VNiVERSiDAD D SALAMANCA

Trabajo Fin de Grado

ANÁLISIS ESTADÍSTICO DE DATOS TRANSCRIPTÓMICOS COMPLEJOS DE MUESTRAS HUMANAS Y USO DE UN MÉTODO DE DECONVOLUCIÓN PARA IDENTIFICAR TIPOS CELULARES ESPECÍFICOS

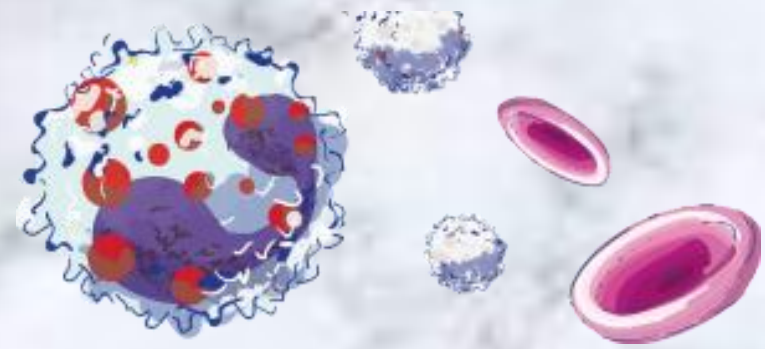
STATISTICAL ANALYSIS OF LARGE-SCALE COMPLEX TRANSCRIPTOMIC DATA OF HUMAN SAMPLES AND USE OF A DECONVOLUTION METHOD TO IDENTIFY SPECIFIC CELL TYPES

Laura Gutiérrez García

Tutores

José Manuel Sánchez Santos

Javier de las Rivas Sanz



GRADO DE ESTADÍSTICA

FACULTAD DE CIENCIAS

Trabajo Fin de Grado

ANÁLISIS ESTADÍSTICO DE DATOS TRANSCRIPTÓMICOS COMPLEJOS DE MUESTRAS HUMANAS Y USO DE UN MÉTODO DE DECONVOLUCIÓN PARA IDENTIFICAR TIPOS CELULARES ESPECÍFICOS

STATISTICAL ANALYSIS OF LARGE-SCALE COMPLEX TRANSCRIPTOMIC DATA OF HUMAN SAMPLES AND USE OF A DECONVOLUTION METHOD TO IDENTIFY SPECIFIC CELL TYPES

Laura Gutiérrez García

Tutores

Dr. José Manuel Sánchez Santos

Dr. Javier de las Rivas Sanz

Dr. Javier de las Rivas Sanz

Dr. José Manuel Sánchez Santos

Laura Gutiérrez García

Salamanca, 2020

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE GRÁFICOS

1. INTRODUCCIÓN

A lo largo de los años, el análisis del transcriptoma ha derivado en grandes avances para el entendimiento de la actividad biológica de los genes. Para ello, los investigadores se han apoyado en nuevas tecnologías disponibles mejorando los estudios realizados y reduciendo las limitaciones con las técnicas que iban surgiendo. Fundamentalmente, la transcriptómica se ocupa del estudio de los niveles de expresión génica en los tejidos siendo su principal objetivo entender cómo y por qué los genes de distintos tipos de células se transforman y la repercusión de estas enfermedades en el desarrollo de algunas enfermedades como el cáncer («Definición de transcriptómica - Diccionario de cáncer - National Cancer Institute», s. f.).

La técnica más novedosa y de creciente uso en la actualidad es la denominada secuenciación de RNA de célula única (scRNA-seq) que pertenece a las tecnologías de secuenciación de nueva generación (NGS: *Next-Generation-Sequencing*) y está permitiendo el estudio de muestras heterogéneas teniendo en cuenta la variabilidad célula a célula y llegando a permitir el poder definir los tipos celulares que están presentes en cada muestra (Hwang, Lee y Bang, 2018). No obstante, a pesar de sus múltiples ventajas, las limitaciones del coste, tiempo, la dificultad para disociar tejidos "fibrosos y pequeños" (como las aneurismas) y los factores de confusión derivados de los "efectos por lote" (Hicks, Teng y Irizarry, 2015), han dado lugar al planteamiento de un método computacional que identifique los tipos celulares existentes sin la necesidad de aislar en el laboratorio las células individualmente.

Recientemente, el método de la deconvolución se ha introducido en este escenario y se han propuesto varios algoritmos a fin de desvelar los tipos celulares ocultos en las muestras de estudio. En el presente trabajo se tratarán las técnicas de deconvolución más relevantes, comparando y concluyendo la más adecuada para las situaciones expuestas.

Comenzaremos el trabajo definiendo los conceptos básicos de la transcriptómica junto con las técnicas más frecuentes en sus análisis de datos. A continuación, se expondrá el problema de la deconvolución desarrollándolo matemáticamente y distinguiendo los dos tipos principales (deconvolución parcial y completa) con los algoritmos destinados a su resolución para el cálculo de una matriz de proporciones celulares (SVR, NNLS, modelo lineal, simplex, ICA).

En la segunda sección, se plantearán los objetivos fundamentales del estudio para comenzar después una tercera sección destinada a la comparación de los métodos en la que se exponen el origen de los datos utilizados, la importancia de una matriz de firmas de referencia (caracterización celular) en la deconvolución parcial, los métodos escogidos para la comparación (*CIBERSORT*, *MIND*, *dtangle*, *linseed* y *deconICA*) y el procedimiento seguido para ejecutar los métodos y el análisis de los mismos mediante las medidas del error (RMSE), correlación de Pearson, divergencia de Kullback-Leibler y los contrastes no paramétricos de Wilcoxon y Friedman.

En el cuarto capítulo, se presentarán los resultados obtenidos tras la evaluación de los métodos en dos conjuntos de datos: simulación y GSE64385, en los que MIND y *linseed* respectivamente, alcanzan los mejores valores en cuanto al error y divergencia. Además, en las muestras biológicas se tienen en cuenta las estimaciones de los tipos celulares indicando que no existe ninguna relación entre los linajes de células inmunes. Por último, en la discusión se interpretarán los resultados anteriores haciendo referencia a artículos similares extrayendo las conclusiones fundamentales que han podido inferirse en la realización de este trabajo.

1.1. EL ANÁLISIS DEL TRANSCRIPTOMA

Desde sus inicios, la biología molecular, entendida como una rama de la biología, ha supuesto un soporte básico en el estudio e investigación de los procesos inherentes a los seres vivos. A simple vista, los seres vivos poseen ciertas características físicas que les hacen diferenciarse entre especies. Sin embargo, las células que componen su interior son las responsables de que se observe una cierta similitud entre ellos.

El interés por el estudio de las células fue aumentando a lo largo de los años, pues relacionado con ellas, se encuentra el proceso de transmisión de la información hereditaria y el desarrollo de la propiedad de replicación, siendo esta última, una propiedad fundamental para el paso de la herencia de una generación a otra. Hoy en día, se sabe que los genes (segmentos de DNA) son los encargados de transmitir esa información genética, pero también, se tiene el conocimiento de que no se trata de una acción aislada, sino que es el resultado de una combinación de fases llevadas a cabo por las dos grandes moléculas del ser humano: DNA y RNA.

A medida que avanzaban las investigaciones respecto a este tema, se iban aportando mayores logros y desvelando nuevas características asociadas a este proceso. A pesar de esto, no fue hasta después de 1950 cuando los científicos hallaron la forma del DNA. Rosalind Franklin fue una de las primeras químicas que contribuyó al descubrimiento de la estructura del DNA y, junto con otros investigadores como Watson y Crick, se estableció su forma de doble hélice (Fierro Correa, 2001). Este hecho condujo también a un conocimiento más amplio acerca de la transmisión genética: el DNA (ácido desoxirribonucleico) está compuesto por cuatro nucleótidos: adenina(A), timina (T), guanina (G) y citosina (C), y su unión por puentes de hidrógenos de forma complementaria (A con T y G con C) es la que configura la estructura que lo caracteriza.

Así, gracias al conocimiento de esta unión de los nucleótidos, se puede descifrar su secuencia en una hebra conociendo la cadena complementaria. La unión de esas dos hebras es lo que se denomina proceso de hibridación y en él se basarán varias de las técnicas que estudian la secuencias específicas de DNA y de RNA (Lleonart, Sánchez, Martín-Duque y Ramón y Cajal, 1997).

Esta última molécula (RNA, ácido ribonucleico) está formada por una sola cadena y es la encargada de transcribir la información genética del DNA. Está formada por los mismos nucleótidos del DNA, a excepción de la timina que es sustituida por el uracilo (U). Puede ser de tres tipos: mensajero (mRNA), ribosomal (rRNA) y de transferencia (tRNA), y cada uno tiene determinadas funciones en la célula, algunas de ellas relacionadas con las proteínas («Definición de ARN - Diccionario de cáncer - National Cancer Institute», s. f.).

La fase de transcripción de la información genética es descrita, junto con la traducción, por el llamado dogma central de la biología molecular, el cual identifica al mRNA como encargado de la síntesis de proteínas en la etapa de traducción (Patiño y Robinson Ramírez Pineda, 2006). De esta manera, el DNA, mediante el proceso de transcripción a mRNA, codifica las instrucciones para la formación de las proteínas indicando en sus segmentos un código específico para cada una de ellas. Estos segmentos son los referidos antes con el nombre de genes y su combinación única conforma lo que se conoce como genoma humano. El estudio del genoma ha desencadenado la unión de diversas disciplinas y su contribución conjunta al análisis de las secuencias de RNA (transcriptoma) ha permitido el descubrimiento de la activación o expresión de los genes en la célula.

Transcriptómica

La transcriptómica estudia los niveles de expresión génica de forma simultánea para determinar perfiles de expresión diferenciados que permitan la caracterización de cada tipo celular, tejido o estado patológico (Cos, 2010). Su entendimiento ha sido esencial, puesto que, con ella, se ha facilitado la comprensión de los elementos funcionales del genoma y se han descubierto componentes moleculares en las células y tejidos que, de una forma u otra, han ayudado a entender el desarrollo y evolución de ciertas enfermedades.

Según Z. Wang, Gerstein y Snyder (2009), la transcriptómica tiene tres objetivos principales que se podrían resumir como:

- Catalogar todas las especies del transcrito (incluyendo cualquier tipo de RNA)
- Determinar la estructura transcripcional de los genes, tanto en relación con sus direcciones y sitios de inicio en la cadena de DNA, como en cuanto a sus patrones de unión y otras modificaciones postranscripcionales.
- Cuantificar los niveles de expresión teniendo en cuenta distintas condiciones y etapas de desarrollo.

El planteamiento de estos objetivos ha supuesto el desarrollo de diversas técnicas, mayoritariamente bioinformáticas, que han conducido a un gran avance en la cuantificación del transcriptoma, pues es a través de estas herramientas (junto con su análisis estadístico e interpretación) como se obtienen los resultados óptimos y de interés para la investigación. Hoy en día, destaca el uso de técnicas de secuenciación de nueva generación (NGS: *Next-Generation-Sequencing*), pero también, siguen conviviendo con ellas los métodos de hibridación que, aunque más antiguos, pueden resultar la opción más adecuada según el objetivo que se pretenda obtener.

Técnicas

Como ya se dijo en el apartado anterior, existen dos métodos apoyados en dos procesos (hibridación y secuenciación) que han dado lugar a las tres técnicas más importantes destinadas al análisis de datos transcriptómicos:

- **Microarrays**: consiste en una colección de sondas (fragmentos de DNA de una sola hebra identificativa de cada gen) que se disponen sobre un portaobjetos de vidrio en posiciones fijas. Esta tecnología trata de detectar el DNA extraído de una o varias muestras que se adhiere en el proceso de hibridación a las sondas, para identificar el nivel de expresión génica con la cantidad de luz que desprende (ya que previamente, el DNA se marca con una sustancia fluorescente) y, por tanto, cuanto mayor sea el nivel de fluorescencia, mayor es el número de copias de DNA que habrán hibridado, por lo que se obtendrá una mayor expresión de éste (Rivas-Lopez, Sánchez-Santos y De las Rivas, 2005).

Figura 1. Microarray de Affymetrix
Recuperado de https://www.oceanridgebio.com/affymetrix_genechip

- **RNA-Seq:** se considera un método de secuenciación de alto rendimiento perteneciente a lo que se conoce como tecnologías de nueva generación (NGS). Se sirve de la secuenciación masiva para calcular la cantidad de RNA presente en una o varias muestras. Al igual que los microarrays, se puede utilizar para determinar los perfiles de expresión génica. No obstante, difieren en cuanto al rendimiento (mayor en RNA-seq) y en su sensibilidad a las isoformas (que son las diferentes variantes de un gen), así como en la identificación de todos los elementos presentes en el transcriptoma (Rodríguez Cubillos, Perlaza Jiménez y Bernal Giraldo, 2014).
- **Single-Cell RNA-seq (scRNA-seq):** aunque también es un método de secuenciación, es una variante más específica de la técnica RNA-seq que consiste en la cuantificación del transcriptoma en células individuales. Gracias a esta técnica, se ha conseguido estudiar la heterogeneidad celular y caracterizar las diferentes poblaciones celulares dentro de una misma muestra (Andrews y Hemberg, 2018).

Hasta hace unos años, los microarrays se mostraban como la técnica más popular para inferir los perfiles de transcripción, pero sus limitaciones asociadas al conocimiento previo de los genes y su incapacidad para identificar variantes genéticas propició la investigación de nuevos enfoques que solventasen estas dificultades.

Actualmente, RNA-seq es una de las tecnologías más demandadas debido a su capacidad en el descubrimiento de nuevos genes, su aplicación en un amplio rango de cuestiones científicas y su alta sensibilidad. Como muestra de ello, varios artículos han revelado su potencial respaldando sus argumentos con varios experimentos, como los relacionados, por ejemplo, en la activación de células T (tipo de glóbulo blanco) (Zhao, Fung-Leung, Bittner, Ngo y Liu, 2014).

La técnica denominada scRNA-seq se considera la mejor opción para estudiar la diversidad celular por la capacidad de analizar independientemente cada célula dentro de la misma muestra. Prueba de ello fue la errónea consideración de las células hematopoyéticas (células sanguíneas originadas en la médula ósea clasificadas en dos grupos: mieloides y linfoides (Mayani et al., 2007)), como una población homogénea y con un comportamiento similar. Esta visión cambió radicalmente en la última década cuando dentro de este tipo de células, se diferenciaron los dos linajes que conforman este tipo de células gracias a esta tecnología (Proserpio y Lönnberg, 2016). Así, la secuenciación célula a célula ha permitido describir un cuadro más detallado sobre los procesos moleculares prescindiendo de la suposición de la no variabilidad intracelular.

El primer paso de la tecnología scRNA-seq consiste en el aislamiento de las células, el cual no siempre resulta sencillo ni muchas veces, eficiente, ya que precisa de marcadores específicos que nos permitan distinguir las poblaciones. Relacionado con este inconveniente, se suman otros asociados al tiempo y al presupuesto que, en ocasiones, la hacen menos accesible para llevar a cabo las investigaciones.

El conocimiento de los tipos celulares que componen una muestra (por ejemplo, una biopsia) y la necesidad de caracterizarlos, es una urgencia en el área biomolecular y, por tanto, se requieren algoritmos que deconvolucionen esa mezcla de expresión génica obtenida, bien a partir de microarrays o de RNA-seq, para descifrar el número y contribución de las diferentes subpoblaciones celulares presentes en la muestra. Además, la información proporcionada por las células individuales puede ayudar en la búsqueda de genes marcadores que actúen como "firmas" de cada tipo celular o tejido. Estas firmas permiten una mayor especificidad en la caracterización de los supuestos subtipos celulares, no visibles dentro de la muestra.

Como se expondrá en este trabajo, la resolución de este problema está ligado a los métodos matemáticos de deconvolución que, fundamentados en diversos algoritmos, sirven para estimar el porcentaje de representación de cada tipo celular en las subpoblaciones de interés o en el perfil de sus genes.

1.2. DECONVOLUCIÓN

El problema de la deconvolución (definición y antecedentes)

La deconvolución se define como una operación consistente en descomponer una mezcla de señales (salida) en sus componentes individuales (entradas). Habitualmente, se le denomina "el problema inverso a la convolución", en el que, de manera generalizada, se sirve de la transformación o combinación de dos funciones para dar lugar a una tercera, es decir, las entradas son conocidas y el elemento a hallar es la salida.

El origen de este planteamiento está relacionado con lo que Cherry (1953) llamó "*cocktail party problem*", un fenómeno que describía la capacidad de las personas para distinguir varios sonidos a la vez (mezclados) y donde entraban en juego muchas variables. Su intención era inferir por qué el ser humano podía discriminar los distintos sonidos (entrada) mezclados en varias voces, música y ruidos simultáneos (salida). Para ello, se toma el ejemplo de una habitación en la que varias personas ubicadas en sitios diferentes hablan a la vez y el cerebro debe separar la voz particular del individuo con el que mantiene la conversación filtrando las señales sonoras que está percibiendo.

Independientemente del estudio e investigación de autores centrados en las habilidades auditivas y perfilando modelos psicoacústicos (Bronkhorst, 2015), el problema de la deconvolución (aplicado matemáticamente) se redirigió hacia otros campos:

- Sismología: en este contexto, la deconvolución de Euler se considera como una herramienta clave para la determinación de los niveles de profundidad de las superficies de contraste (Orihuela, 2015).
- Óptica: destaca por su uso en la eliminación de distorsiones en microscopia y su capacidad de actuación en lugares de poca luz, para su aplicación en imágenes digitales fluorescentes (Mcnally, Karpova, Cooper y Conchello, 1999).
- Análisis de señales: se desenvuelve en torno a la necesidad de restaurar señales y la recuperación de información que se haya visto influida por el desenfoque y el ruido derivados de la convolución. A lo largo de los años, se han desarrollado varias alternativas y algoritmos para hacer frente al problema, no solo como un sistema de identificación lineal e inverso sino como un modelo Gaussiano en el que se pueda introducir información sobre los parámetros (Vincent et al., 2014).

Además de estas áreas, se extendió también hacia otros ámbitos como la transcriptómica en el análisis de expresión génica, asociando en este campo como "salida", la matriz de "mezclas" con los datos de la expresión de los genes en las diversas muestras. Venet, Pecasse, Maenhaut y Bersini (2001) fueron de los primeros autores en plantear una solución al problema de la descomposición celular por muestras (posteriormente conocido como deconvolución) definiendo las tres matrices clave (mezclas, firmas y proporciones) y proponiendo varios algoritmos destinados a encontrar y diferenciar los tipos celulares que conforman las muestras.

Como se ha descrito en los párrafos superiores, la deconvolución es un método asociado a numerosas disciplinas y que, seguramente, con la evolución y el desarrollo de los métodos computaciones se seguirá exportando a otras muchas más que busquen mejorar el efecto indeseable de filtrado (como en óptica) o bien, se extraiga el concepto generalizado centrado en substraer los elementos particulares de una mezcla.

Desarrollo del problema

La formulación del método de deconvolución, tal y como se ha detallado previamente, se puede entender también como una ecuación entre matrices, en lugar de funciones, ajustándose a la siguiente notación:

Sea T la matriz $(n\ x\ m)$ que contiene los valores medidos de las n características o variables en cada una de las m muestras (en Bioinformática es usual colocar las variables en filas y los individuos en columnas). Ésta sería la matriz de "salida" que queremos descomponer o deconvolucionar. Dentro de cada muestra, se distinguen k grupos que, al estar mezclados, se encuentran representados en proporciones diferentes en cada una de ellas.

Así, la matriz T se descompone en el producto de otras dos de la manera siguiente:

$$T = C \times P \tag{1}$$

La matriz C $(n\ x\ k)$ contiene los valores que tomarían las características o variables en cada grupo k y la matriz P $(m\ x\ k)$ contendría las frecuencias relativas o proporciones de los k grupos en las m muestras.

En el ámbito biológico, la matriz T recoge la expresión de n genes en un número m de muestras. El objetivo es descomponerla en el producto de dos matrices, una matriz C llamada matriz de firmas que contiene los valores de expresión de los n genes en los k tipos celulares, otra matriz P llamada matriz de proporciones, que contenga la abundancia (en porcentaje) de cada tipo celular k en cada muestra m.

Para explicar el método más detalladamente y extrapolándolo a otra área distinta a la biológica, considérese el siguiente ejemplo:

Se cuenta con un total de $m = 3$ aulas (muestras) A, B, y C, donde los alumnos de $k = 2$ modalidades (tipos) Ciencias y Letras, se encuentran mezclados. Se les pregunta a los alumnos sobre su interés en $n = 4$ asignaturas (características) Lengua, Matemáticas, Química e Inglés, en una escala del 1 al 10. La matriz T contiene las puntuaciones medias que se han obtenido para cada asignatura en cada aula. La deconvolución pretende encontrar descomposiciones de dicha matriz en el producto de otras dos, de manera que en una matriz C podamos conocer las puntuaciones medias obtenidas por cada asignatura en cada modalidad (ciencias, letras) y en otra matriz P podamos conocer el porcentaje de alumnos de cada modalidad que hay en cada aula.

