosstalk between cancer cells and cancer-associated fibroblasts via extracellular vesicles. *Cancer Cell Int.* 24 de noviembre de 2022;22:367.
6. Boomershine CS, Chamberlain A, Kendall P, Afshar-Sharif AR, Huang H, Washington MK, et al. Autoimmune pancreatitis results from loss of TGF signalling in S100A4-positive dendritic cells. *Gut.* 1 de septiembre de 2009;58(9):1267-74.
7. Hotamisligil GS. Inflammation and metabolic disorders. *Nature.* diciembre de 2006;444(7121):860-7.
8. Rao VS, Srinivas K, Sujini GN, Kumar GNS. Protein-Protein Interaction Detection: Methods and Analysis. *Int J Proteomics.* 2014;2014:147648.
9. Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. *WIREs RNA.* 2017;8(1):e1364.
10. Cheng C, Chen W, Jin H, Chen X. A Review of Single-Cell RNA-Seq Annotation, Integration, and Cell–Cell Communication. *Cells.* 30 de julio de 2023;12(15):1970.

11. Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, et al. Single cell RNA Seq reveals dynamic paracrine control of cellular variation. *Nature*. 19 de junio de 2014;510(7505):363-9.
12. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med*. 18 de agosto de 2017;9(1):75.
13. Armingol E, Baghdassarian HM, Lewis NE. The diversification of methods for studying cell–cell interactions and communication. *Nat Rev Genet*. junio de 2024;25(6):381-400.
14. Dimitrov D, Türei D, Garrido-Rodriguez M, Burmedi PL, Nagai JS, Boys C, et al. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat Commun*. 9 de junio de 2022;13(1):3224.
15. Luo J, Deng M, Zhang X, Sun X. ESICCC as a systematic computational framework for evaluation, selection, and integration of cell-cell communication inference methods. *Genome Res*. octubre de 2023;33(10):1788-805.
16. Liu Z, Sun D, Wang C. Evaluation of cell-cell interaction methods by integrating single-cell RNA sequencing data with spatial information. *Genome Biol*. 17 de octubre de 2022;23:218.
17. Raredon MSB, Yang J, Kothapalli N, Lewis W, Kaminski N, Niklason LE, et al. Comprehensive visualization of cell–cell interactions in single-cell and spatial transcriptomics with NICHES. *Bioinformatics*. 1 de enero de 2023;39(1):btac775.
18. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods*. diciembre de 2016;13(12):966-7.
19. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, et al. Inference and analysis of cell-cell communication using CellChat. *Nat Commun*. 17 de febrero de 2021;12:1088.

20. Raredon MSB, Yang J, Garritano J, Wang M, Kushnir D, Schupp JC, et al. Computation and visualization of cell–cell signaling topologies in single-cell systems data using Connectome. *Sci Rep.* 9 de marzo de 2022;12:4187.
21. Strategies for addressing pseudoreplication in multi-patient scRNA-seq data | bioRxiv [Internet]. [citado 13 de septiembre de 2024]. Disponible en: <https://www.biorxiv.org/content/10.1101/2024.06.15.599144v1>
22. Wilk AJ, Shalek AK, Holmes S, Blish CA. Comparative analysis of cell–cell communication at single-cell resolution. *Nat Biotechnol.* marzo de 2024;42(3):470-83.
23. Raredon et al. - Comprehensive visualization of cell-cell interacti.pdf [Internet]. [citado 14 de septiembre de 2024]. Disponible en: https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/bioinformatics/39/1/10.1093_bioinformatics_btac775/1/btac775_supplementary_data.pdf?Expires=1729217953&Signature=euG PU0glhF8qWcnGzJVND6fNh4AUx1Rn3SN8top9B4pNS0rQbYKNqPuo4dLraIX3peP5h fv0bqd9xV2xVjOzTgVWdtjiC~NwA-j2-rJDQeCFIkn4GEa6otmESM6B7htwMPMNxrAX~C3af1j~4G4V~03rcmQsyPLyEkEFEz Ye3cTdn38XO4xO-9mObcfpQ~eFV3H5wl3DOwd2wndyi29obiaPkHNIhc5B0wKiQk7WdZQuEK~4zqbliuQ XRRmViQgbmAG5xl8h0wZgBTblitqJe3jfVDzSDlvZcX4BTL~e8NPMnPFqBfedf5oWhK clfUafy3hijcWY2F6a5jfCNNuHyw__&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA
24. Almet AA, Cang Z, Jin S, Nie Q. The landscape of cell–cell communication through single-cell transcriptomics. *Curr Opin Syst Biol.* 1 de junio de 2021;26:12-23.
25. Cabello-Aguilar S, Alame M, Kon-Sun-Tack F, Fau C, Lacroix M, Colinge J. SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. *Nucleic Acids Res.* 4 de junio de 2020;48(10):e55.
26. Rieckmann JC, Geiger R, Hornburg D, Wolf T, Kveler K, Jarrossay D, et al. Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat Immunol.* mayo de 2017;18(5):583-93.

27. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol.* mayo de 2018;36(5):411-20.
28. 10xGenomics. 6k PBMCs from a Healthy Donor [Internet]. 2016. Disponible en: <https://www.10xgenomics.com/datasets/6-k-pbm-cs-from-a-healthy-donor-1-standard-1-1-0>
29. 10xGenomics. 8k PBMCs from a Healthy Donor [Internet]. 2017 [citado 29 de agosto de 2024]. Disponible en: <https://www.10xgenomics.com/datasets/8-k-pbm-cs-from-a-healthy-donor-2-standard-2-0-1>
30. Kanemaru K, Cranley J, Muraro D, Miranda AMA, Ho SY, Wilbrey-Clark A, et al. Spatially resolved multiomics of human cardiac niches. *Nature.* 2023;619(7971):801-10.
31. Saviano A, Henderson NC, Baumert TF. Single-cell genomics and spatial transcriptomics: Discovery of novel cell states and cellular interactions in liver physiology and disease biology. *J Hepatol.* 1 de noviembre de 2020;73(5):1219-30.
32. Xiong X, Kuang H, Liu T, Lin JD. A single-cell perspective of the mammalian liver in health and disease. *Hepatol Baltim Md.* abril de 2020;71(4):1467-73.
33. Wang L, Yu P, Zhou B, Song J, Li Z, Zhang M, et al. Single-cell reconstruction of the adult human heart during heart failure and recovery reveals the cellular landscape underlying cardiac function. *Nat Cell Biol.* enero de 2020;22(1):108-19.
34. Payen VL, Lavergne A, Alevra Sarika N, Colonval M, Karim L, Deckers M, et al. Single-cell RNA sequencing of human liver reveals hepatic stellate cell heterogeneity. *JHEP Rep.* 21 de marzo de 2021;3(3):100278.
35. Takahashi H, Kato S, Murata M, Carninci P. CAGE- Cap Analysis Gene Expression: a protocol for the detection of promoter and transcriptional networks. *Methods Mol Biol Clifton NJ.* 2012;786:181-200.
36. Guerrini MM, Oguchi A, Suzuki A, Murakawa Y. Cap analysis of gene expression (CAGE) and noncoding regulatory elements. *Semin Immunopathol.* 1 de enero de 2022;44(1):127-36.

37. Quintana JF, Chandrasegaran P, Sinton MC, Briggs EM, Otto TD, Heslop R, et al. Single cell and spatial transcriptomic analyses reveal microglia-plasma cell crosstalk in the brain during *Trypanosoma brucei* infection. *Nat Commun.* 30 de septiembre de 2022;13:5752.
38. Jou V, Peña SM, Lehoczy JA. Regeneration-specific promoter switching facilitates Mest expression in the mouse digit tip to modulate neutrophil response. *bioRxiv.* 15 de junio de 2024;2024.06.12.598713.
39. Sankowski R, Süß P, Benkendorff A, Böttcher C, Fernandez-Zapata C, Chhatbar C, et al. Multiomic spatial landscape of innate immune cells at human central nervous system borders. *Nat Med.* 2024;30(1):186-98.
40. Rajachandran S, Zhang X, Cao Q, Caldeira-Brant AL, Zhang X, Song Y, et al. Dissecting the spermatogonial stem cell niche using spatial transcriptomics. *Cell Rep.* 25 de julio de 2023;42(7):112737.
41. Sinha S, Sparks HD, Labit E, Robbins HN, Gowing K, Jaffer A, et al. Fibroblast inflammatory priming determines regenerative versus fibrotic skin repair in reindeer. *Cell.* 8 de diciembre de 2022;185(25):4717-4736.e25.
42. Castranio EL, Hasel P, Haure-Mirande JV, Ramirez Jimenez AV, Hamilton BW, Kim RD, et al. Microglial INPP5D limits plaque formation and glial reactivity in the PSAPP mouse model of Alzheimer's disease. *Alzheimers Dement J Alzheimers Assoc.* junio de 2023;19(6):2239-52.
43. Yu X, Xie L, Ge J, Li H, Zhong S, Liu X. Integrating single-cell RNA-seq and spatial transcriptomics reveals MDK-NCL dependent immunosuppressive environment in endometrial carcinoma. *Front Immunol.* 4 de abril de 2023;14:1145300.