Organizando los datos como en la ecuación (*1*), podríamos tener que:

$$n = 4\ asignaturas \quad m = 3\ aulas \quad k = 2\ modalidades$$

$$T = \begin{array}{c} \\ Asig. \\ L \\ M \\ Q \\ I \end{array} \begin{array}{c} \begin{array}{cccc} & Aula & & \\ A & & B & C \end{array} \\ \begin{pmatrix} 4.95 & 8.53 & 6.58 \\ 8.2 & 4.9 & 6.7 \\ 6.35 & 3.33 & 4.98 \\ 5.9 & 7.55 & 6.65 \end{pmatrix} \end{array} = \begin{array}{c} \\ Asig. \\ L \\ M \\ Q \\ I \end{array} \begin{array}{c} \begin{array}{cc} & Mod. \\ L & C \end{array} \\ \begin{pmatrix} 9.5 & 3 \\ 4 & 10 \\ 2.5 & 8 \\ 8 & 5 \end{pmatrix} \end{array} \times \begin{array}{c} \\ Mod. \\ L \\ C \end{array} \begin{array}{c} \begin{array}{ccc} & Aula & \\ A & & B \end{array} \quad C \\ \begin{pmatrix} 0.3 & 0.85 & 0.55 \\ 0.7 & 0.15 & 0.45 \end{pmatrix} \end{array} = C \cdot P$$

Cada elemento t_{ij} (valoración observada de las asignaturas por aula) se puede expresar como una combinación lineal de la valoración media por modalidad teniendo en cuenta su proporción en la muestra:

$$t_{ij} = \sum_{k=1}^{2} c_{ik} \cdot p_{kj} + e_{ij} \quad \begin{matrix} i = 1, \dots, 4 \\ j = 1, \dots, 3 \\ k = 1,2 \end{matrix} \tag{2}$$

Donde e_{ij} es un error añadido de la valoración de las asignaturas por muestra, t_{ij} es la valoración de la asignatura i por la clase j, c_{ik} es la valoración de la asignatura i por los de la modalidad k y p_{kj} es la proporción de alumnos de la modalidad k en la clase j.

Para que este problema pueda resolverse, es necesario añadir una restricción a la matriz de proporciones P:

$$\sum_{k=1}^{2} p_{kj} = 1 \quad \begin{matrix} k = 1,2 \\ j = 1,2,3 \end{matrix} \tag{3}$$

La suma por columnas de esta matriz tiene que ser igual a la unidad porque cada una de ellas se puede interpretar como la proporción o probabilidad que tiene esa "modalidad" de presentarse en las distintas "aulas".

En este caso particular, se conocen todos los datos que desglosan la matriz de entrada T, pero generalmente, se intentan calcular los valores de alguna de las otras dos matrices (C o P) o de las dos a la vez, según la información que se conozca. Partiendo de este punto, se pueden distinguir dos tipos de deconvolución asociados a varias técnicas de resolución del problema.

Tipos de deconvolución

Según expone Gaujoux (2013) en su presentación de un método para deconvolucionar muestras biológicas, se pueden distinguir dos tipos de deconvolución de acuerdo a los datos previos que se dispongan:

- **Deconvolución parcial:** es aquella en la que, además de la matriz T, se añade el conocimiento de una de las dos matrices producto (C o P). De esta forma, en función de si se dispone de una matriz u otra, se fijan dos objetivos:
  - Estimar las proporciones P, a partir de la matriz de firmas C: en este caso, los genes deben representar una cantidad mayor que el número de coeficientes que se desean estimar (proporciones de cada tipo celular). Así, el número de incógnitas a calcular sería $m \times k$.
  - Estimar las firmas C de cada tipo celular sabiendo sus proporciones P en las muestras: al igual que antes, hay más muestras que firmas a estimar. No obstante, habría que tener en cuenta la cantidad de genes expresados en las muestras que, normalmente, al superar el número de muestras, convierte esta situación en una estimación menos precisa que la anterior. Aquí, se estiman un total de $n \times k$ valores.
- **Deconvolución completa:** dada una matriz de mezclas T, se infieren tanto la matriz de firmas (C) como la de proporciones (P). Hallar una solución para dicho planteamiento se convierte en un procedimiento difícil de afrontar y, por tanto, muchas veces se requiere la incorporación de nuevas restricciones que simplifiquen y acoten las operaciones o de información biológica complementaria que ayude en la determinación de genes marcadores (filas que componen la matriz de firmas).

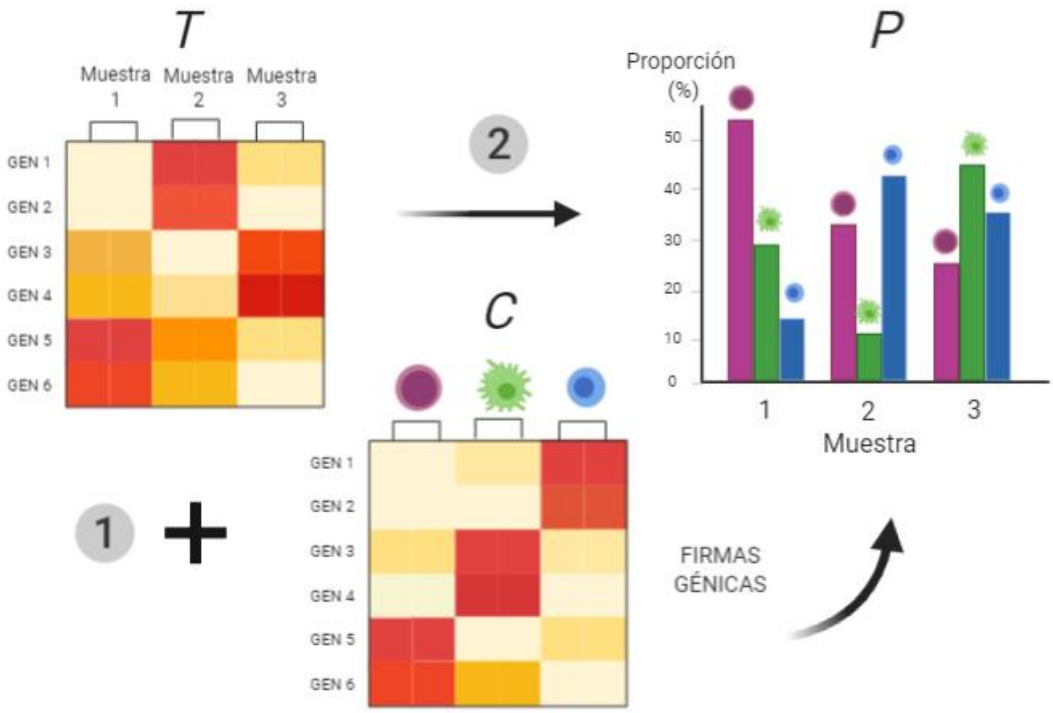


Figura 2. Tipos de deconvolución
1) Deconvolución parcial 2) Deconvolución completa
(Imagen creada con Biorender.com)

Es común que, para la obtención de tales genes marcadores, se empleen datos de "single-cell RNA-seq" en los que se observe la expresión génica en varios tipos celulares y, para cada tipo, se escoja un determinado número de genes que presente un valor diferenciado del resto (habitualmente son los que tienen una mayor expresión en ese tipo celular). No obstante, se debe investigar más a fondo en esta materia, ya que, aún no se conoce con certeza cómo generalizar un procedimiento común a cualquier conjunto de datos o especificar las circunstancias que deben cumplirse para usar esa matriz.

Posibles soluciones: algoritmos

A partir de la clasificación de los métodos de deconvolución en completa y parcial, se han desarrollado diversos algoritmos adaptados a la naturaleza de los datos de entrada. A causa de su gran diversidad, únicamente se explicará el desarrollo de los cinco algoritmos en los que se basarán los métodos seleccionados para este trabajo, que son los más referenciados en la bibliografía científica sobre este tema.

Regresión con máquinas de soporte vectorial (SVR: *Support Vector Regression*)

Las máquinas de soporte vectorial (SVM: *Support Vector Machine*) son un conjunto de algoritmos supervisados que establecen un modelo para asignar a cada individuo de una muestra su clase correspondiente. Para ello, el algoritmo de SVM trata de ir aprendiendo qué clase corresponde a cada observación con una muestra de entrenamiento. Esto lo hace pintando puntos a ambos lados de un hiperplano (línea que separa el espacio en dos mitades) que los divide. Los vectores de soporte se identifican con los puntos más próximos a ese hiperplano (Awad y Khanna, 2015).

Principalmente, este método se usa para problemas de clasificación binaria y regresión, siendo este último caso el que se emplea en la deconvolución de tipos celulares. Al igual que en cualquier modelo de SVM, la técnica SVR considera un hiperplano de separación, pero en lugar de predecir clases como en la clasificación, la predicción es un número real que se estima con un modelo de regresión. Cuando se trata de modelos lineales, el hiperplano corresponde a la ecuación de una recta: $y = wx + b$. A ambos lados, se sitúan unos márgenes a una distancia ε de la recta. Los puntos que se encuentren más próximos a la delimitación de los márgenes, son los vectores de soporte (v) y los que se salgan de ese recinto se utilizarán para calcular su distancia a la banda representado por un valor denominado épsilon que se tendrá en cuenta para la función de pérdida a minimizar:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi + \xi^*) \qquad (4)$$

donde C es una constante positiva que, al aumentar, hace que el error tienda a 0 y, ξ y ξ^*son las variables que controlan el error de entrenamiento al medir la distancia de los puntos que se quedan fuera a ambos lados de la banda.

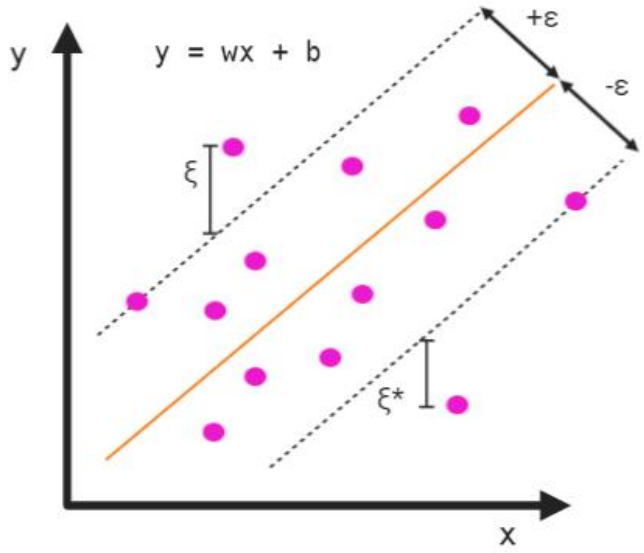


Figura 3. Hiperplano del modelo lineal en el algoritmo SVR.
(Imagen creada con Biorender.com)

En modelos no lineales, se usa un *kernel* (núcleo) para poder representar los datos en un espacio multidimensional y así separarlos linealmente.

Otro aspecto a tener en cuenta es el parámetro a determinar (ε o v), ya que, conduce a dos enfoques para resolver el problema con SVR. El primero se utiliza cuando se desea controlar el error y, el segundo cuando se busca obtener un modelo sencillo con pocos vectores de soporte.

En la deconvolución de muestras biológicas, uno de los métodos aplicado en este proyecto se decanta por la opción de v-SVR identificando como los vectores de soporte a los genes seleccionados de la matriz de firmas y normalizando al final los coeficientes obtenidos en la regresión para que sumen uno y representen los vectores de las proporciones celulares relativas en las muestras.

Mínimos cuadrados no negativos (NNLS: *Non Negative Least Squares*)

El algoritmo de mínimos cuadrados no negativos constituye una particularización del problema de mínimos cuadrados en el que se añade la restricción de no negatividad a los coeficientes. Consiste en minimizar la suma de cuadrados de la diferencia de los valores predichos y los observados. Es decir, dada una matriz A y un vector observado y, el objetivo es encontrar:

$$min_x \|Ax - y\|^2 \quad sujeto\ a\ que\ x \geq 0 \tag{5}$$

También, se define como la minimización de la distancia Euclídea, ya que, la norma que se suele usar es la Euclídea (Avila Cobos, Vandesompele, Mestdagh y De Preter, 2018)

Al aplicarlo en la deconvolución de tipos celulares, se suele añadir otra condición que establece que la suma de las proporciones sea igual a uno (teniendo un rango de 0 a 1). En este contexto, los datos predichos corresponden al producto $C \times P$ y los observados a la matriz T. Con este planteamiento, se pueden abarcar los dos subtipos de deconvolución parcial, que suponen conocidas una de las dos matrices de la multiplicación. Así, si lo que se pretende es calcular la matriz de proporciones, su planteamiento sería:

$$min_P \|C \times P - T\|^2\ sujeto\ a\ que\ P \geq 0 \tag{6}$$

Como inconveniente, puede ocurrir que el método no converja al mínimo absoluto y, por tanto, no se alcance la solución óptima.

Modelo Lineal

Los modelos lineales constituyen una generalización de varios subtipos (como los modelos mixtos que mezclan efectos fijos y aleatorios) basados en regresión. Su aplicación varía en función de la naturaleza de la variable a predecir (variable respuesta) que puede ser continua o discreta y la relación de esta con la variable independiente (variable explicativa), ya que, si no es lineal, es necesario realizar un ajuste con una función de vínculo que linealice dicha relación.

La formulación más sencilla de este modelo es la que considera una relación lineal:

$$y_i = \beta_0 + \beta_1 \cdot x_i \tag{7}$$

donde y es la variable respuesta, x es la variable explicativa y β_0 y β_1, los coeficientes del modelo.

Este caso es el que se observa en la ecuación del problema de deconvolución de tipos celulares, en el que se supone que la expresión de un determinado gen n (n_n) en una muestra determinada, se puede expresar como la combinación lineal de su expresión en el tipo celular k (n_{kn}) junto con su proporción correspondiente en la muestra (p_k):

$$n_n = \sum_{k=1}^{K} p_k \cdot n_{kn} \tag{8}$$

De esta forma, para obtener el modelo final, además de la matriz de expresión génica mezclada (Y_n) se tiene en cuenta también el conocimiento aportado por una matriz de referencias (Z_{krn}) que recoge la expresión de v muestras puras (siendo r una de ellas) en los n genes para cada tipo celular k:

$$Y_n = \mu + \theta_n + \gamma log_2(n_n) + \varepsilon_n$$
$$Z_{krn} = \alpha + \theta_n + \gamma log_2(n_{kn}) + \varepsilon_{krn} \quad (9)$$

Las matrices de expresión de mezclas y de referencia se encuentran en escala logarítmica, por eso, la expresión del gen (n_n) también se ha de transformar a dicha escala. Los parámetros μ y α corresponden a la media de expresión para cada matriz, mientras que θ y ε son el sesgo y la desviación en las ecuaciones. Por último, γ es un factor cuyo valor depende del origen de los datos y que se especifica en el desarrollo del método que recurre a este algoritmo.

Así, como resultado de la combinación de las ecuaciones superiores se establecen los genes marcadores que conforman el conjunto G_k y que determinarán la proporción de cada tipo celular en las muestras.

Simplex

En programación lineal, el método Simplex es el más utilizado para resolver los problemas de optimización (Lewis, 2008). Gráficamente, consiste en representar en el plano las variables de decisión (cada una en un eje) y, sobre ellas, representar las restricciones correspondientes. La región que delimita la figura resultante conforma el espacio de soluciones, siendo las coordenadas de las esquinas (o vértices) las óptimas. Matemáticamente, se plantea como:

$$Max\{z = c^T x\} \;\; sujeto\ a\ Ax \leq b, x \geq 0 \quad (10)$$

siendo $c = (c_1, \dots, c_n)$ los coeficientes de la función objetivo (z) a maximizar (o minimizar) y que ponderan al vector $x = (x_1, \dots, x_n)$ de variables. Las restricciones están determinadas por la matriz A de dimensiones $m \times n$ y el vector de constantes $b = (b_1, \dots, b_m)$ que son los que, en este tipo de ejercicios, establecen los recursos disponibles. Por último, las variables están condicionadas por la restricción de no negatividad que impone un límite inferior para los ejes.

El ámbito de actuación de este algoritmo se ha centrado, fundamentalmente, en la industria donde los principales fines son: maximizar los ingresos (o minimizar los costes) y encontrar las proporciones para la composición de sustancias de acuerdo con unos recursos limitados. Este último fin relacionado con las proporciones de los componentes de una mezcla, se asocia con la investigación de muestras biológicas para cuantificar los tipos celulares presentes en una muestra donde tras un filtrado basado en la mutua linealidad de los genes y un análisis de descomposición en valores singulares (SVD: *Singular Value Descomposition*) que define el número de tipos celulares presentes en la mezcla. Esta información se proyecta en un espacio simplex en el que se asocian las esquinas observadas con las proporciones y genes marcadores de los tipos celulares a inferir.

El algoritmo del Simplex es uno de los cinco desarrollados en este apartado, que se emplea en el tipo de deconvolución completa.

Análisis de Componentes Independientes (ICA: *Independent Compoment Analysis*)

Técnica computacional de datos multivariantes ligada a la resolución del problema de la deconvolución completa en el ámbito del procesamiento de señales. Su principal objetivo consiste en descubrir cuáles son las componentes estadísticamente independientes en una matriz de mezclas (sonidos simultáneos como en el fenómeno "*cocktail party problem*" o expresión génica en el caso de este estudio).

Es similar al análisis de componentes principales (PCA: *Principal Component Analysis*), solo que, en lugar de buscar la no correlación entre las variables, se busca la independencia (que es una condición más fuerte) y los datos no siguen una distribución gaussiana.

Dado un vector de muestras de k componentes independientes (s_k), se puede expresar como un modelo de variables latentes:

$$x_{ij} = \sum_{k=1}^{K} a_{kj} \times s_{ik} \tag{11}$$

O de forma matricial:

$$x = As \tag{12}$$

Una vez que se ha estimado A, se puede calcular su inversa (W) y obtener las componentes independientes mediante:

$$s = Wx \tag{13}$$

Siendo x el vector de muestras mezcladas, W una matriz de separación invertible y s el vector a estimar con las variables latentes (los tipos celulares).

La independencia máxima entre las componentes independientes se consigue a través de una "función contraste" ϕ, es decir, se trata de hallar la matriz W que maximice ϕ(W) para que las variables sean lo más independientes posible (Hornillo Mellado, 2005).

En el análisis de datos transcriptómicos, X es la matriz ($m \times n$) de la expresión génica observada en las muestras, A ($m \times k$) contiene las puntuaciones de los tipos celulares en cada muestra y S ($k \times n$) conforma los pesos de cada gen en los tipos celulares (Kairov et al., 2017).

El análisis de componentes independientes es el último algoritmo utilizado en este estudio orientado a la deconvolución completa, pero también, existen otros enfocados a este tipo como son: el Análisis de Componentes Principales (PCA: *Principal Component Analysis*) que reduce la gran cantidad de variables (genes) en un número más pequeño de componentes (los tipos celulares), la factorización de matrices no negativa (NMF: *Nonnegative Matrix Factorization*) basada en la propiedad de no negatividad de las matrices y en la aplicación de mínimos cuadrados alternantes (ALS: *Alternative Least Squares*) y, finalmente, algoritmos basados en aproximaciones Bayesianas que buscan maximizar una función de verosimilitud (Avila Cobos et al., 2018).

Con estas herramientas, dada la complejidad del problema a resolver, las soluciones aportadas no son tan precisas como las que proporcionan los algoritmos de deconvolución parcial, es por eso, por lo que, en los dos últimos algoritmos definidos cabría esperar peores resultados que en los primeros.

2. OBJETIVOS

El objetivo fundamental de este trabajo es el de realizar un estudio comparativo de métodos analíticos de la señal de expresión génica global (señal transcriptómica) de muestras humanas (obtenidas de biopsias), que incluyen habitualmente una mezcla compleja de células distintas y en las que interesa inferir la proporción de los tipos celulares presentes en ellas. De este modo, se trata de comparar y evaluar métodos de descomposición de señales complejas o de mezclas conocidos con el nombre de ***métodos de deconvolución.***

Para ello, se seleccionan cinco métodos con sus algoritmos matemático-estadísticos de deconvolución correspondientes, en los que, tras evaluar sus resultados, se responderán otros objetivos secundarios relacionados con el principal:

I. Enunciar el problema de deconvolución.
II. Determinar el algoritmo más propicio de acuerdo con la situación de partida.
III. Estudiar las fortalezas y debilidades de cada método.
IV. Hallar las diferencias entre los modelos estableciendo un criterio común de comparación.
V. Analizar y contrastar en un conjunto de datos específico los tipos celulares inmunes más difíciles y más sencillos de estimar para cada método.

A su vez, se plantearán las posibles causas y consecuencias de las estimaciones obtenidas aportando información adicional que refrende los argumentos y busque soluciones óptimas a las dificultades y limitaciones experimentadas.

3. COMPARACIÓN DE MÉTODOS

En total, son cinco métodos los que se estudian en este trabajo, analizando sus resultados y comparando sus estimaciones para determinar cuál o cuáles analizan y descomponen mejor los datos propuestos. Sus nombres son: ***CIBERSORT*** (Newman et al., 2015), ***MIND*** (J. Wang, Devlin y Roeder, 2020), ***dtangle*** (Hunt, Freytag, Bahlo y Gagnon-Bartsch, 2019), ***linseed*** (Zaitsev, Bambouskova, Swain y Artyomov, 2019) y ***deconICA*** (Czerwińska, 2018). Los tres primeros están orientados a resolver el problema de la deconvolución parcial donde se necesita una matriz de referencia que aporte información para deconvolucionar la matriz con la expresión global, y los dos últimos abordan el problema de la deconvolución completa, por tanto, la única matriz que necesitan estos métodos es la de mezclas.

CIBERSORT y *MIND* se apoyarán en la misma matriz de firmas (validada por este primer método) para encontrar la solución óptima cuando se calcula su precisión en datos reales, mientras que *dtangle* obtendrá una matriz de referencia a partir de datos de muestras que contienen un solo tipo celular (muestras puras).

3.1. DATOS

Los datos para analizar en este estudio provienen de dos fuentes: por una parte, se genera un conjunto de datos simulado que prueba la capacidad para deconvolucionar de cada método de un modo más genérico y, por otra, se estudian varias muestras de pacientes humanos, a fin de evaluar dichos métodos en un contexto biológico real.

SIMULACIÓN

La evaluación de la fiabilidad de los algoritmos constituye una parte esencial en la comparación de métodos. Por eso, antes de probar su precisión en un conjunto de datos real, se simulan las tres matrices integrantes de la función de deconvolución para controlar de manera más exacta el error cometido y las proporciones estimadas por cada método. La simulación de estas matrices se lleva a cabo con una función implementada en R del paquete "*deconICA*", creado para presentar un método de deconvolución completa de tipos celulares y del cual se hablará más adelante (Czerwińska, 2018).

La función de simulación se basa en otra existente desarrollada por el autor del paquete "*CellMix"* que crea un modelo NMF (*Nonnegative Matrix* Factorization) aleatorio para simular las matrices. Además, este paquete también proporciona los siete métodos de deconvolución más referenciados hasta ese momento (Gaujoux, 2013). A diferencia de la función incluida en *CellMix*, la función de *deconICA* (`simulate_gene_expression`) no usa una distribución Uniforme para generar la matriz de firmas, sino que permite elegir la distribución o bien, dejar por defecto la Binomial Negativa que es la que más se asemeja a la realidad biológica.

En las poblaciones celulares, al igual que en los organismos, la probabilidad de encontrar una célula de un tipo determinado es mayor si ya se ha encontrado otra de ese mismo tipo, por eso, una de las distribuciones que mejor se adapta a dicha propiedad, es la Binomial Negativa que mide el número de pruebas hasta hallar n éxitos, en nuestro caso, la cantidad n de células de un tipo (Gómez López, 1984).

La función de simulación del paquete tiene la forma:

```
simulate_gene_expresssion(k, n, m, markers = x)
```

Según la sintaxis superior (fichero *Simulated_data_5_deconvolution_methods.R*), habría que indicar el número k de tipos celulares, los n genes, las m muestras y los x marcadores que posee cada tipo celular. El procedimiento que sigue para construir las tres matrices consta de varios pasos:

1) En primer lugar, se crea la matriz de firmas según la distribución Binomial Negativa (o la que se introduzca como argumento a la función) asignando nombre a los genes y a los tipos celulares.
2) Según el número x de genes marcadores que se haya elegido, se origina una lista aleatoria con tantos componentes como tipos celulares y, dentro de ellos, se les asigna ese número de genes marcadores por cada tipo celular.
3) Una vez indicados los genes marcadores, se modifica la matriz de firmas en función del punto de corte establecido (por defecto, *mfold* = 2) para que se diferencien los genes marcadores del tipo celular k respecto de los otros en al menos el doble de señal.
4) Posteriormente, se genera la matriz de proporciones con una distribución Uniforme de dimensiones $k \times n$.

5) Por último, la matriz de mezclas T se obtiene del producto de la matriz de firmas y la de proporciones, añadiendo un "ruido" procedente de una distribución Normal de parámetros de media 0 y desviación 0.05.

Así, se obtiene como resultado una lista de cuatro elementos que contiene las matrices T, C (con la firma de todos los genes de la matriz de mezclas T) y P, junto con la lista ($k \times x$) de los genes marcadores por cada tipo celular. La matriz de firmas que se busca en este tipo de experimentos no es la formada por todos los genes sino la filtrada con los genes marcadores establecidos.

DATOS REALES

La búsqueda de un conjunto de datos real que se adapte a los requerimientos de cada función se considera también una labor importante dentro de la evaluación de los métodos de deconvolución. Ante esto, se debe tener en cuenta la necesidad de contar con varias muestras en las que se conozca (o se pueda inferir de alguna manera) la proporción de cada tipo celular a estimar. Sin este conocimiento previo, no se podría calcular la precisión de los métodos, ni decidir cuál de ellos es el que más se aproxima a las cantidades observadas.

Por otra parte, atendiendo al tipo de entrada que requiere cada método, se incluye un tipo de dato u otro, es decir, en todos los métodos, se introduce como argumento la matriz de mezclas T, pero no siempre se precisa de una matriz de firmas C (como en los métodos de *linseed* y *deconICA*), o esta firma no se considera la mejor opción como referencia. Este último es el caso de *dtangle* donde, preferiblemente se han de introducir varias muestras puras de los tipos celulares que se deseen deconvolucionar para que el algoritmo seleccione aquellos genes marcadores que le aporten más información sobre la discriminación entre los tipos.

Teniendo en cuenta las condiciones anteriores y su empleo en varios artículos relacionados con el tema de estudio (Chen, 2019), se tomó la decisión de utilizar el conjunto de datos que aparece en el trabajo de Becht et al. (2016) y que, en la plataforma de datos genómicos GEO (*Gene Expression Omnibus*), se accede mediante el número: GSE64385 (*https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64385*).

En él se incluyen los perfiles de expresión génica (analizados con la técnica de microarray de *Affymetrix*) de doce muestras humanas: dos de ellas puras, asociadas a la línea celular de cáncer de colón HCT116 y las diez restantes compuestas por la mezcla de un total de cinco tipos celulares inmunes: linfocitos B, T y NK, monocitos y neutrófilos, además de una cantidad constante de células referentes al linaje de cáncer de colón.

A pesar de contar con la presencia de seis tipos celulares, la tumoral (HCT116) no se tendrá en cuenta para solventar el problema, puesto que, en este tipo de estudios, los tipos celulares objetivo suelen ser los del sistema inmunitario. Además, es de este linaje de células inmunes del que se posee una firma génica.

Como no se puede eliminar la presencia del tipo celular HCT116 en las muestras y dado que las doce tienen la misma cantidad de este tipo celular (10 ng), se le puede considerar como un "ruido" dentro de ellas para descubrir también la eficacia de los métodos cuando un tipo celular que no sea de interés no está presente en la matriz de referencias o, en el caso de la deconvolución completa, determinar un número de tipos celulares menor al observado.

GEO permite descargar en R este *dataset* como un objeto que contiene toda la información referente a la matriz de expresión, la nomenclatura de los genes (tanto sondas como su símbolo *HUGO* asociado («Home | HUGO Gene Nomenclature Committee», s. f.)) y la abundancia medida en nanogramos (ng) de cada tipo celular.

La obtención de este objeto se realiza en R descargando dos librerías y a través de la siguiente función (fichero *Real_data.R*):

```
library(Biobase)

library(GEOquery)

GSE64385 <- getGEO("GSE64385", GSEMatrix = TRUE, AnnotGPL =
TRUE)[[1]]
```

Los valores a introducir en la función son: el nombre del número de serie en GEO (GSE64385) y el operador lógico *TRUE* o *FALSE* para indicar si se necesitan las columnas de la matriz que vienen por defecto (`GSEMatrix = TRUE`) y la información actualizada que proporciona *Entrez Gene* (base de datos del NCBI con información específica sobre los genes). Así, manipulando el contenido del objeto GSE, se obtienen la matriz de mezclas T y la de proporciones P.

3.2. CARACTERIZACIÓN CELULAR

La matriz de firmas (C) es considerada una parte importante en la resolución del problema de deconvolución, especialmente en la categorizada como "no completa". En este tipo específico, se requiere de información previa de los genes marcadores y, cuánto más caractericen y diferencien éstos a los diversos tipos celulares, mejor será la estimación que se podrá hacer de las proporciones en las que contribuye cada tipo celular.

El proceso de creación de esta matriz es, todavía, algo complejo, ya que, existe discrepancia en cuanto a la cantidad de datos necesaria para representar la diversidad celular que debe reflejar dicha matriz. Autores como Vallania et al. (2018), defienden la incorporación de datos de varias plataformas de microarrays y de distintos tipos de pacientes sanos y enfermos a fin de evitar sesgos y mejorar la estimación de las proporciones. Sin embargo, en las firmas que presenta *CIBERSORT* (*LM22*) (método que se desarrollará más adelante), sólo se tienen en cuenta muestras de pacientes sanos y una única plataforma de microarrays (*Affymetrix*), lo cual no ha sido un impedimento para obtener unas estimaciones precisas cuando se ha comparado con otros métodos (Avila Cobos, Alquicira-Hernandez, Powell, Mestdagh y Preter, 2020)

En referencia a la publicación citada, Cobos et al. (2020) también coinciden con la afirmación anterior referente a la influencia de la selección de los genes marcadores en la deconvolución de tejidos y muestras complejas, pero no cuando se trata de información procedente de la técnica scRNA-seq porque ya aporta el conocimiento necesario para caracterizar a los tipos celulares. El estudio individualizado de células está permitiendo grandes avances en el conocimiento de la heterogeneidad celular y en la resolución de este problema, pero como ya se dijo en la introducción, aún existen ciertas dificultades (tiempo y coste) que implican la búsqueda de otros medios.

De esta forma, el trabajo asociado a la creación de una matriz de referencia conlleva una investigación en cuanto a los tipos celulares de estudio y el acuerdo de una técnica estandarizada que sea igual para cualquier tejido o linaje de células. Es por esto por lo que, hasta que se investigue más en este campo, se empleará la matriz de firmas *LM22* para los dos métodos que la utilizan (*CIBERSORT* y *MIND*) en el caso real, ya que es la más utilizada en el problema de la deconvolución de tipos celulares hasta la fecha.

El tercer método que necesita una caracterización celular es *dtangle*, sin embargo, para obtener el resultado más preciso, su algoritmo requiere la introducción de varias muestras puras de los distintos tipos celulares a deconvolucionar, siendo el propio método el que determina los genes que más caracterizan cada tipo. Dado que este tipo de datos es más fácil de encontrar con la técnica scRNA-seq, se escoge el conjunto formado por las mismas muestras que se menciona en el artículo referente al método (Hunt et al., 2019).

3.3. MÉTODOS DE DECONVOLUCIÓN

En esta sección, se presentan los métodos de deconvolución escogidos para analizar en este trabajo. En total, son cinco métodos los que se comparan, tres de ellos aplicados a resolver el problema de la deconvolución parcial y, los dos últimos, la completa. En todos los casos, se estima la matriz P (proporciones de cada tipo celular en las muestras), ya que, es esta matriz la que más interesa en este tipo de problemas.

CIBERSORT

Es un método de deconvolución parcial desarrollado por Newman et al. (2015) a fin de cuantificar la proporción relativa de cada tipo celular en una muestra. Para ello, ponen a disposición del usuario una aplicación web y un guion en el programa estadístico R, en los que se pueden introducir tanto datos procedentes de microarrays como de RNA-seq.

Además, a pesar de tratarse de una deconvolución no completa, *CIBERSORT* facilita una matriz de firmas denominada "*LM22*" con la que realizar todo el proceso, en el caso de querer determinar las proporciones en células inmunes. Ésta consta de 22 tipos celulares incluidos dentro de las dos grandes ramas de la hematopoyesis: linfoide y mieloide y que se describirán más adelante en un apartado dedicado a esta matriz.

También, ofrece la posibilidad de crear una firma específica con los datos propios, importando un fichero de muestras de referencia y uno de fenotipos. El primero está formado por la expresión génica observada en varios tipos celulares (contando con varias réplicas en cada uno de ellos) y el segundo indica las muestras que pertenecen a cada tipo celular y si se desea comparar con otra clase o no. De esta forma, se obtendrían las proporciones de los tipos celulares para cada muestra, en función de la expresión de los genes que se encuentran en ambas matrices de entrada (firmas y mezclas).

En su propio artículo (Newman et al., 2015), compara este método computacional con la citometría de flujo, un método experimental que determina el porcentaje de células a partir de un tinte sensible a la luz con el que se puede medir la respuesta de las células a la luz («Definición de citometría de flujo - Diccionario de cáncer - National Cancer Institute», s. f.) . Como resultado, se obtiene una correlación significativa entre ambos, al inferir la proporción de los tipos celulares.

Al igual que otros métodos de deconvolución, el objetivo de *CIBERSORT* es resolver el sistema de ecuaciones lineales descrito en (1) y, para lograrlo, emplea el algoritmo de v-SVR (v-*Support vector regression*) perteneciente a las técnicas de aprendizaje supervisado de las máquinas de soporte vectorial (SVM). El motivo de escoger esta particularización de SVR se encuentra en la necesidad de imponer un límite inferior en el número v de vectores de soporte (genes marcadores) y de uno superior para los errores de entrenamiento. Se parte de tres valores para v (0.25, 0.5 y 0.75) y se elige aquél con el que se obtenga un menor error al comparar el resultado observado y el predicho.

Las proporciones relativas a cada tipo celular están representadas por los coeficientes no negativos de la regresión normalizados a sumar la unidad.

La función empleada para llevar a cabo la deconvolución en R es:

```
res -> CIBERSORT (sig_matrix, mixture_matrix, perm = 100)
```

Siendo *sig_matrix* la matriz de firmas C (por ejemplo *LM22*) y *mixture_matrix*, la matriz de mezclas T con los símbolos de los genes en la primera columna y, como cabecera, los nombres de las muestras. El último argumento corresponde al número de permutaciones que se desee incluir en el análisis, en referencia al cálculo de la distribución empírica de la correlación de los coeficientes para realizar el análisis estadístico correspondiente a la obtención del p-valor. Para este último parámetro, se recomienda un número mayor o igual que 100 porque la operación que se lleva a cabo consiste en generar tantas muestras aleatorias de la matriz T como permutaciones indicadas para calcular en cada una el proceso de deconvolución.

Al final, se obtiene como resultado una matriz compuesta por:

- Las contribuciones de cada tipo celular en las muestras de las que se obtiene la P estimada:
  ```
  P_EST <- res[,1:5]
  ```
- El p-valor correspondiente a la aplicación del método de remuestreo de Montecarlo en cada muestra y que supone como hipótesis nula:
  $$H_0 = La\ matriz\ T\ no\ contiene\ ningún\ tipo\ celular\ presente\ en\ C$$
- La correlación y la raíz del error cuadrático medio (RMSE) entre el valor observado y el predicho por muestra.

LM22

CIBERSORT propone su matriz de firmas *LM22* para deconvolucionar células inmunes (leucocitos). Esta matriz está formada por un total de 22 tipos celulares utilizando la nomenclatura *HUGO* para designar a los genes (rasgo a tener en cuenta para el correcto funcionamiento del algoritmo).

Los tipos celulares que presenta, forman parte del proceso de hematopoyesis, en el que, las células sanguíneas se forman a partir de una sola denominada célula madre. De acuerdo con la clasificación a la que atiende este proceso, los tipos celulares que conforman la matriz de firmas son:

- Rama linfoide:
  - Linfocitos B: maduros (o *naïve*), de memoria y células plasmáticas.
  - Linfocitos T: CD4 (*naïve*, de memoria activos e inactivos), cooperadores foliculares, reguladores y gamma-delta.
  - Células asesinas naturales o "*natural killer*": activas e inactivas.

- Rama mieloide:
    - Mastocitos: activos e inactivos.
    - Mieloblastos: neutrófilos, eosinófilos y monocitos (células dendríticas y macrófagos).

El formato de *LM22* es un fichero de texto delimitado por tabuladores y organizado de la siguiente manera: primera columna con los nombres de los genes y la primera fila con los de las muestras, de esta forma, el resto de los valores se corresponden con la expresión en cada muestra. Para construirla, se seleccionaron los genes que se habían expresado diferencialmente utilizando un contraste bilateral *t* para varianzas distintas, y que se hallaban por encima de un punto de corte establecido dentro de cada muestra. Posteriormente, se establecen varias cribas de acuerdo con una puntuación de enriquecimiento (*Enrichment Scores*: ES) para los genes.

Si los tipos celulares a deconvolucionar no perteneciesen al sistema inmune, se emplearía otra matriz de firmas distinta a *LM22*, pero conservando la misma estructura con la que se organiza ésta. Es importante mencionar que la nueva actualización de *CIBERSORT* denominada *CIBERSORTX*, permite crear esta matriz con datos de scRNA-seq. Dado que no todas las técnicas permiten el empleo de este tipo de datos, en este estudio, se analizará únicamente la versión tradicional, el cual emplea el mismo algoritmo que la nueva versión (SVR).

MIND

A diferencia de otros métodos de deconvolución, el principal objetivo de *MIND (**M**ulti-measure **IN**dividual **D**econvolution)* es estimar perfiles de expresión génica específicos por tipo celular en cada sujeto. Por ejemplo, los precursores de este método (J. Wang et al., 2020) lo aplican a datos de expresión génica de varios sujetos y de distintas zonas del cerebro. No obstante, antes de llegar a determinar tales perfiles, es necesario estimar las fracciones de cada tipo celular, y para calcularlas, aunque cualquier algoritmo puede ser válido, se utiliza el de mínimos cuadrados no negativos (NNLS: *Non Negative Least Squares*).

Al tratarse de una publicación reciente (2020), llama la atención que se emplee un algoritmo tradicional sin añadir ninguna modificación. Por eso, se decidió tenerlo en cuenta para incluirlo en este trabajo. Además, *MIND* comparte sus funciones en código abierto para poder ejecutarlas en R. Así, para poder deconvolucionar una matriz de mezclas a partir de la de firmas con el algoritmo NNLS, se introducen como argumentos estas dos matrices a la función relativa al mismo y se calculan las proporciones para cada tipo celular.

Las dos matrices deben tener el mismo número n de genes, ya que, de no ser así, el algoritmo no puede realizar los cálculos. Por eso, antes de ejecutar *MIND*, conviene seleccionar en la matriz T los genes (marcadores) que forman la matriz de firmas C, reduciendo de esta manera, la dimensión de las filas de la matriz de mezclas a las mismas que la de la matriz de firmas. La función implementada por el programa es:

```
est_frac(sig = C, bulk = T)
```

donde C es la matriz de firmas con los genes marcadores y T, la matriz de mezclas filtrada con dichos genes marcadores. Como resultado se obtiene la matriz de proporciones de cada tipo celular correspondiente a cada muestra:

```
P_EST <- res.mind
```

siendo *res.mind* el nombre del objeto que almacena el resultado de `est_frac()`.

DTANGLE

dtangle es un método orientado a resolver el problema de la deconvolución parcial empleando datos de genes marcadores y de una matriz de referencia asociada a la matriz de firmas. Se basa en un modelo lineal que emplea los dos tipos de escalas más utilizados en este problema (logarítmica y lineal). Parte de dos suposiciones:

- La cantidad de mRNA de una mezcla es la suma de la cantidad que presenta cada tipo celular (propiedad de la escala lineal).
- Las expresiones medidas en escala lineal conforman el modelo lineal, pero con un ajuste posterior que transforma a escala logarítmica los datos.

Su planteamiento convierte a *dtangle* en un método eficiente y robusto que aprovecha las ventajas proporcionadas por ambos tipos de escalas y, evitando algunos problemas reflejados en otros métodos al utilizarlas de una manera distinta a la que *dtangle* propone. Además, los datos pueden ser de microarrays o RNA-seq y el modo en el que obtiene las estimaciones se puede indicar con otra medida que no sea la media, como por ejemplo la mediana, así, se añaden como argumentos estas características a su función en R:

```
dtangle(Y = t(T), references = t(C), data_type = "microarray_gene", summary_fn = "mean")
```

Las matrices de mezcla T y de firmas C deben proporcionase traspuestas al algoritmo y, al igual que en el método anterior (*MIND*), los genes que conforman ambas matrices deben ser los mismos. Como salida, proporciona una lista de cuatro elementos:

- La matriz con las proporciones estimadas:
  ```
  P_EST <- dt_out$estimates
  ```
- Lista de marcadores por tipo celular con su posición en la matriz T.
- Número de marcadores escogido por cada tipo celular.
- Parámetro de sensibilidad gamma (γ) basado en el tipo de datos de entrada (tiene en cuenta fallos producidos en la lectura de cuantificación).

Como aplicación práctica, se usa en el estudio de la respuesta del sistema inmune en la enfermedad de Lyme: como matriz de referencias usa *LM22*, la matriz de firmas que proporciona *CIBERSORT* (método explicado previamente) y, como genes marcadores, se seleccionan el 10% de los que más se expresan diferencialmente en cada tipo celular. Al final del estudio, se comprueba que las proporciones de los tipos celulares se corresponden con el conocimiento previo que se tiene de esta enfermedad, por eso, el empleo de esta herramienta puede resultar una buena opción para los investigadores que busquen estudiar cambios histológicos en la composición celular (Hunt et al., 2019).

LINSEED

*Linseed (**LIN**ear **S**ubspace identification for gene **E**xpresion **D**econvolution)* es un método que, a diferencia de los expuestos en los párrafos superiores, es capaz de abordar el problema de la deconvolución completa sin necesidad de aportar ningún conocimiento previo sobre genes marcadores o firmas celulares que identifiquen cada tipo celular. A veces, el conseguir este tipo de información es complicado, por eso, conviene incluir algún método en el trabajo que se ajuste a tal situación.

Para poder estimar las proporciones de cada tipo celular, *linseed* considera que las señales de expresión que se darían en ellos son aditivas linealmente, es decir, que la contribución de cada tipo celular es proporcional a su fracción relativa en las muestras. Esta suposición se

fundamenta en la propiedad de mutua linealidad de los genes pudiendo representarlo en un subespacio lineal denominado simplex. De esta forma, los vectores de expresión de cada gen (normalizados por filas) se pueden representar como una combinación lineal de las proporciones de los tipos celulares. Gráficamente, las esquinas (o vértices) del simplex serían las firmas y las proporciones de los tipos celulares.

De entre varios algoritmos computacionales, los autores escogen SISAL por su capacidad de identificación de la estructura del simplex (selección de manera recursiva de las variables de entrada "hacia atrás" careciendo de conocimiento previo) y por su adaptabilidad al ruido de los datos (Zaitsev et al., 2019).

Para realizar la deconvolución con *linseed* en R, es necesario ejecutar varios pasos que requieren la intervención del usuario (*Anexo, Gráficos S5-S15, Procedimiento de linseed*):

1) Crear el objeto de tipo linseed con la matriz de mezclas T:
```
lo <- LinseedObject$new(T)
```
2) Construir la red de colinealidad evaluando sus coeficientes, correlación de Spearman y calculando el p-valor correspondiente (en este caso 0.01) para filtrar los genes con los que se queda el modelo:
```
lo$calculateSpearmanCorrelation()
lo$calculateSignificanceLevel(100)
lo$significancePlot(0.01)
lo$filterDatasetByPval(0.01)
```
3) Elegir el número de tipos celulares a estimar según el gráfico representado (por ejemplo, $k = 5$):
```
lo$svdPlot()
lo$setCellTypeNumber(5)
```
4) Proyectar en el subespacio simplex y deconvolucionar según las coordenadas del vector de las esquinas halladas:
```
lo$project("full")
lo$projectionPlot(color = "filtered")
lo$project("filtered")
lo$smartSearchCorners(dataset = "filtered", error = "norm")
lo$deconvolveByEndpoints()
```
5) Visualizar las proporciones de forma gráfica o numérica con una matriz compuesta por los tipos celulares por filas y por columnas, las muestras:
```
plotProportions(lo$proportions)
lo$proportions
```

Al obtener una lista, se accede a la matriz P estimada a través de uno de sus elementos y, para que aparezcan las muestras por filas y los tipos celulares en columnas, se traspone el resultado:

```
P_EST <- t(lo$proportions)
```

DECONICA

Al igual que *linseed*, *deconICA* (***Decon****volution of omic data trrough* ***I****mmune* ***C****omponent* ***A****nalysis*) es un método que se aplica a la deconvolución no supervisada (o completa) y, por tanto, el único argumento a introducir en su función es la matriz de mezclas T. Este método se basa en el algoritmo *FastICA* propuesto por Hyvärinen (1999) para realizar el análisis de componentes independientes.

La matriz T $(n \times m)$ se puede expresar como una combinación lineal de componentes no-Gaussianos (independientes) a partir de su factorización con las matrices S $(k \times n)$ y A $(m \times k)$. De aquí, la matriz de interés es la S, ya que, de ella se obtendrán las firmas y genes marcadores para después calcular la matriz de proporciones P.

La función que lleva a cabo la deconvolución en R, posee la opción de llamar al programa Matlab para implementar la estabilización *Icasso* (Himberg y Hyvärinen, 2003). Este procedimiento ejecuta varias veces el algoritmo con inicializaciones distintas, agrupa las componentes, define los centroides de los grupos y estima lo compactos que son. Sin embargo, dada la temporización del trabajo fin de grado y los problemas que podría ocasionar el recurrir a otro programa informático, se decidió no seleccionar dicha opción y usar la que ofrece R por defecto. La función que realiza la deconvolución se estructura de la siguiente manera:

```
deconica -> run_fastica (T, overdecompose = FALSE, with.names = FALSE, gene.names = row.names(T), samples = colnames(T),  n.comp = k, R = TRUE)
```

En ella, conviene especificar además de la matriz T, si se conoce o no el número de tipos celulares presentes en la muestra para llevar a cabo la descomposición (*overdecompose*), si la primera columna de T contiene los nombres de los genes (*with.names*) o si se pasa como argumento el vector con ellos (*gene.names*), el nombre de las muestras (*samples*), el número de tipos celulares en el caso de que la descomposición no se haga por defecto (*n.comp*) y si el algoritmo se va a ejecutar en *R* o en *Matlab* (*R*).

Como resultado, se obtiene una lista con las siguientes matrices:

- X $(n \times m)$: datos preprocesados después de aplicar el análisis de componentes principales (PCA).
- K $(m \times k)$: matriz que proyecta los datos a las primeras k componentes principales antes de aplicar la transformación lineal.
- W $(k \times k)$: matriz de separación de la mezcla.
- A $(k \times m)$: contribuciones estimadas de cada componente.
- S $(n \times k)$: ponderaciones para los tipos celulares.
- log.counts $(n \times m)$: matriz inicial en escala logarítmica, sin genes repetidos y antes de estandarizar.

De todas ellas, las principales son S y *log.counts* que son las que se usarán en funciones posteriores para determinar los genes marcadores y la matriz P estimada.

Después de calcular la descomposición de las matrices, se utilizan otras dos funciones para hallar la matriz de puntuaciones:

- `deconica_markers_x <- generate_markers(deconica, x)`
  Esta función extrae del objeto que contiene la matriz S (*deconica*), x genes marcadores para cada tipo celular, siendo éstos donde se observa un mayor valor de expresión.
- `deconica_scores <- get_scores(deconica$log.counts, deconica_markers_x)`
  A partir de los genes marcadores obtenidos con la función anterior, filtra la matriz con la expresión logarítmica y obtiene una matriz de puntuaciones con valores positivos representando la abundancia de cada tipo celular en las muestras.

Por último, para convertir dichas puntuaciones en porcentajes, se divide cada valor entre la suma de todos los de la fila (misma muestra):

```
P <- deconica_scores/rowSums(deconica_scores)
```

3.4. PROCEDIMIENTO

El tratamiento de los datos, la aplicación de los métodos y el análisis comparativo de los mismos se lleva a cabo con el software estadístico *R*-versión 3.6.1 (*https://www.r-project.org/),* a través del cual se realizan también todos los gráficos que se muestran en los resultados. Este programa permite la instalación de los paquetes creados por los autores de los métodos de deconvolución, así como el cálculo de medidas estadísticas para evaluar su precisión.

PREPROCESAMIENTO

En la simulación, el primer paso antes de obtener las matrices integrantes de la ecuación de deconvolución, es identificar los elementos que conforman la lista resultante de la función `simulate_gene_expresssion()` (fichero *Simulated_data_5_deconvolution_methods.R*):

- ***Expression***: matriz $n \times m$ que contiene los valores de expresión de los genes en las muestras. Corresponde con la matriz T.
- ***Marker.genes***: lista con los nombres de los x genes marcadores seleccionados para cada tipo celular.
- ***Basis_matrix***: matriz $n \times k$ con los valores de expresión de todos los genes en cada tipo celular.
- ***Prop***: matriz $k \times m$ formada por las proporciones de cada tipo celular en las muestras. Suma uno por columnas y se identifica con la matriz P.

Tras asociar las matrices T y P con sus elementos de la lista, faltaría construir la matriz de firmas C, la cual se obtiene al filtrar las filas de la matriz base con los nombres de los genes marcadores:

```
C <- basis_matrix[mark.genes,]
```

En los datos reales, las matrices a extraer del objeto de *GEO* son T y P. Al contrario que antes, se sigue un proceso algo más complejo para construirlas, ya que, hay que realizar algunas transformaciones (fichero *Real_data.R*):

- **Matriz T**: primero, se extrae la matriz de expresión en la que los nombres de las filas son las sondas en lugar de los genes. Para solucionarlo, se asigna a una variable el nombre del gen (con nomenclatura *HUGO*) y se cambia el nombre de las filas por esta variable (*Gene_Symbol*):

  ```
  T <- exprs(GSE64385)
  Gene_Symbol <- GSE64385@featureData@data[["Gene symbol"]]
  row.names(T) <- Gene_Symbol
  ```

  Así, se obtendría la matriz de expresión T con los datos crudos (sin normalizar) y con la misma nomenclatura que utiliza la matriz de firmas (*LM22*).

- **Matriz P**: se recogen en una tabla las características fenotípicas de cada muestra que proporcionan información relevante sobre los tipos celulares mezclados y su abundancia en cada muestra:

```
cell_prop <- pData(GSE64385)[, c(1, 2, 10, 11, 12, 13, 14, 15, 16, 17)]
```

Después, una vez localizado dicho contenido, al no ser un elevado número de muestras (12 pacientes) y tipos celulares (5), se traspasa manualmente a un *dataframe*, añadiendo tantas columnas como tipos celulares. Por ejemplo, con el tipo celular NK:

```
cell_prop.clean$NK <- c (0, 0, 10, 5, 2.5, 1.3, 0.6, 10, 0.6, 1.3, 2.5, 5)
```

El último paso es pasar los valores absolutos a relativos (porcentaje):

```
P <- cell_prop.clean / rowSums(cell_prop.clean)
```

La matriz de firmas C que se utiliza es "*LM22*" (proporcionada como fichero de texto por *CIBERSORT*). Esta matriz está compuesta por 22 tipos celulares inmunes y, como en este ejemplo solo están presentes cinco de ellos, se eliminan los restantes. Si el símbolo del gen no está como nombre de filas, se cambia para que los métodos no den ningún error al no encontrar una primera columna numérica:

```
C <- LM22

C <- C [, -c (4,15:22)]

row.names(C) <- C [ , 1] #En este caso el símbolo del gen está en la primera columna
```

La matriz de referencia de *dtangle* no es *LM22*, sino que es la formada por el conjunto de varias muestras puras que se encuentra en su propio repositorio: *https://wm1693.app.box.com/s/wjiiblczvo5p5wdt5siml8i6cff87ic3/folder/84154009615.*

En su artículo, se emplea también el mismo *dataset*, por eso, se acude a su página web para descargar el fichero tipo "*rda"* que contiene la matriz de expresión con todas las muestras mezcladas y puras (esta matriz recibe el nombre de *data*). En total son 763 muestras puras que se localizan a partir de la fila 13: 29 células NK, 56 linfocitos B, 298 neutrófilos, 271 linfocitos T y 105 monocitos.

Para este método, se cambia la matriz T por la de *data* (unión de T más las muestras puras) y se le pasa como argumento el índice de las muestras puras (compuestas por un solo tipo celular inmune) (fichero *dtangle_Real_data.R*):

```
pure_samples <- list(NK = c(13:42), Bcells = c(43:99), Neutrophils = c(100:398),Tcells = c(399:670), Monocytes = c(671:776))
```

Por tanto, la función en este conjunto de datos particular quedaría modificada de la siguiente manera:

```
dt_out <- dtangle(Y = data, pure_samples = pure_samples)
```

Al incluir las posiciones de las muestras puras en el conjunto de datos, ya no sería necesario introducir la matriz de referencia. Y, para realizar las comparaciones, como la

función proporciona la abundancia relativa en todas las muestras (incluidas las puras), se seleccionan únicamente las correspondientes a las muestras mezcladas:

```
P_EST <- dt_out$estimates
P_EST <- P_EST [1:12,]
```

DECONVOLUCIÓN

Cada método propuesto exige unos criterios de transformación de las matrices distinto, por eso, en este apartado se especifican los cambios necesarios a realizar en los datos procesados:

Tabla 1. Transformaciones de las matrices en cada método de deconvolución

	SIMULACIÓN	***GSE64385***
CIBERSORT	Añadir una primera columna a T con el nombre de los genes	
DTANGLE	Filtrar la matriz T con los genes de C Trasponer ambas matrices	Cambiar la matriz T a escala logarítmica
MIND	Filtrar la matriz T con los genes de C	
LINSEED	Conviene usar la matriz T filtrada para no originar problemas de memoria	

Tras haber ejecutado la función correspondiente, los resultados se almacenan (como se explicó en el desarrollo de cada método) en objetos de distintas clases: *CIBERSORT* y *MIND* generan como salida una matriz de la que, en el caso de *CIBERSORT*, hay que seleccionar las cinco primeras columnas que son las asociadas a las estimaciones de cada tipo celular. *Linseed* y *dtangle* dan como salida una lista en la que se accede a las proporciones estimadas con el operador $ y, en *deconICA*, la salida de la función *get_scores()* es la que se guarda como matriz P estimada. A partir de este punto, ya se pueden analizar los resultados y comparar las estimaciones de cada método.

ANÁLISIS

La evaluación de los métodos se realiza usando varios estadísticos que nos sirven para comparar el grado de asociación y error de los distintos métodos de deconvolución en las situaciones planteadas:

Coeficiente de correlación de Pearson

Mide el grado de dependencia lineal entre dos variables numéricas indicando mediante el signo el sentido de la asociación (Menhenhall, Beaver y Beaver, 2016).

$$r = \frac{S_{XY}}{S_X S_Y} \qquad -1 \leq r \leq 1 \tag{14}$$

Con este coeficiente se pretende calcular la correlación entre las proporciones estimadas (*Est*) y las observadas (*Real*) empleando en R la función `scores_corr_plot(Est, Real, method = "number", tl.col = "black")` (dentro del paquete *deconica*).

Raíz del error cuadrático medio (*RMSE: Root Mean Square Error*)

Muestra el error cometido entre los valores observados (y_j) y los predichos ($\hat{y}_j$) ponderando aquellos que se distancien más de la media (Camarillo-Peñaranda, Saavedra-Montes y Ramos-Paja, 2013).

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{15}$$

Este valor se calcula con la función implementada en R `rmse()` (del paquete *hydroGOF*) a la que se llama con la función creada manualmente `error(Est, Real, muestra = TRUE)`. El error se mide tanto por muestras como por tipo celular (`muestra = TRUE` o `FALSE`), para así determinar cuáles son los tipos de células que peor estima o en qué muestras le ha sido más difícil al algoritmo determinar los porcentajes. Así, en la ecuación (*15*), $n = m$ o k según corresponda.

Divergencia de Kullback-Leibler

También denominada entropía relativa, es una medida de similitud entre dos distribuciones de probabilidad definida como:

$$D_{KL}(p||q) = \sum_{x \in X} p(x)\log\left(\frac{p(x)}{q(x)}\right) \tag{16}$$

A pesar de que se suele considerar como una distancia, no es simétrica, ya que la divergencia de P a Q no siempre es la misma que la de Q a P. Sus valores son no negativos, de tal forma que, cuanto más se aproxima a 0, más se parecen dichas distribuciones (Cover y Thomas, 2005).

La función (fichero *DIVERGENCIA.R*) empleada para calcularla es: `divergenciaKL(Est, Real)` creada a partir de la función `KL()`(del paquete *philentropy*) donde la distribución P corresponde con las proporciones reales (*Real*) en las muestras y la Q con las estimadas (*Est*). Al contrario que en el error, la divergencia solo se estima por muestras, ya que, al tratarse de una comparación entre distribuciones de probabilidad, se debe cumplir la restricción de sumar uno que se tiene en cuenta al hallar los porcentajes en las muestras.

Test no paramétricos

Una vez obtenidos los valores correspondientes al error y a la divergencia, se estudia si existen diferencias significativas al realizar la deconvolución con un método u otro y, para ello, se emplean dos contrastes estadísticos no paramétricos a un nivel de significación del 5%:

- **Test de Friedman para muestras relacionadas**: con esta prueba se analiza si existen diferencias significativas en cuanto al error o la divergencia entre los cinco métodos:

  `friedman.test(valor, método, muestra)`

  Indicando en la función el valor que se desea comparar (error o divergencia) y el método y muestra al que pertenece.

  La hipótesis que se plantea es la siguiente:

  $H_0 = El\ error\ o\ divergencia\ se\ considera\ igual\ en\ todos\ los\ métodos$

- **Test de Wilcoxon para datos apareados:** mediante este contraste, se busca entre qué pares de métodos hay diferencias:

```
pairwise.wilcox.test(valor, método, paired = TRUE,
p.adjust.method = "holm")
```

En esta función, además del valor y el método, hay que especificar que las muestras están relacionadas y qué tipo de método se escoge para calcular la corrección del p-valor. En este caso, se elige el método de Holm que introduce algunas mejoras respecto al de Bonferroni.

Por parejas se busca contrastar:

$$H_0 = No\ hay\ diferencias\ en\ el\ error\ (o\ divergencia)\ entre\ el\ método\ A\ y\ el\ B$$

En los datos reales, es importante recordar que las dos primeras muestras corresponden al linaje celular de cáncer de colon HCT116, el cual no se incluye dentro de la matriz de firmas ni como muestra pura en el método *dtangle*. Además, en todos los métodos salvo en *linseed*, las proporciones por muestra han de cumplir la condición de suma uno, por tanto, en estas primeras muestras, los valores que se obtendrían en el error y la divergencia serían valores extremos que impedirían una correcta visualización cuando se representasen dichas medidas en gráficos. Por este motivo, en la evaluación del GSE64385, se omiten las dos primeras filas, comparando las diez restantes.

4. RESULTADOS

Previamente a la ejecución y presentación de los resultados obtenidos en los cinco métodos, se muestra el enunciado del problema de deconvolución con los datos escogidos:

El punto de partida en cualquier tipo de deconvolución es la matriz de mezclas T (*Gráfico 1*)En ella, se recoge la expresión de los genes en cada muestra, mostrando un nivel de intensidad mayor o menor que será analizado posteriormente para asociarlo a los genes marcadores tanto en los datos simulados como en los reales.

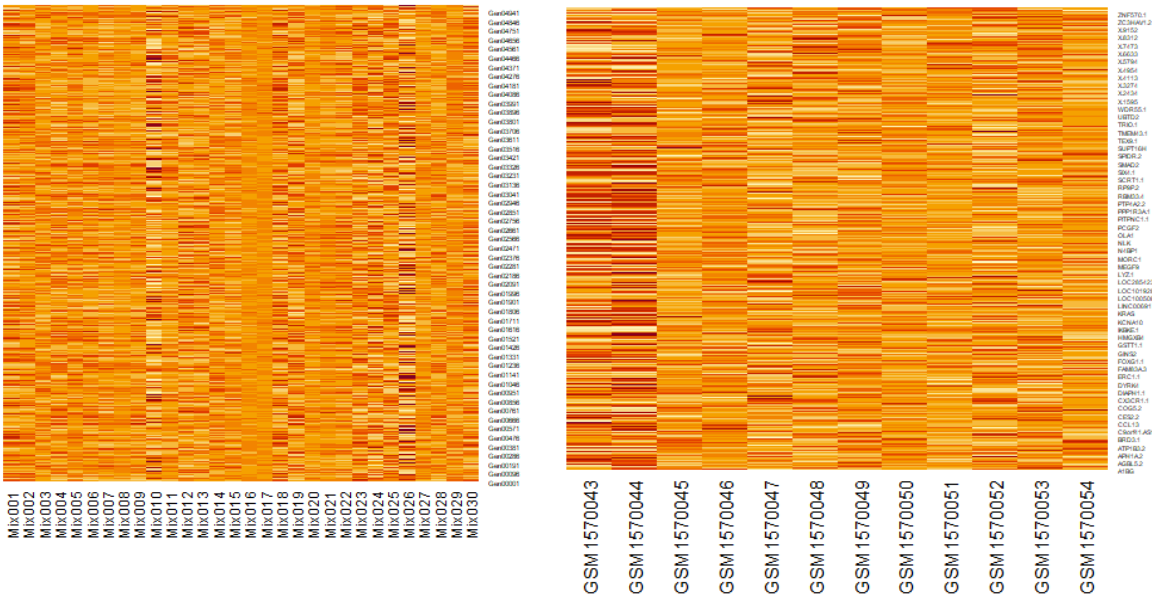


Gráfico 1. Heatmaps matriz T
muestras simuladas (izquierda) y reales (derecha)

En los datos reales (GSE64385), al ser una matriz de menor dimensión, se aprecia mejor el nivel de expresión en cada muestra pudiendo observar una clara diferencia en las dos primeras, que son las asociadas al tipo celular HCT116.

La segunda matriz a tener en cuenta es la matriz de firmas o de referencia C, donde los genes poseen un mayor valor de expresión según el tipo celular del que sean marcadores (*Gráfico 2*).

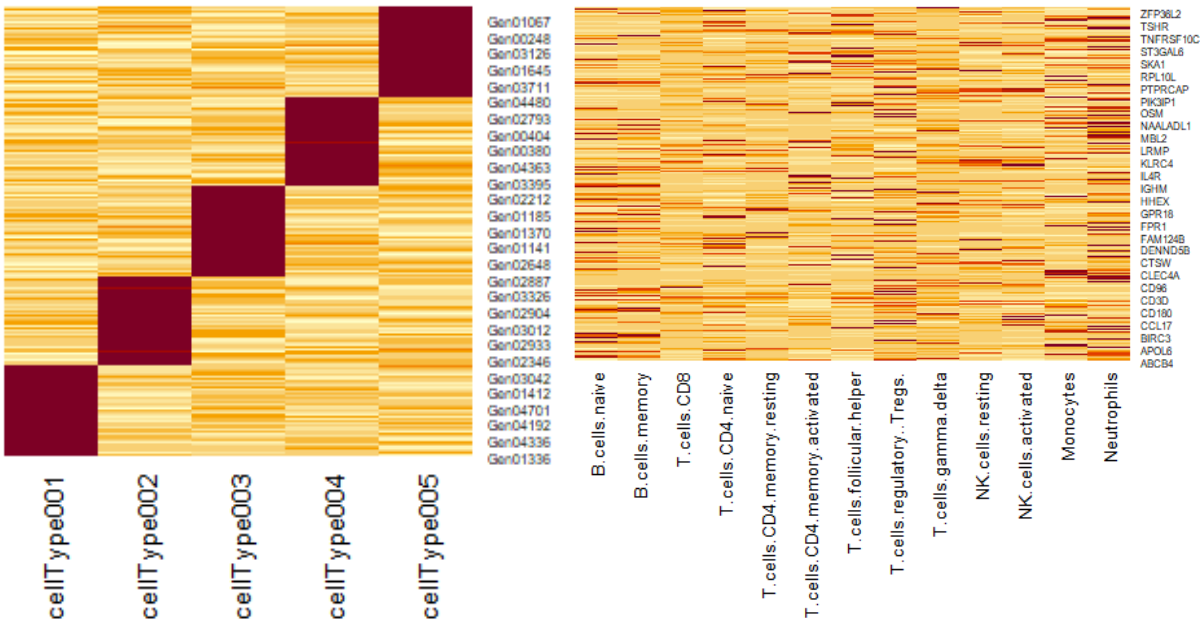


Gráfico 2. Matrices C
datos simulados (izquierda) y datos reales, *LM22* (derecha)

La matriz *LM22* (matriz de firmas en el GSE64385) está constituida por 22 tipos celulares inmunes, sin embargo, en el gráfico superior, se muestran solo 13 de ellos. La razón de este hecho se debe a que el *dataset* de muestras biológicas cuenta con la presencia de estos tipos determinados, por tanto, para evitar efectos de confusión en los métodos, se eliminan los restantes. Otro aspecto a tener en cuenta es la subdivisión de los tipos celulares, los linfocitos B se representan en el *heatmap* mediante los linfocitos B *naïve* y los de memoria, por eso, como varias firmas pertenecen a un mismo grupo, la solución por la que se opta es no modificar la expresión de los genes en esos subtipos (porque esto puede suponer pérdida de información para la matriz de firmas y reducción de precisión en el cálculo de las proporciones) y, posteriormente, ajustar los porcentajes mediante la suma como se indicó en el procedimiento.

De este modo, cada tipo celular se encuentra ligado a un determinado número de genes marcadores que ayudarán en la predicción de la matriz de proporciones P: cuando se trate de una deconvolución parcial, estos genes se obtendrán de la matriz de firmas C y, en el caso de la completa, de la matriz de mezclas T.

Al final, tras aplicar los métodos de deconvolución, se pueden descifrar los porcentajes relativos estimados para cada tipo celular en las muestras, a través de la visualización de la matriz P. Esta matriz se compone de tantas filas como muestras formen la matriz inicial de expresión T, por eso, del primer *dataset* se obtienen un total de 30 gráficos con la representación de la abundancia relativa en cada muestra y, en el segundo, 10, ya que, las dos primeras muestras no están compuestas por ningún tipo celular inmune.

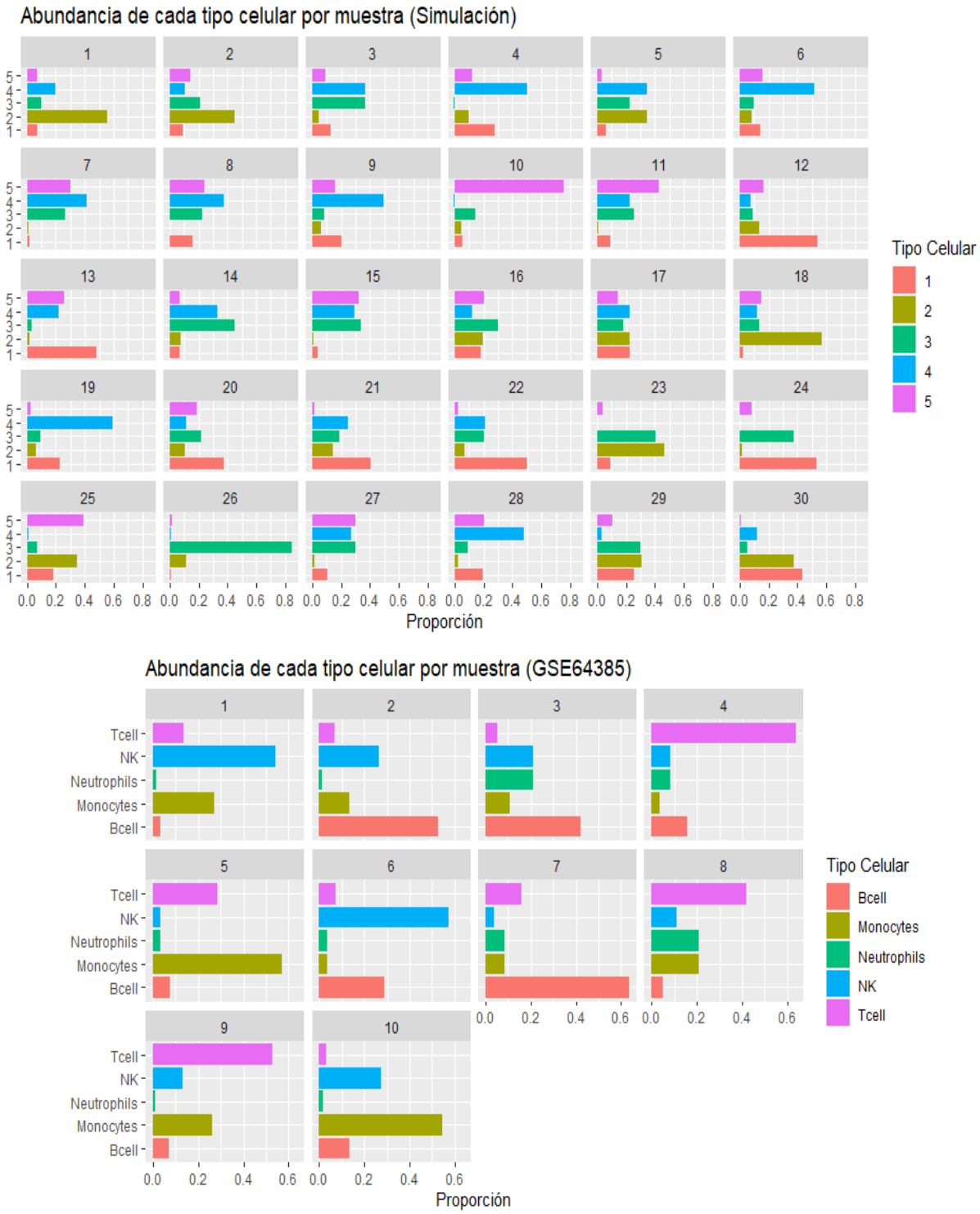


Gráfico 3. Diagramas de barras matrices P
datos simulados (superior) y reales (inferior)

En el *Gráfico 3* aparecen representadas las proporciones correspondientes a cada conjunto de datos en las muestras, variando los porcentajes de los tipos celulares en cada una y observando generalmente la predominación de un tipo concreto. Estas frecuencias relativas pueden verse con más detalle mediante la visualización de gráficos de sectores individuales como por ejemplo la muestra 2 en los datos simulados y la 5 en el caso real (*Gráfico 4*):

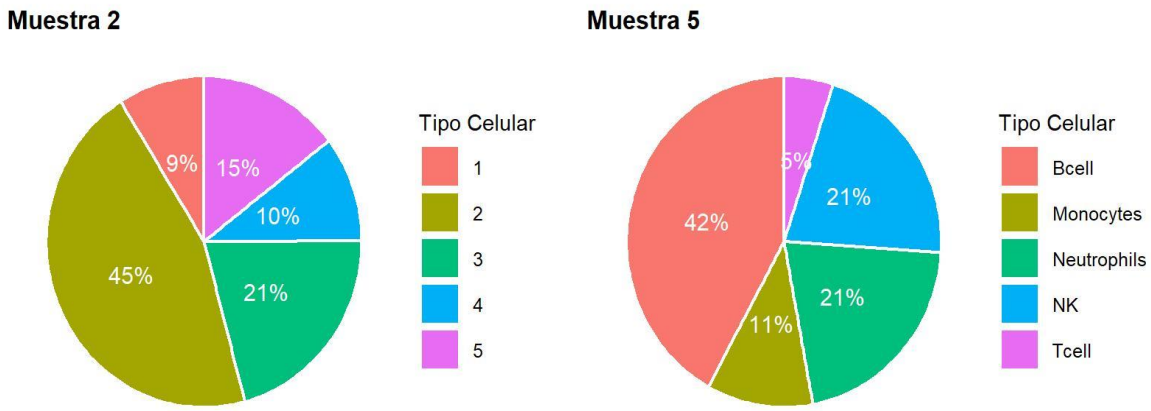


Gráfico 4. Proporciones en 2 muestras simulados (muestra 2) y reales (muestra 5)

En los datos simulados, el tipo celular con mayor abundancia (45%) es el 2, mientras que, en el GSE64385, los linfocitos B son los que presentan un mayor porcentaje (42%). En cada muestra, ya sea de un conjunto de datos u otro, las proporciones adquieren diferentes valores predominando en ellas tipos celulares distintos (el resto de gráficos pueden encontrarse en el *Anexo, Gráficos S1-S4, Proporciones muestras simuladas y GSE64385*).

Para cada método se estiman dichas proporciones obteniendo diferentes medidas para la correlación, error y divergencia. A continuación, se exponen los resultados obtenidos en cada uno de manera individual, distinguiendo la fuente de obtención de los datos.

CIBERSORT

CIBERSORT es uno de los métodos mejor referenciados en la literatura científica en cuanto a la estimación de tipos celulares inmunes. Además, como en cualquier otro método de deconvolución parcial, permite el uso como argumento de una matriz de firmas propia para poder extrapolar el problema a otros tipos celulares distintos a los recogidos en su matriz de firmas *LM22*.

En este trabajo, al analizar dos conjuntos de datos (simulados y reales) se emplean dos matrices de referencia (una propia para los datos simulados y *LM22* para los reales) evaluando su eficiencia en ambas situaciones. La comprobación de las estimaciones de la matriz P se puede realizar mediante la representación de un *heatmap* (*Gráfico 5*) en el cual se puedan identificar los tipos celulares predominantes en cada muestra.

En los datos simulados (imagen superior) se puede observar una gran semejanza entre ambos, coincidiendo prácticamente de manera exacta las proporciones celulares (reflejadas en la intensidad del color) en cada muestra. En el GSE64385 (imagen inferior), cabe destacar que, a pesar de equivocarse en las dos primeras muestras, estas corresponden al tipo celular no incluido en la matriz de firmas y, además, la matriz de proporciones se encuentra supeditada a la restricción de sumar uno por muestras, por lo que reparte las cantidades entre los tipos celulares que reconoce.

Gráfico 5. Heatmaps CIBERSORT
matriz P simulada (superior) y GSE64385 (inferior)

También, es cierto que, en este caso, la estimación no parece tan precisa como antes, ya que se aprecian diferencias en cuanto a la abundancia de los tipos celulares más representados, siendo los linfocitos B y los neutrófilos los peor estimados.

En relación con esta última interpretación, se puede comprobar numéricamente esa suposición con el cálculo de la raíz del error cuadrático medio por tipo celular. En la siguiente tabla, se presenta esta medida en cada conjunto de datos, omitiendo las dos primeras muestras en los datos reales ya que, este método solo estima los tipos celulares que componen la matriz de firmas que emplea para la deconvolución.

Tabla 2. RMSE por tipo celular en cada conjunto de datos en *CIBERSORT*

SIMULACIÓN				
CT1	**CT2**	**CT3**	**CT4**	**CT5**
0.03217041	0.02636729	0.02863227	0.03711309	0.03457494
GSE64385				
NK	**Bcell**	**Neutrophils**	**Tcell**	**Monocytes**
0.0751831	0.2246892	0.2661066	0.1060680	0.1049114

En los datos simulados, el error es mínimo, se encuentra aproximadamente en un 3% en todos los tipos celulares, lo cual representa un valor bajo en comparación con el segundo conjunto de datos (GSE64385) donde los neutrófilos alcanzan un error de casi un 27% seguidos por los linfocitos B con un 22.5%. Tal y como se observó en el *heatmap*, se confirma que estos tipos celulares pertenecientes a linajes distintos (linfoide y mieloide) sean los peor estimados en estas muestras.

DTANGLE

Siendo uno de los métodos más recientes, *dtangle* resuelve, al igual que *CIBERSORT*, el problema de la deconvolución parcial con la especificación de una matriz de referencia obtenida a partir de muestras puras. Por eso, en los datos reales, en lugar de usar *LM22* como en los demás métodos, se aportan los niveles de expresiones en muestras puras. En los datos simulados, esta condición no afecta al algoritmo porque al originarse mediante una simulación, las firmas generadas proceden de los mismos datos de expresión de la matriz de mezclas y pueden, por tanto, considerarse como muestras puras.

Comparando las matrices P estimadas con las originales, se observa como mediante este método, existe una mayor discrepancia que con el anterior (*Gráfico 6*).

En la primera situación, aunque se aprecia una intensidad de color distinta, se puede establecer una asociación entre los tipos celulares más abundantes en cada muestra (representados de un color rojo más fuerte) que coinciden con los observados en el caso original. Las diferencias, sin embargo, parecen mayores cuando se visualiza el *heatmap* de los datos reales, ya que, aunque no se tuvieran en cuenta las dos primeras muestras (como en *CIBERSORT*), los linfocitos T son los que presentan una mayor proporción en la estimación y, tal y como se puede ver en la matriz original, estos no siempre son el tipo celular más abundante en todas las muestras. Igualmente ocurre con los linfocitos B que, según la matriz estimada, son los que poseen una menor abundancia en las muestras y, en la matriz P original, se muestra la variación de su cantidad relativa en cada muestra.

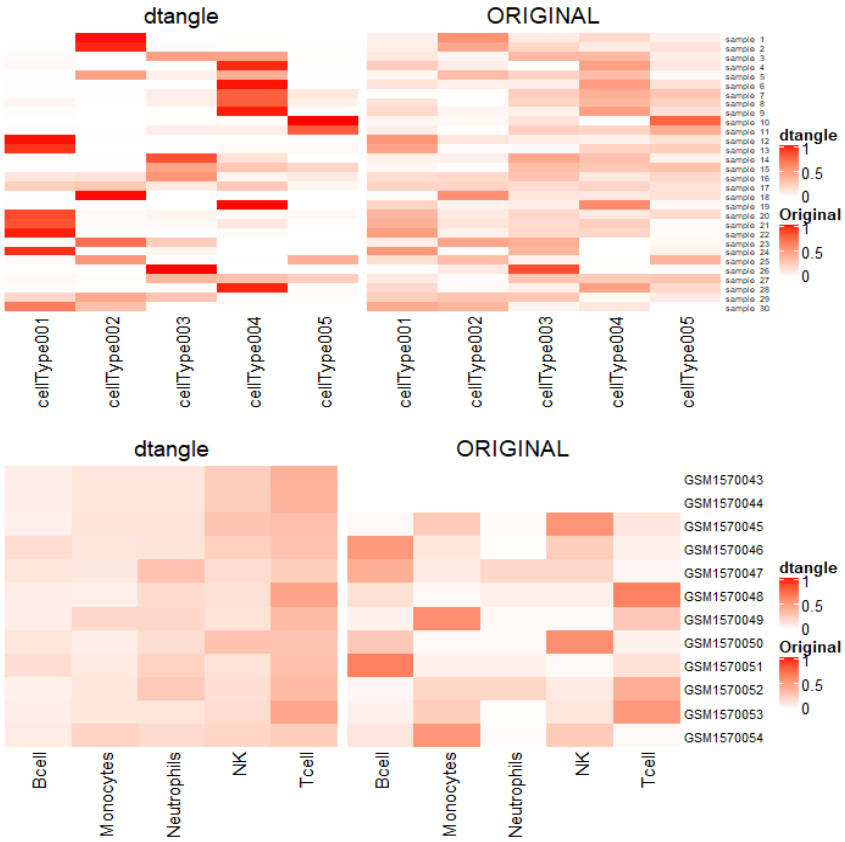


Gráfico 6. Heatmaps dtangle
matriz P simulada (superior) y GSE64385 (inferior)

Si calculamos el RMSE por cada tipo celular siguiendo el mismo procedimiento que en el método anterior:

Tabla 3. RMSE por tipo celular en cada conjunto de datos en *dtangle*

SIMULACIÓN				
CT1	**CT2**	**CT3**	**CT4**	**CT5**
0.2266170	0.1704622	0.1528671	0.2307306	0.1359358
GSE64385				
NK	**Bcell**	**Neutrophils**	**Tcell**	**Monocytes**
0.1290658	0.2128128	0.1337328	0.1759773	0.1673039

El mayor error se comete en el tipo celular 4 en los datos simulados con un valor de un 23% y, en los reales, las medidas de los valores del error más elevadas son similares con un 21.3% en los linfocitos B y un 17.6% en los linfocitos T, ambos procedentes del mismo linaje.

MIND

Es el último método empleado en este estudio que requiere de una matriz de firmas para obtener la matriz de proporciones. La principal función de *MIND* consiste en determinar perfiles de expresión específicos de cada sujeto de acuerdo a la posición del tejido analizado. No obstante, en su paquete de R, se encuentra la función destinada a deconvolucionar y que se aplica al cálculo de las proporciones celulares para posteriormente, compararlas con el resto de métodos. De manera individual, se sigue el procedimiento realizado también con los otros métodos de comprobar la similitud de los valores estimados y los observados a través de los *heatmaps* (*Gráfico 7*):

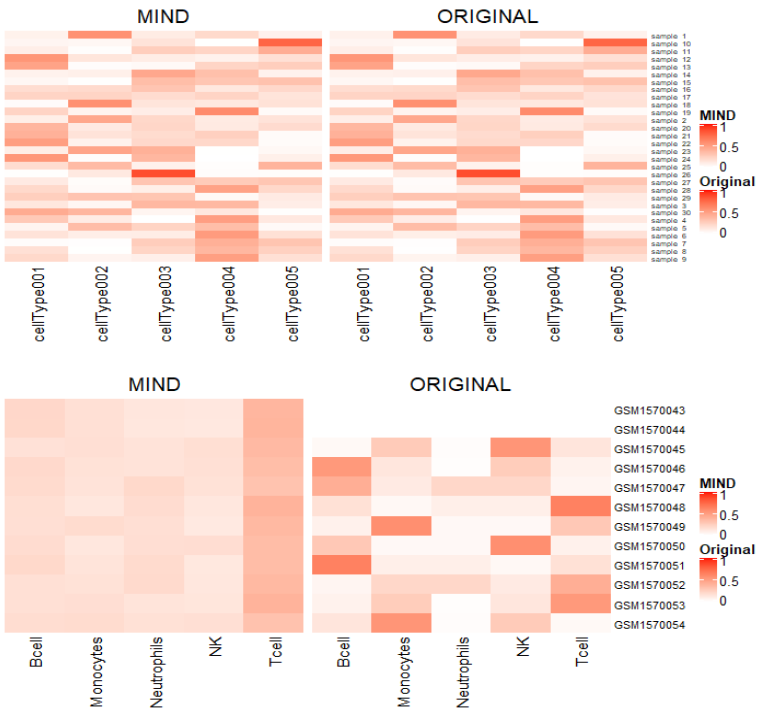


Gráfico 7. Heatmap MIND
matriz P simulada (superior) y GSE64385 (inferior)

Es evidente que, en la simulación, la estimación puede considerarse una réplica de la original, pero en el caso real, las proporciones que calcula *MIND* se alejan no solo en las dos primeras muestras, sino también en todas las demás, dando la impresión de que existe la misma proporción de cada tipo celular en todas ellas con una mayor predominación de los linfocitos T. Excluyendo las dos primeras muestras en el conjunto de datos real, se obtienen los siguientes valores numéricos referentes al error cometido:

Tabla 4. RMSE por tipo celular en cada conjunto de datos en *MIND*

SIMULACIÓN				
CT1	**CT2**	**CT3**	**CT4**	**CT5**
0.001624655	0.001631106	0.001475697	0.001517361	0.001521688
GSE64385				
NK	**Bcell**	**Neutrophils**	**Tcell**	**Monocytes**
0.1837083	0.2022758	0.1141074	0.2165913	0.1797239

En la primera situación, el error es prácticamente inexistente, quedándose muy por debajo del 1% en todos los tipos celulares y como ya se dijo, replicando de una forma casi perfecta las proporciones. En la segunda, se observa un acusado aumento de esta medida en todas las células inmunes siendo, como en *CIBERSORT* y *dtangle*, los linfocitos T los que poseen un mayor error.

LINSEED

El conocimiento previo de una matriz de referencia que ayude en la deconvolución de la matriz de mezclas T, no siempre es posible, por eso, se desarrollan métodos como *linseed* capaces de resolver lo que se ha denominado como una deconvolución completa. Una particularidad de su algoritmo es la ausencia de la restricción de sumar uno en las muestras, lo que a veces implica una pérdida de precisión cuando todos los tipos celulares componen la muestra o un aumento cuando no se da tal situación.

En los *heatmaps* (*Gráfico 8*), la primera característica distintiva de *linseed* respecto a los demás, es el orden de los tipos celulares estimados. En ambos conjuntos de datos, es necesario asociar según el coeficiente de correlación (medida que se mostrará en la comparación de los métodos) el *cluster* que corresponde a cada tipo celular. Tras la asignación, se puede descubrir cierta similitud entre las proporciones ya que, tanto en los datos simulados como en el GSE64385, los tipos celulares que más predominan en cada muestra de la matriz estimada coinciden con los de la observada. No obstante, en ambos conjuntos de datos, algunos tipos celulares como el 3 (asociado al tipo celular 5 estimado en el primero) y los neutrófilos, presentan una diferencia notable en el valor de la proporción en las muestras.

Por otra parte, es importante mencionar que, dado que los otros métodos imponen la condición de sumar uno en sus muestras, en *linseed* también se lleva a cabo este ajuste en la simulación, pero no en los datos reales. De esta forma, se puede ver como este método permite detectar la ausencia de las células inmunes elegidas para la deconvolución en las dos primeras muestras. Y así lo demuestra el cálculo del error, en el que se observa (al contrario que en los otros métodos anteriormente explicados) una disminución al cambiar de una simulación a un *dataset* real.

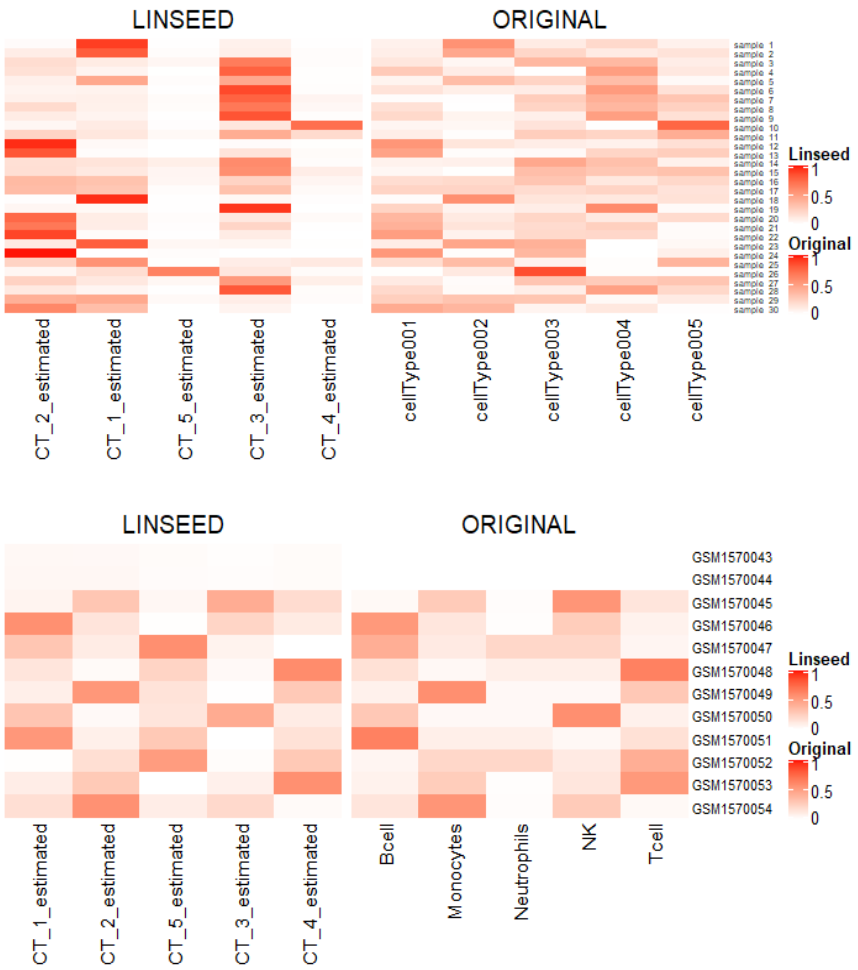


Gráfico 8. Heatmaps linseed
matriz P datos simulados (superior) y GSE64385 (inferior)

Tabla 5. RMSE por tipo celular en cada conjunto de datos en *linseed*

SIMULACIÓN				
CT1	**CT2**	**CT3**	**CT4**	**CT5**
0.1901508	0.1471919	0.2076991	0.1999178	0.1451368
GSE64385				
NK	**Bcell**	**Neutrophils**	**Tcell**	**Monocytes**
0.07993337	0.05482646	0.16053542	0.05067541	0.02514752

En los primeros datos, todos los errores son similares, sin destacar ninguno con mucha diferencia sobre los demás, aunque como ya se mencionó antes, el tipo celular 3 es el que se estima con un mayor error. En los segundos, los neutrófilos son los que superan en medida de error al resto de tipos celulares. En este método, cabe destacar que, al contrario que en los anteriores, el error aumenta de forma considerable en la simulación y disminuye en el caso real (aun habiendo incluido los errores de las primeras muestras).

DECONICA

Es el último de los cinco métodos elegidos para comparar en este trabajo y que, al igual que *linseed*, no requiere de una matriz de firmas, ya que resuelve el problema de la deconvolución desde un enfoque no supervisado.

En la ejecución de su algoritmo, es conveniente indicar si se conoce, el número de tipos celulares en los que se ha de descomponer la muestra, puesto que si no, se impondrá por defecto 100 si $m > 100$ o m si $100 \geq m$, recordando que m hacía referencia al número de muestras que conformaban la matriz de mezclas: 30 y 12 en los datos empleados para realizar el análisis. Las componentes independientes que infiere corresponden, como en *linseed*, a un tipo celular determinado y, al no contar con ninguna referencia, se establece la asociación según la correlación obtenida.

En la representación de los *heatmaps* (*Gráfico 9*), lo primero que se observa es que los tipos celulares que poseen un mayor porcentaje en la matriz original, también lo presentan en las estimadas, pero con un valor inferior, dando lugar a una menor diferencia entre la abundancia de un tipo celular y otro.

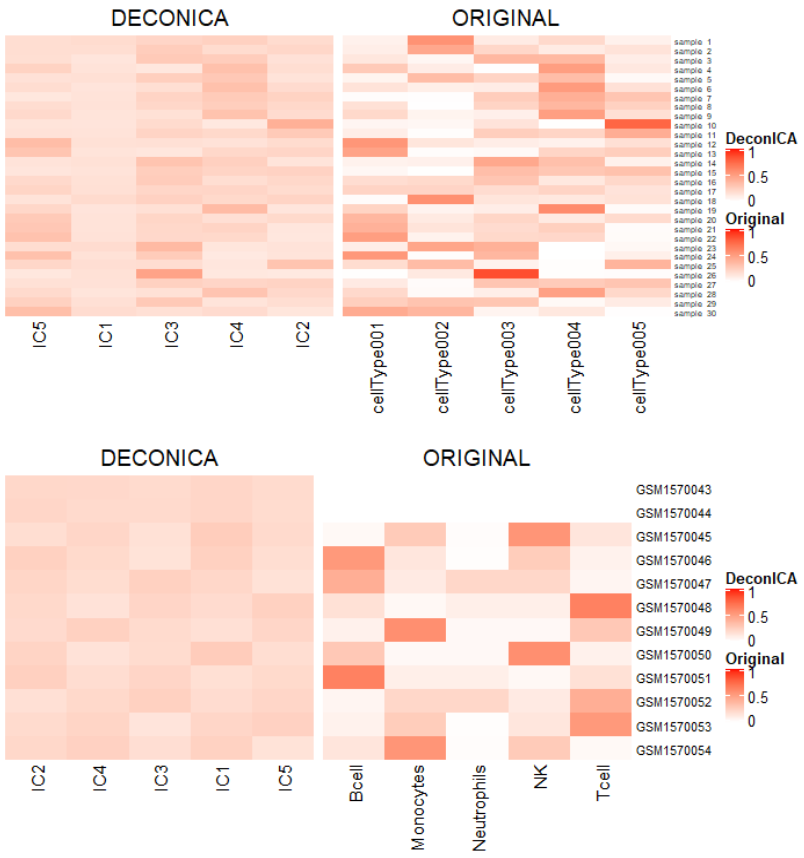


Gráfico 9. Heatmaps deconICA
matriz P datos simulados (superior) y GSE64385 (inferior)

En los datos simulados, el tipo celular 2 se estima con una proporción casi idéntica en todas las muestras, lo cual no ocurre en los datos originales. Por otro lado, los porcentajes estimados para el GSE64385, parecen seguir el mismo patrón en ambas matrices (teniendo en cuenta la diferencia relativa a la disminución de la abundancia) excepto en las dos primeras muestras que, al indicar al principio el número de componentes en los que se tienen que separar las muestras, influye en un aumento del error en las mismas.

Aunque *deconICA* calcula puntuaciones para cada tipo celular en las muestras, éstas se transforman en proporciones de acuerdo con la condición de sumar uno. Si no se realizase esa modificación en las puntuaciones, el patrón de colores que define el *heatmap* del método, se aproximaría más al original, incluyendo las dos primeras muestras en el caso de los datos biológicos no simulados (*Anexo, Gráficos S17 y S18, Heatmap deconICA puntuaciones*).

Las diferencias entre los *heatmaps* mostradas en el *Gráfico 9* se reflejan también numéricamente en el cálculo del error:

Tabla 6. RMSE por tipo celular en cada conjunto de datos en *deconICA*

SIMULACIÓN				
CT1	**CT2**	**CT3**	**CT4**	**CT5**
0.09856240	0.15979651	0.10003665	0.10987384	0.09988678
GSE64385				
NK	**Bcell**	**Neutrophils**	**Tcell**	**Monocytes**
0.1516094	0.1861384	0.1289641	0.1834599	0.1586312

Como ya se visualizó en los gráficos, el tipo celular 2 en los datos simulados es el que se estima con un mayor error (16% aproximadamente) mientras que, en el resto, el error se acerca al 10%. En el conjunto de datos real (excluyendo las dos primeras muestras), el error aumenta en comparación al caso anterior. Aquí, las células peor estimadas pertenecen a un mismo linaje y son los linfocitos B y T, superando un error de un 18%.

COMPARACIÓN

Tras analizar los métodos de forma individual, se detectan algunas semejanzas y diferencias entre ellos que pueden ser comparadas también desde una perspectiva estadística. En esta sección se emplean otras medidas como la correlación de Pearson y la divergencia de Kullback-Leibler, junto con otra ya utilizada, el RMSE para contrastar si, en estas dos últimas, las diferencias observadas en los métodos son significativas o no.

Correlación de Pearson

Este coeficiente representa una medida de gran utilidad, especialmente en los métodos de deconvolución completa, pues gracias a los resultados que proporciona, se pueden identificar los grupos o componentes inferidos con el tipo celular con el que presenten una mayor correlación. Los valores hallados varían de un conjunto de datos a otro, por eso, a continuación, se expone un resumen de lo obtenido en cada uno de ellos:

- Simulación: en este caso, la mayoría de los métodos alcanzan correlaciones de más de un 90% (*Gráfico 10*), siendo muy próximas a uno, incluso en los métodos de deconvolución completa, donde el que obtiene una correlación menor en comparación con los demás es *linseed*. También, llama la atención el caso de *MIND* que presenta un valor de 1 en todos los tipos celulares, pero al igual que en el resto, siguen existiendo algunas correlaciones algo elevadas (alrededor del 40% en valor absoluto) con otros tipos lares distintos al que corresponde.

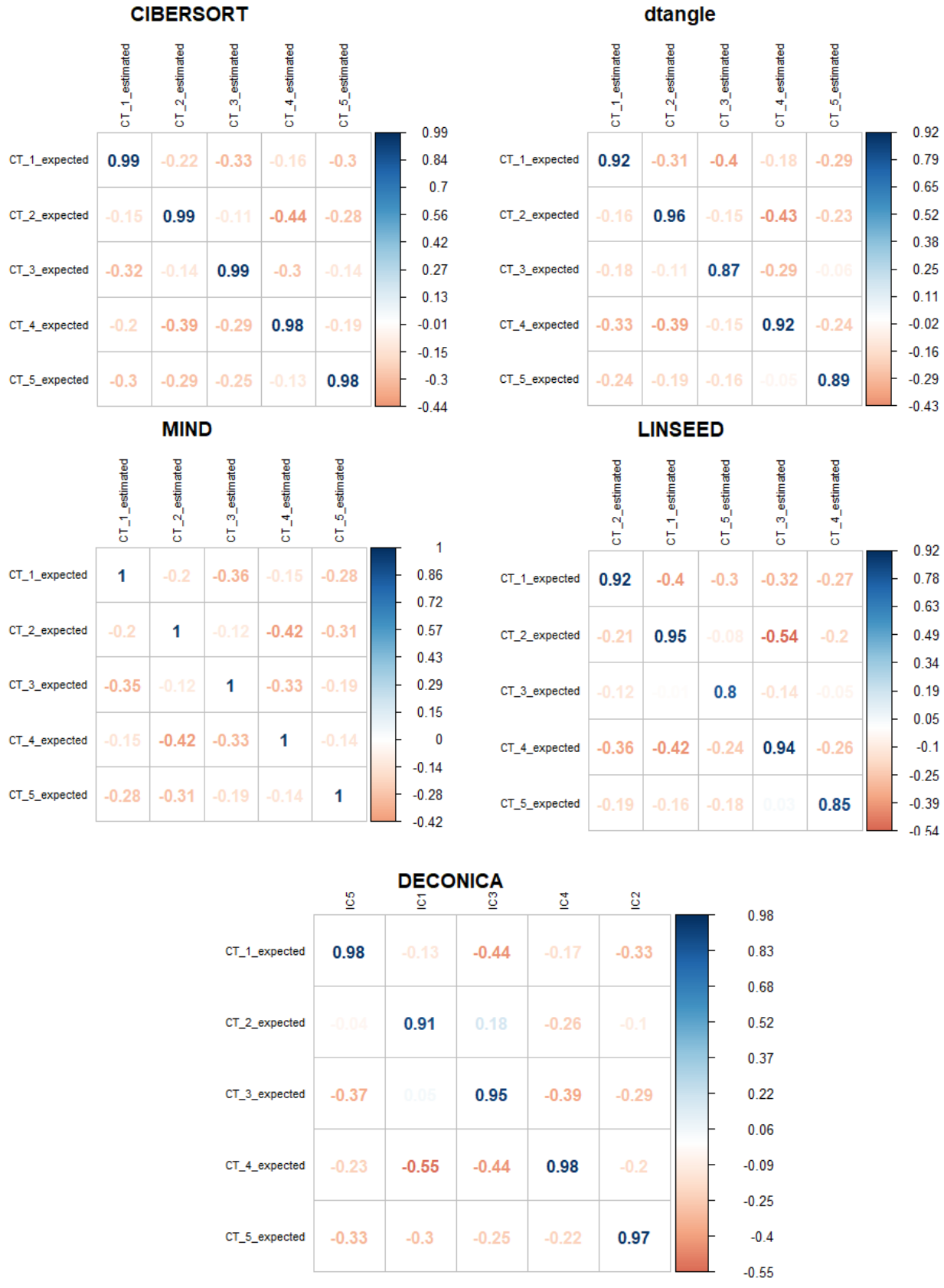


Gráfico 10. Correlación datos simulados

- GSE64385: en los datos reales, se puede ver como generalmente, las correlaciones empeoran y comienza a haber más diferencias entre métodos (*Gráfico 11*):

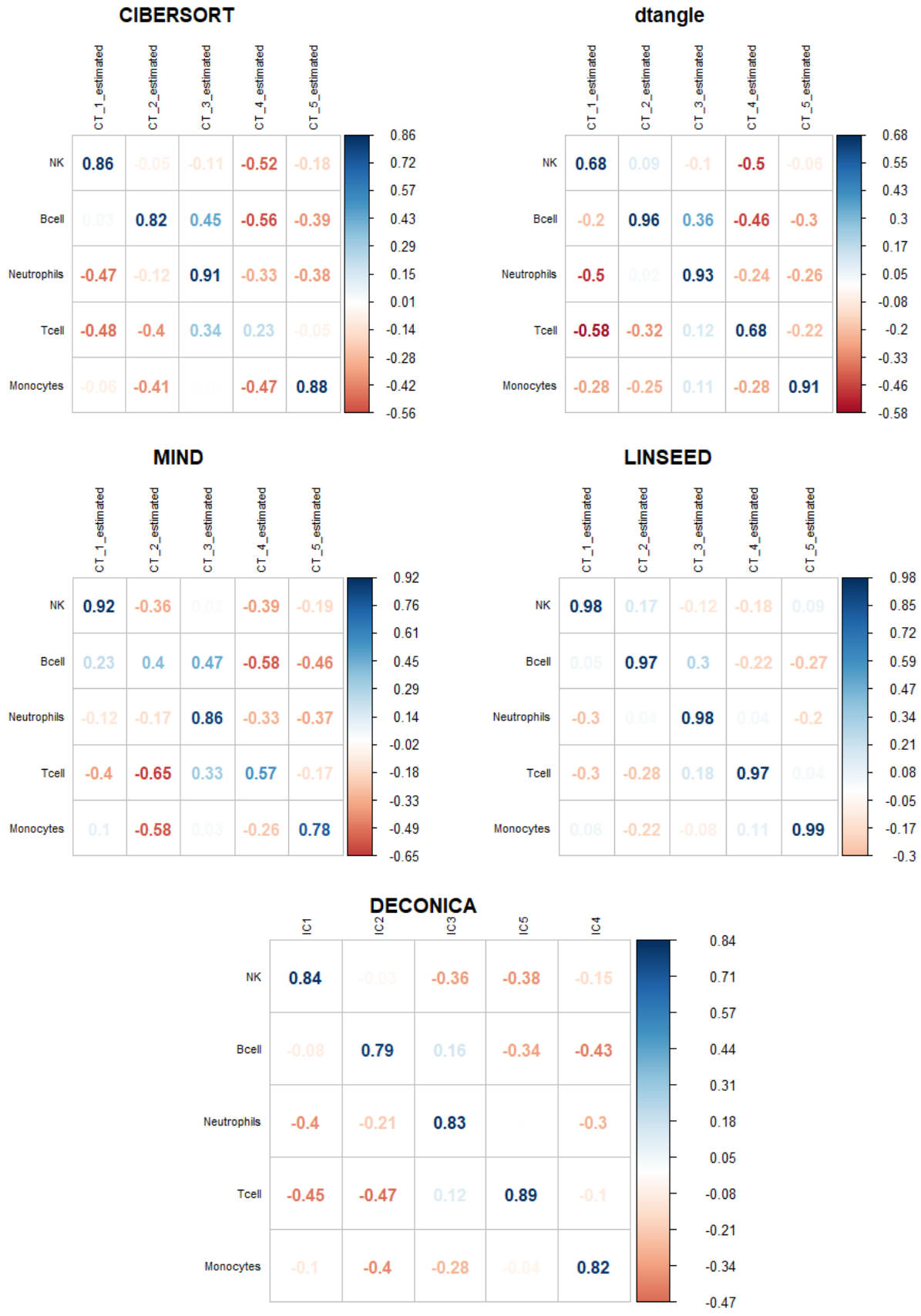


Gráfico 11. Correlación GSE64385

CIBERSORT posee correlaciones cercanas a 1 en todos los tipos celulares (aunque con cierta confusión respecto a otros con correlaciones próximas al 50%) salvo en los linfocitos T, para los que no encuentra ninguna correlación al tipo celular 4, que es el que corresponde a dicho grupo. Por otra parte, *dtangle* sí establece una correlación con todos los tipos celulares que estima, sin embargo, en las células NK y en los linfocitos T, los valores bajan hasta un 69%, observando también correlaciones de un 50% con otros tipos celulares que no corresponden (como ocurre en *CIBERSORT*). En *MIND*, no se halla ninguna asociación clara a los linfocitos T, ni tampoco a los linfocitos B, que tienen una correlación similar con todos los tipos celulares estimados.
En los métodos de deconvolución completa, sobresale *linseed* con correlaciones superiores al 97% en todas las células inmunes y sin ninguna relación con otro tipo celular y, en *deconICA*, los resultados se sitúan aproximadamente en un 80%, pero como ya ocurría en los otros, existe cierta confusión con aquellos tipos en los que alcanza valores de un 40% en los tipos estimados que no corresponden.

Como ya se observó en la simulación, en estos dos últimos métodos, el orden de los tipos celulares o componentes estimados se fija según el tipo celular con el que presente mayor asociación.

A las diferencias observadas entre los métodos en ambos *datasets*, habría que añadir las que se pueden dar dentro de un mismo método, es decir, en *dtangle*, por ejemplo, cabe la posibilidad de emplear la mediana en lugar de la media como método de agregación de las expresiones génicas y, aunque no se obtengan resultados muy dispares, en los datos reales se aprecia una leve mejora en la correlación de algunos tipos celulares (*Anexo, Gráfico S16, Gráfico correlación mediana dtangle).* Lo mismo ocurre con *deconICA* que, si se mantienen las puntuaciones originales, también aumenta dicha medida para casi todos los tipos celulares en el GSE64385 (*Anexo, Gráfico S19, Correlación puntuaciones deconICA (GSE64385)*). No obstante, como estas diferencias son mínimas, en el resto de las comparaciones se continúa con las medidas escogidas que son las más apropiadas y recomendadas por los propios métodos.

Error cuadrático medio (RMSE)

En el cálculo del error por tipo celular en los datos reales, se omitieron las dos primeras muestras para evitar valores extremos, por este mismo motivo, también se eliminarán en el error por muestra, ya que, en la mayoría de los métodos, los errores máximos se obtendrían en ellas.

Teniendo en cuenta dicho criterio, se calcula para cada conjunto de datos, el error que cometen los métodos en cada muestra, sin observar ninguna coincidencia en ellos en cuanto a la muestra donde se obtiene el mayor valor (salvo *CIBERSORT* y *dtangle* en el GSE64385) (*Anexo, Tabla S3, Muestras máximo RMSE*).

Entre los métodos, tomando como medida de referencia la mediana se muestran algunas diferencias que pueden ser contrastadas con las pruebas no paramétricas de Friedman (ANOVA no paramétrico) y Wilcoxon (post-hoc no paramétrico):

Tabla 7. Mediana del RMSE por muestras en cada método y conjunto de datos

DATASET	***CIBERSORT***	***dtangle***	***MIND***	***linseed***	***deconICA***
SIMULACIÓN	0.027	0.2047	0.0013	0.1832	0.1009
GSE64385	0.1587	0.1742	0.2001	0.0657	0.1619

Antes de realizar estos contrastes, es importante señalar que, en la simulación, *MIND* es el método con el menor error, mientras que, en el GSE64385, su valor asciende hasta alcanzar la máxima cifra dentro de los métodos comparados. Por el contrario, en otros métodos no se observa un cambio tan acusado, muestra de ello es *CIBERSORT*, que siempre se mantiene en una posición intermedia sin considerarse el método que mejor o peor estima en estos *datasets*.

Para poder determinar si verdaderamente existe una diferencia significativa en esta medida, se emplea el test de Friedman como método ANOVA no paramétrico para contrastar si existen globalmente diferencias entre los métodos: en él, se obtiene un $p - valor < 2.2e^{-16}$en los datos simulados y, un $p - valor = 0.02441$ en los datos reales. Ambos resultados son menores que 0.05 (nivel de significación escogido para establecer el contraste) por lo que se puede inferir que estas diferencias observadas entre los métodos sí son significativas.

A continuación, para descubrir entre qué métodos se dan estas diferencias, se calcula el test de Wilcoxon de muestras apareadas dos a dos, que presenta los siguientes resultados:

Tabla 8. P-valores obtenidos en el Test de Wilcoxon para el RMSE por muestras

SIMULACIÓN				
	cibersort	**deconica**	**dtangle**	**linseed**
deconica	1.9e-08	-	-	-
dtangle	1.9e-08	0.00042	-	-
linseed	1.9e-08	6.2e-05	0.73034	-
mind	1.9e-08	1.9e-08	1.9e-08	1.9e-08
GSE64385				
	cibersort	**deconica**	**dtangle**	**linseed**
deconica	1.000	-	-	-
dtangle	1.000	1.000	-	-
linseed	0.020	0.088	0.156	-
mind	1.000	1.000	0.633	0.156

En el primer caso, se encuentran diferencias entre todos los métodos (*p-valor < 0.05*), por eso, se pueden ordenar de menor a mayor error, observando las medianas, de la siguiente manera: *MIND* < *CIBERSORT* < *deconICA* < *linseed* < *dtangle* siendo *MIND* el método que mejor estima en cuanto al RMSE.

En el segundo, solo hay diferencias significativas entre *CIBERSORT* y *linseed*, obteniendo con este último un valor para el error más pequeño que con el primero. En ese sentido, únicamente se podría afirmar que, según los valores obtenidos en las medianas, *linseed* < *CIBERSORT*, es decir, *CIBERSORT* posee un mayor error que *linseed* en el GSE64385.

Divergencia de Kullback-Leibler

Las proporciones relativas a los tipos celulares en las muestras, se pueden interpretar también como probabilidades de encontrar dichas células en las muestras, por eso, la divergencia de Kullback-Leibler es un buen indicador que refleja la similitud entre ambas distribuciones de probabilidad, la de las estimaciones y la verdadera.

Tabla 9. Summary divergencia de Kullback-Leibler
para cada método en los datos simulados (fila superior) y el GSE64385 (fila inferior)

	MIN	***Q1***	***MEDIANA***	***MEDIA***	***Q3***	***MAX***
CIBERSORT	0.0006	0.0145	0.0306	0.0462	0.0493	0.2398
	0.2957	0.4962	0.6330	2.9954	0.8881	15.0475
dtangle	0.0880	0.2717	0.5181	0.5005	0.7540	0.9008
	0.0742	0.5041	0.6830	2.90439	0.9352	14.5098
MIND	0.00001	0.00004	0.0001	0.00026	0.0002	0.0027
	0.1310	0.5478	0.7680	2.9662	1.0416	14.4391
LINSEED	0.1560	0.4413	0.5281	0.5616	0.6600	1.0452
	0.02032	0.1176	0.1605	2.5570	0.4176	14.4147
DECONICA	0.0246	0.1977	0.3319	0.4188	0.5734	1.2715
	0.2171	0.5642	0.6133	2.8404	0.7030	14.2899

El rango de valores de la divergencia en ambos *datasets* es muy distinto: en la simulación (filas superiores de cada método), las máximas divergencias son las obtenidas con los métodos de deconvolución completa *deconICA* (1.2715) y *linseed* (1.0452), ambos inferiores a 2, mientras que, en el GSE64385, los mayores valores se encuentran en dos métodos de deconvolución parcial: *CIBERSORT* (15.0475) y *dtangle* (14.5098) alcanzando una divergencia superior a 14.

Otros valores a comentar son los que presentan las medias y las medianas, distanciándose algo más en los datos reales que en los simulados, por lo que, dado el amplio rango de valores reflejado en dicha distancia, conviene estudiar la mediana en lugar de la media para evitar la influencia de *outliers* y valores extremos. Además, si se analizan los valores obtenidos en el tercer cuartil (Q_3) (especialmente en los datos reales que son los que presentan una mayor diferencia en cuanto al valor máximo) son todos inferiores a 1 (o muy cercano a 1 como *MIND*), lo que significa que el 75% de las muestras en todos los métodos posee una distribución bastante similar a la observada.

En referencia a estas medidas, cabe destacar el método *linseed* que, al contrario que el resto, reduce su divergencia al deconvolucionar las muestras biológicas (su mediana disminuye de 0.5281 en la simulación a una de 0.1605 en el GSE64385). Estas diferencias mostradas en la *Tabla 9* se pueden apreciar también al representarlas gráficamente para cada conjunto de datos y así, contrastar posteriormente, la significación estadística que poseen (*Gráfico 12****¡Error! No se encuentra el origen de la referencia.***):

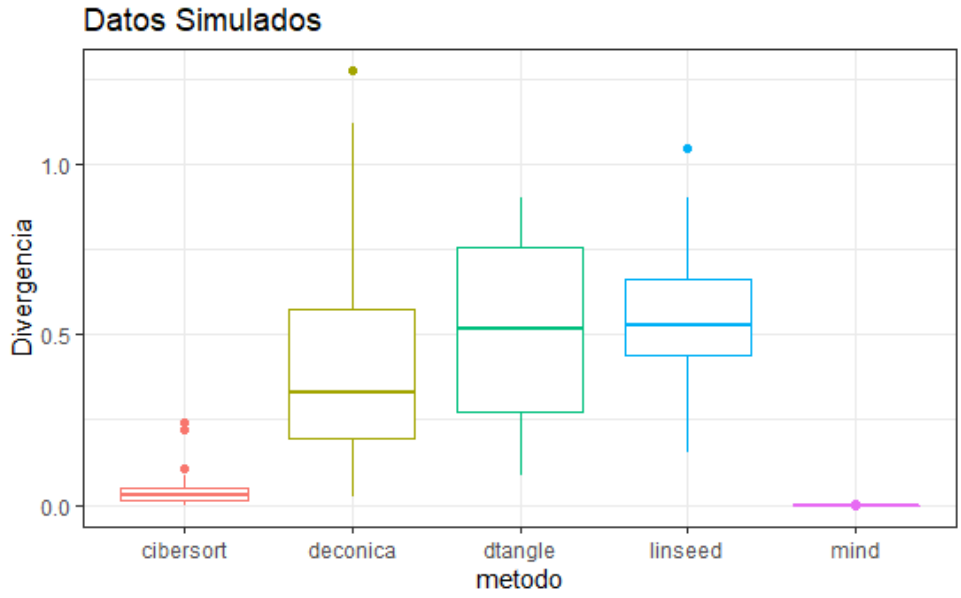


Gráfico 12. Boxplot de las distribuciones de Kullback-Leibler (Simulación)

En el diagrama de cajas (referente a los datos de la simulación), se puede ver la dispersión que presentan los valores de divergencia en los métodos, observando que, tanto la dispersión como las medianas de *CIBERSORT* y *MIND* se encuentran muy por debajo en comparación con las de los demás métodos.

Para comprobarlo estadísticamente, se recurre al análisis con las pruebas de Friedman y Wilcoxon, obteniendo resultados que indican la existencia de diferencias entre ambos métodos y con el resto: $p - valor\ Friedman < 2.2e^{-16}$, $p - valor\ Wilcoxon < 1.9e^{-8}$ (para *CIBERSORT* y *MIND* con los otros cuatro métodos, cada uno). Por tanto, en la simulación, al hallar diferencias significativas, estos dos métodos son los que más se asemejan a la distribución de probabilidad real, situándose *MIND* por encima de *CIBERSORT*.

Por último, falta la evaluación de los resultados en las muestras biológicas en las que, al igual que se hizo con las simuladas, se estudia esta medida gráficamente (*Gráfico 13*):

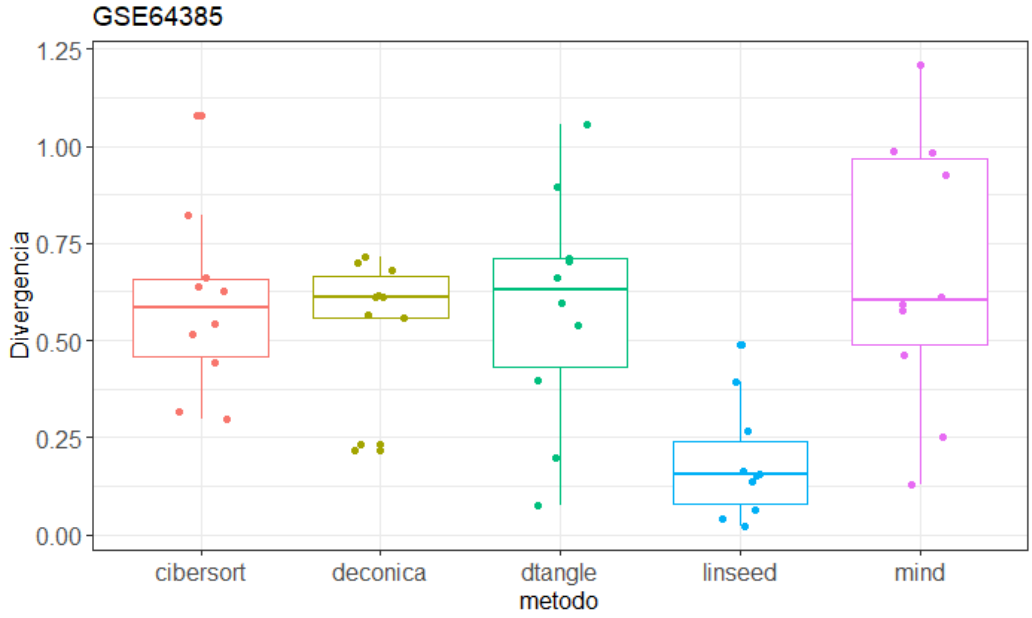


Gráfico 13. Boxplot con puntos de las distribuciones de Kullback-Leibler (GSE64385)

El número de muestras en las que se analiza la divergencia es un total de diez (excluyendo las dos primeras del tipo celular HCT116) por eso, gracias al reducido tamaño de muestra, se elige para representar los valores, un gráfico de cajas con puntos para observar la dispersión. En él, se aprecia que *linseed* es el método que ha obtenido una menor divergencia y dispersión. En cuanto a los demás, el rango de valores es más amplio, especialmente en *MIND* que como se introdujo en la descripción de la *Tabla 9*, es el que alcanza el valor máximo. En *CIBERSORT* y *dtangle*, el comportamiento es similar, aunque con puntos más repartidos a lo largo de todo su recorrido en este último. En *deconICA*, se observan dos puntos con una menor divergencia y una amplitud del recorrido menor que en los anteriores, sin embargo, la mayoría de los puntos se encuentra en lo que representaría la mediana de los otros métodos.

Estas diferencias pueden ser contrastadas con la prueba de Friedman, dando lugar a un $p - valor = 0.0081$ (<0.05) y, al resultar significativo, se procede con el contraste dos a dos con el test de Wilcoxon, en el cual, el único p-valor significativo es el resultante de la comparación entre *CIBERSORT* y *linseed* ($p - valor = 0.02$). Así, lo que se puede afirmar en este *dataset* es que *linseed* se caracteriza por una menor divergencia que *CIBERSORT* (mismo caso que en el error).

5. DISCUSIÓN Y CONCLUSIONES

La deconvolución es un problema fundamental que, enfocada al análisis transcriptómico, permite identificar los tipos celulares presentes en una o varias muestras descomponiendo la heterogeneidad que las caracteriza. En este trabajo, se han reproducido y aplicado cinco de esos métodos, todos ellos destinados a resolver el problema desde una perspectiva biológica y estimando la matriz de proporciones celulares a través del conocimiento de una matriz de firmas y otra de niveles de expresión génica en la deconvolución parcial o solo con la información de esta última matriz, en el caso de la completa.

Métodos como *xCell* (Aran, Hu y Butte, 2017) que calcula puntuaciones de enriquecimiento (*Enrichment Scores*) o *csSAM* (Shen-orr et al., 2010) que estima la matriz de firmas, no pueden ser comparados con los ya escogidos (*CIBERSORT*, *dtangle*, *MIND*, *linseed* y *deconICA*) puesto que no aportan como salida la matriz de proporciones, pero sí podrían emplearse en estudios posteriores orientados a construir una matriz de referencia que aprovechase los avances que ofrece la tecnología scRNA-seq.

En ese sentido, al analizar los métodos seleccionados, como afirma Vallania et al. (2018) en su propuesta de una matriz base integrada por datos de múltiples plataformas, se considera esencial aportar una matriz de referencia válida a los métodos de deconvolución no completa para minimizar el error y mejorar la correlación de las proporciones estimadas y observadas. Otra perspectiva a tener en cuenta es emplear datos de scRNA-seq (cuando sean accesibles). Con ellos, se puede construir una matriz de firmas más específica en función de las muestras que se quieran deconvolucionar. Métodos como *MuSiC* (X. Wang, Park, Susztak, Zhang y Li, 2019) y *DigDLSorter* (Torroja y Sanchez-Cabo, 2019) recurren a este tipo de datos para incluirlos como referencia en sus algoritmos. Sin embargo, no se ha de olvidar la deconvolución completa, ya que, a causa de la variabilidad presente en las muestras de estudio, sus algoritmos son los únicos que no se encuentran influidos por las firmas de tipos celulares ausenten en la matriz de mezclas.

Retomando la evaluación de los métodos elegidos, es habitual realizar su análisis con medidas tradicionales como el error (RMSE) y la correlación de Pearson (Avila Cobos et al., 2020). Sin embargo, el uso de otras medidas como la divergencia KL y el empleo de técnicas no paramétricas (test de Friedman y Wilcoxon) resultan de gran utilidad en las comparaciones entre ellos, pudiendo determinar estadísticamente si presentan diferencias significativas.

Por último, es importante mencionar la rápida aparición de nuevos métodos que aporten su solución al problema, por eso, actualizar las comparaciones entre ellos es una tarea fundamental que supondrá una gran ayuda a la hora de optar por un método u otro. Así, en lo que a este estudio se refiere y, después de comprobar la precisión de los métodos en muestras simuladas y humanas, se puede concluir que:

- Dentro de los métodos de deconvolución parcial, *CIBERSORT* es de los que presenta un menor error y divergencia en ambos conjuntos de datos (aunque no significativos en el caso real y, realizando una peor estimación que *linseed*). Por otra parte, *MIND*, a pesar de su casi perfecta estimación en los datos simulados (debida a la similitud entre su algoritmo y el empleado para la simulación de muestras), la dispersión mostrada en estas dos medidas con el GSE64385 genera dudas en cuanto a la fiabilidad de la precisión, ya que sus valores varían en función de las distintas muestras. En *dtangle*, los valores son bastante elevados, acercándose a *MIND* en las muestras reales y superando a *CIBERSORT* en las simuladas (convirtiéndose en el peor método). No obstante, para este método se podría esperar una mejora en la estimación si se posee información más específica de los tipos celulares a deconvolucionar con unas muestras puras más representativas, al igual que en los métodos que emplean la matriz de firmas *LM22* deberían ser probados con otra matriz base teniendo en cuenta lo comentado anteriormente sobre la inclusión de datos de varias plataformas.
- La deconvolución completa, planteada en un principio como una dificultad para los métodos asociados, ha supuesto un gran descubrimiento que ha posicionado a *linseed* como el método con menor error y divergencia (junto con una mayor correlación) en el conjunto de datos biológicos detectando, además, la ausencia de los tipos celulares inmunes en las dos primeras muestras. *DeconICA*, también obtiene unas puntuaciones que muestran la falta de esas células, pero al transformarlas en porcentaje para poder comparar sus estimaciones con el resto de los métodos, sus valores empeoran por considerar únicamente la puntuación máxima por filas y no la del total.
- En vista a los resultados obtenidos en los tipos celulares, no se puede hacer una distinción entre linajes, ya que, en ningún método se aprecia una mejora relevante al estimar la rama linfoide o mieloide. Sin embargo, en todos los métodos se observa un error más elevado en los linfocitos B (incluyendo también los T en los métodos *MIND* y *dtangle*) salvo en *linseed*, en el que los neutrófilos son los peor estimados. En cuanto al mejor, existe mayor discrepancia, pudiéndose establecer tres grupos: NK ("*natural killer*") en *CIBERSORT* y *dtangle*, neutrófilos en *MIND* y *deconICA* y, monocitos en *linseed*. En la simulación, también se encuentra cierta coincidencia: el tipo celular 5 corresponde a los tipos celulares mejor estimados en todos los métodos excepto en *CIBERSORT* y, el 4, el que peor en todos, menos en *linseed* y *deconICA*.

Como conclusión final, los métodos de deconvolución estudiados, a pesar de sus limitaciones, presentan una solución bastante próxima a la realidad celular que permite la cuantificación de los tipos celulares sin tener que recurrir a las técnicas experimentales con el coste y tiempo que éstas suponen. Además, el método más adecuado dependerá del conocimiento previo que se disponga (firmas celulares y número de tipos presentes) y del tipo de datos de entrada (microarray, RNA-seq o scRNA-seq).

6. REFERENCIAS

Andrews, T. S. y Hemberg, M. (2018). Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine*, *59*, 114-122. https://doi.org/10.1016/j.mam.2017.07.002

Aran, D., Hu, Z. y Butte, A. J. (2017). xCell : digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology*, *18*(1), 220. https://doi.org/10.1186/s13059-017-1349-1

Avila Cobos, F., Alquicira-Hernandez, J., Powell, J., Mestdagh, P. y Preter, K. De. (2020). Comprehensive benchmarking of computational deconvolution of transcriptomics data. *bioRxiv*. https://doi.org/10.1101/2020.01.10.897116

Avila Cobos, F., Vandesompele, J., Mestdagh, P. y De Preter, K. (2018). Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics (Oxford, England)*, *34*(11), 1969-1979. https://doi.org/10.1093/bioinformatics/bty019

Awad, M. y Khanna, R. (2015). *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Apress. https://doi.org/10.1007/978-1-4302-5990-9

Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., … de Reyniès, A. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology*, *17*(1), 218. https://doi.org/10.1186/s13059-016-1070-5

Bronkhorst, A. W. (2015). The cocktail-party problem revisited : early processing and selection of multi-talker speech. *Attention, perception & psychophysics*, *77*(5), 1465-1487. https://doi.org/10.3758/s13414-015-0882-9

Camarillo-Peñaranda, J. R., Saavedra-Montes, A. J. y Ramos-Paja, C. A. (2013). Recomendaciones para Seleccionar Índices para la Validación de Modelos. *TecnoLógicas*, 109-122. https://doi.org/10.22430/22565337.372

Chen, L. (2019). Mathematical Modeling and Deconvolution for Molecular Characterization of Tissue Heterogeneity Mathematical Modeling and Deconvolution for Molecular Characterization of Tissue Heterogeneity. Arlington, Virginia: Virginia Tech.

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*(5), 975-979.

Cos, M. G. (2010). Nuevos métodos de diagnóstico molecular. *GH CONTINUADA*, *9*(4), 160-164.

Cover, T. M. y Thomas, J. A. (2005). *Elements of Information Theory*. John Wiley & Sons. https://doi.org/10.1002/047174882X

Czerwińska, U. (2018). *Unsupervised deconvolution of bulk omics profiles : methodology and application to characterize the immune landscape in tumors*. Sorbonne Paris Cité.

Definición de ARN - Diccionario de cáncer - National Cancer Institute. (s. f.). Recuperado a partir de https://www.cancer.gov/espanol/publicaciones/diccionario/def/arn

Definición de citometría de flujo - Diccionario de cáncer - National Cancer Institute. (s. f.). Recuperado 2 de abril de 2020, a partir de https://www.cancer.gov/espanol/publicaciones/diccionario/def/citometria-de-flujo

Definición de transcriptómica - Diccionario de cáncer - National Cancer Institute. (s. f.). Recuperado 4 de mayo de 2020, a partir de https://www.cancer.gov/espanol/publicaciones/diccionario/def/transcriptomica

Fierro Correa, J. A. (2001). Breve historia del descubrimiento de la estructura del ADN. *Rev. Méd. Clín. Condes*, *12*, 71-75.

Gaujoux, R. (2013). An introduction to gene expression deconvolution and the CellMix package, 1-45.

Gómez López, H. (1984). La distribución binomial negativa en el estudio de poblaciones de insectos. *Revista Facultad Nacional de Agronomía Medellín*, *37*(2), 3-11.

Hicks, S. C., Teng, M. y Irizarry, R. A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv*. https://doi.org/10.1101/025528

Himberg, J. y Hyvärinen, A. (2003). Icasso: software for investigating the reliability of ICA estimates by clustering and visualization. *In Neural Networks for Signal Processing - Proceedings of the IEEE Workshop*, 259-268. https://doi.org/10.1109/NNSP.2003.1318025

Home | HUGO Gene Nomenclature Committee. (s. f.). Recuperado 20 de mayo de 2020, a partir de https://www.genenames.org/

Hornillo Mellado, S. (2005). *Sobre el Análisis en Componentes Independientes de Imágenes Naturales*. Universidad de Sevilla. Recuperado a partir de http://hdl.handle.net/11441/16056

Hunt, G. J., Freytag, S., Bahlo, M. y Gagnon-Bartsch, J. A. (2019). Dtangle: Accurate and robust cell type deconvolution. *Bioinformatics (Oxford, England)*, *35*(12), 2093-2099. https://doi.org/10.1093/bioinformatics/bty926

Hwang, B., Lee, J. H. y Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine*, *50*(8), 96. https://doi.org/10.1038/s12276-018-0071-8

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626-634. https://doi.org/10.1109/72.761722

Kairov, U., Cantini, L., Greco, A., Molkenov, A., Czerwinska, U., Barillot, E. y Zinovyev, A. (2017). Determining the optimal number of independent components for reproducible transcriptomic data analysis. *BMC Genomics*, *18*, 712. https://doi.org/10.1186/s12864-017-4112-9

Lewis, C. (2008). Linear Programming: Theory and Applications. *Whitman College Mathematics Department*.

Lleonart, M. E., Sánchez, R., Martín-Duque, P. y Ramón y Cajal, S. (1997). Técnicas de hibridación, clonación y secuenciación de ácidos nucleicos en el diagnóstico anatomopatológico. *Revista Española de Patología*, *30*(3), 249-257.

Mayani, H., Flores-Figueroa, E., Pelayo, R., Montesinos, J. J., Flores-Guzmán, P. y Chávez-González, A. (2007). Hematopoyesis. *Cancerología*, *2*, 95-107.

Mcnally, J. G., Karpova, T., Cooper, J. y Conchello, J.-A. (1999). Three-Dimensional Imaging by Deconvolution Microscopy. *Journal of Methods*. https://doi.org/10.1006/meth.1999.0873

Menhenhall, W., Beaver, R. y Beaver, B. (2016). *Introducción a la probabilidad y estadística*. *Cengage Learning* (13.ª ed.). https://doi.org/10.1177/1533210110392951

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., … Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, *12*(5), 1-10. https://doi.org/10.1038/nmeth.3337

Orihuela, N. (2015). Deconvolución de Euler de datos gravimétricos del segmento central de la zona de borde sur de la Placa Caribe. *Boletín de Geología*, *37*(2), 25-39.

Patiño, P. J. y Robinson Ramírez Pineda, J. (2006). El dogma central de la biología molecular.

Proserpio, V. y Lönnberg, T. (2016). Single-cell technologies are revolutionizing the approach to rare cells, 225-229. https://doi.org/10.1038/icb.2015.106

Rivas-Lopez, M. J., Sánchez-Santos, J. M. y De las Rivas, J. (2005). Estructura y análisis de microarrays. *Boletín de Estadística e Investigación Operativa*, *21*(2), 10-15.

Rodríguez Cubillos, A., Perlaza Jiménez, L. y Bernal Giraldo, A. (2014). RNA-Seq Data Analysis in Prokaryotes: A Review for Non-experts. *Acta Biológica Colombiana*, *19*(2), 131-142. https://doi.org/10.15446/abc.v19n2.41010

Shen-orr, S. S., Tibshirani, R., Khatri, P., Bodian, D. L., Staedtler, F., Perry, N. M., … Butte, A. J. (2010). Cell type – specific gene expression differences in complex tissues. *Nature Methods*, *7*(4). https://doi.org/10.1038/nmeth.1439

Torroja, C. y Sanchez-Cabo, F. (2019). Digitaldlsorter: Deep-learning on scrna-seq to Deconvolute Gene Expression Data. *Frontiers in Genetics*, *10*, 978. https://doi.org/10.3389/fgene.2019.00978

Vallania, F., Tam, A., Lofgren, S., Schaffert, S., Azad, T. D., Bongen, E., … Khatri, P. (2018). Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nature Communications*, *9*(1), 4735. https://doi.org/10.1038/s41467-018-07242-6

Venet, D., Pecasse, F., Maenhaut, C. y Bersini, H. (2001). Separation of samples into their constituents using gene expression data. *Bioinformatics (Oxford, England)*, *17*, 279-287.

Vincent, E., Bertin, N., Gribonval, R., Bimbot, F., Vincent, E., Bertin, N., … Bimbot, F. (2014). From blind to guided audio source separation : How models and side information can improve the separation of sound. *IEEE Signal Processing Magazine*, *31*(3), 107-115.

Wang, J., Devlin, B. y Roeder, K. (2020). Using multiple measurements of tissue to estimate subject- and cell-type-specific gene expression. *Bioinformatics (Oxford, England)*, *36*(3), 782-788. https://doi.org/10.1093/bioinformatics/btz619

Wang, X., Park, J., Susztak, K., Zhang, N. R. y Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, *10*, 380. https://doi.org/10.1038/s41467-018-08023-x

Wang, Z., Gerstein, M. y Snyder, M. (2009). RNA-Seq : a revolutionary tool for transcriptomics. *Nature reviews.Genetics*, *10*(1), 57-63. https://doi.org/10.1038/nrg2484

Zaitsev, K., Bambouskova, M., Swain, A. y Artyomov, M. N. (2019). Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nature Communications*, *10*, 2209. https://doi.org/10.1038/s41467-019-09990-5

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. y Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PloS one*, *9*(1). https://doi.org/10.1371/journal.pone.0078644

SUMMARY

Over the years, transcriptomics have had a fundamental role in genome understanding because thanks to its studies in gene expression levels, it has allowed to determine differentiated expression profiles discovering molecular components that help in the comprehension of some diseases development like cancer.

Transcriptomic data analysis is carried by several techniques including microarrays, RNA-seq and scRNA-seq. The former is supported by a hybridization process and the others in a sequencing process.

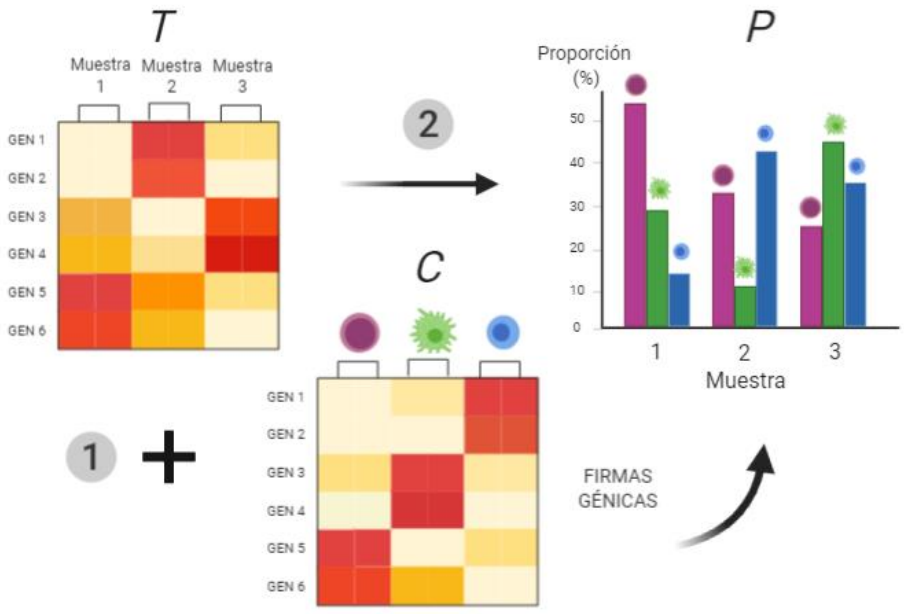
Despite their popularity and effectiveness, all these techniques have limitations. Depending on previous knowledge and the aim of the study, the best technique is chosen providing that it has been adapted to research conditions. For example, it is known that scRNA-seq technology is the most suitable so that it studies cellular heterogeneity. However, it has some disadvantages associated with cost and time and therefore, it is necessary to find a method that investigates such problem using data from the two previous techniques.

To meet the challenge, deconvolution methods, also used in fields such as seismology and signal pre-processing, suggest some algorithms that analyse gene expression in numerous samples struggling to identify the different cell types mixed in them. The problem could be solved from two formulations: on the one hand, partial deconvolution has the knowledge of a signature gene matrix C that provides useful information to deconvolute mixed matrix T (which contents expression levels in each sample) in a proportion matrix P. On the other hand, in complete deconvolution, matrices C and P are inferred directly by the input matrix T.

Regardless of which formulation is taken, deconvolution equation formula is:

$$T = C \times P$$

where T $(n \times m)$ = measured expression values of n genes in m samples and it is decomposed into product of C = signature matrix of n genes in the k cell types and P = mixing proportions of k cell types in m samples.



In this project, algorithms of five deconvolution methods are assessed (three from a partial perspective and two from a complete one) through a simulated and a real dataset, thereby prove whether there are any differences between cell types estimations. Methods and their algorithms are defined bellow:

CIBERSORT

Partial deconvolution method that provides a signature matrix called LM22 to deconvolve immune cell types. It has a web application that, in case of other cell type's estimations, let users yield other signature matrix or introduce their own reference matrix. Furthermore, as with subsequent methods, there is an R script that runs its v-SVR algorithm.

Support Vector Regression (SVR) is a supervised algorithm regarded as a particular type of Support Vector Machine (SVM). It estimates its predictions using a separation hyperplane so that the nearest points to its margins are the support vectors (v). These ones represent the marker genes of the signature matrix in cellular deconvolution.

Related to data pre-processing, this method has some special requirements: gene symbols must be in the first column of T and sample names on its header, similarly, signature matrix C is organised with gene by rows and cell types by columns.

LM22

LM22 signature matrix had previously been cited as the proposed basis matrix of *CIBERSORT* for immune cell type deconvolution. This is formed by 22 cell types expressions organized according with their linage are:

- Lymphoid lineage:
  - B lymphocyte: naïve and memory B cells
  - T lymphocyte: CD4 (*naïve*, memory activated and resting), follicular helper, regulatory and gamma-delta T cells.
  - Natural killer cell (activated and resting)
- Myeloid lineage:
  - Mast cells (activated and resting)
  - Myeloblasts: neutrophils, eosinophils and monocytes (dendritic cells and macrophages)

This project uses this basis matrix as reference in partial deconvolution methods when it is being analysed a real dataset. The reason for this is the challenge to build a matrix of these properties with a standard procedure and with scRNA-seq data recollected. Because of such drawbacks, *LM22* is taken as reference for all partial deconvolution methods with the exception of one (*dtangle*). This one takes from its repository single cell data to make the algorithm chose the gene markers.

MIND

This deconvolution method hast the main objective to estimate subject- and cell-type-specific gene expression. However, a preliminary step before obtaining such profiles is to calculate cell proportions with NNLS (Non Negative Least Squares) algorithm. This consists in minimize the sum of squares of the differences between predicted and observed values adding non-negativity restriction for the coefficients and in cell type deconvolution, the sum of proportions within each sample must be one.

T and C matrices must have the same number of genes; therefore, it is necessary to filter both matrices with intersection genes before starting the algorithm.

DTANGLE

Last method in this study that solves deconvolution problem using a partial approach. It is based on a linear model in which gene expression in one sample is supposed to be the result of a lineal combination of its expression in cell type k, weighted by the proportion it represents in the sample. Moreover, the most often scales are used: a linear model is applied on linear scale expressions and after that; it is adjusted by a logarithmic transformation of the data.

As it has already mentioned in the explanation of LM22 signature matrix, real dataset deconvolution in *dtangle* is performed by the expression of several pure samples from which marker genes are taken to estimate cell type proportions. Finally, it is possible to specify the type of input data (microarray or RNA-seq) taking into account its source variability.

LINSEED

Method that covers complete deconvolution and consequently, it is not necessary to include a signature matrix with cell type information that helps in the deconvolution process.

Linseed is based on gene mutual linearity, that is, gene expression signal of each cell type is proportional to its relative fraction in samples. Graphically, this can be represented on a linear subspace called simplex whose corners (optimal points) are cell type signatures and proportions.

In order to obtain the proportion matrix, this method follows some steps:

1. Generate a linseed object with T matrix.
2. Build a gene collinearity network.
3. Determinate number of cell types.
4. Project the data to the simplex and deconvolve the dataset.
5. Visualize estimated proportions.

DECONICA

Method based on *FastICA* algorithm to perform an Independent Component Analysis (ICA) and solve the complete deconvolution problem. ICA is a computational method related to Principal Component Analysis (PCA) that reveals hidden factors in a mixture. It is more powerful than PCA because it looks for independent components instead of correlated ones. Independency is achieved by a contrast function such that a separation matrix is found for the mixing samples vector that maximize their values.

As well as in previous methods, decomposition function (deconvolution) is executed in R. It is convenient to specify the number of cell types we want to infer because by default, this number will be set to 100 or the number of the samples in the mixture, depending on its size.

The result is a score matrix that must be converted into percentages in order to benchmark its estimations with the other methods.

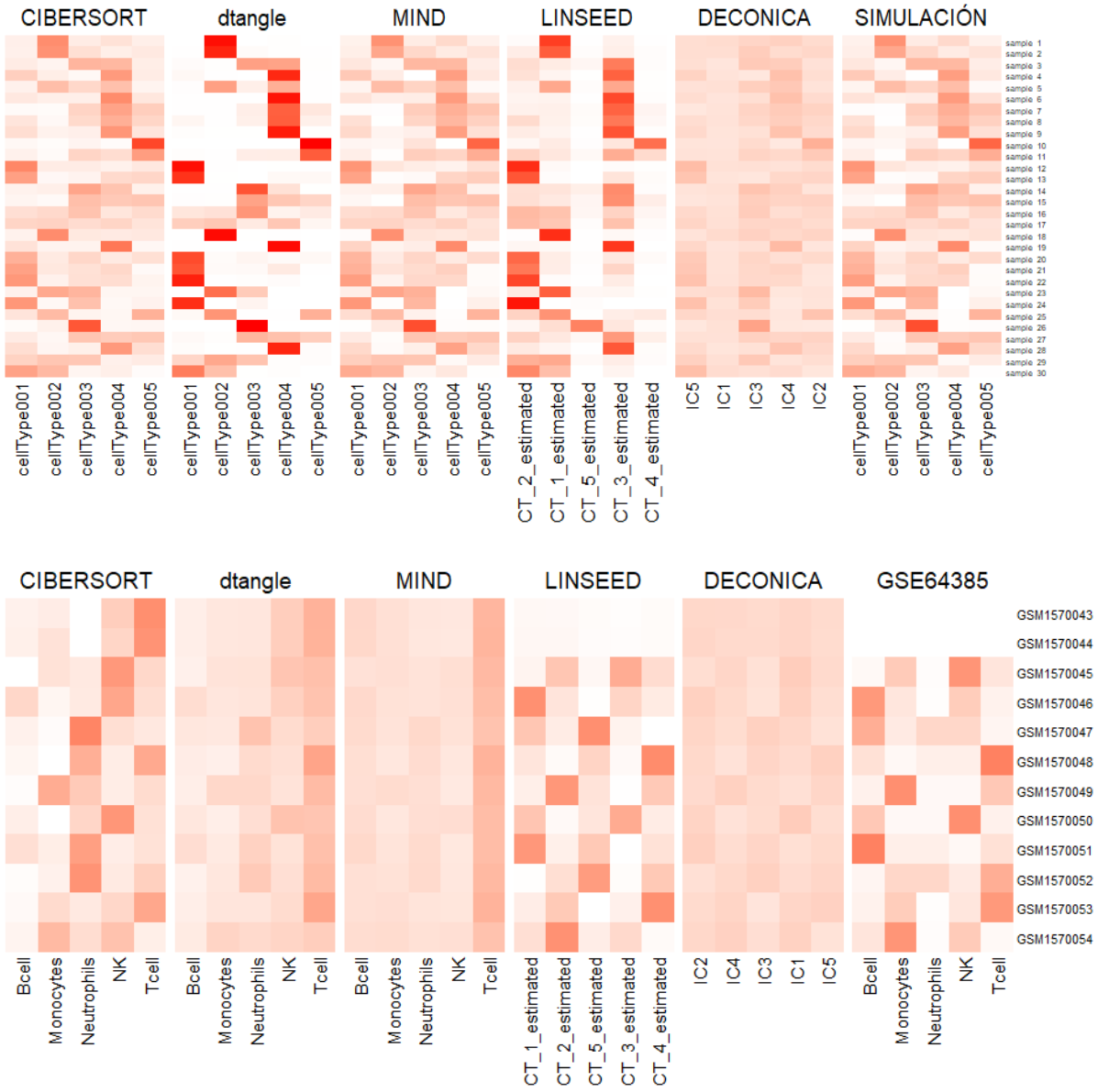
BENCHMARKING

The fundamental purpose of this study is to estimate cell type proportions for the different samples employing five deconvolution methods. In all cases, P matrix is estimated so, a common criterion is established: output matrix is represented in *heatmaps* and error measurements (RMSE), divergence (Kullback-Leibler) and correlation values (Pearson coefficient) are calculated. It is also analysed if there are some significant differences with non-parametric tests of Friedman and Wilcoxon. Besides, methods are tested in two datasets:

- Simulation: a random dataset is generated conformed by the three key matrices throughout a non-negative matrix factorization (NMF) model.
- GSE64385: it is a real dataset that contains the expression of six cell types: HCT116 cancer colon cell line and five immune cell populations (B cells, T cells, NK cells, monocytes and neutrophils). The first cell type is mixed in a constant quantity in the twelve samples except in the two first samples that are pure cells of this particular type. Respect to the comparison, it is only quantified immune cells proportions, but it is taken into account what is the method capable of detect lacking of immune cells in the first samples.

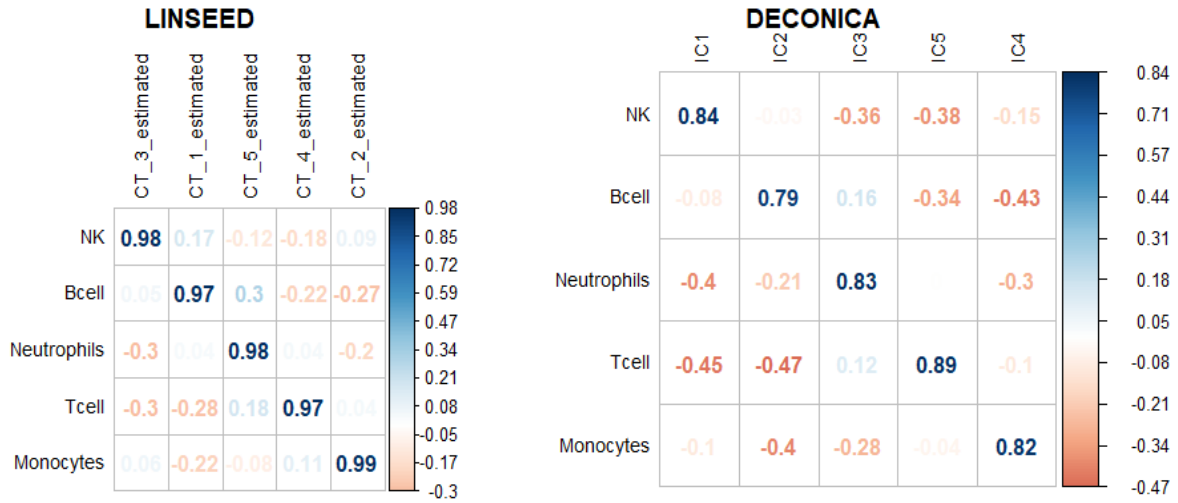
Cell type estimations accuracy varies among datasets. Generally, it is observed a better result in the simulation than in real data (except in linseed). This could be tested with a *heatmap* representation where the worst cell type estimations in each method are shown and which of them is the most convenient in each situation.

In simulated data, *MIND* (partial deconvolution) is the method that represents proportions with the smallest error (and greater correlation) which may be due to the similarities between the simulation algorithm and the one that *MIND* uses in deconvolution process. In real data, *linseed* (complete deconvolution) is the only one that, as sum to one condition is not included, faithfully represents the two first samples. If score matrix had been represented instead of percentages, *deconICA* would have been able to detect the problem with these samples.



As mentioned earlier, *heatmaps* could be used to discover some coincidences in terms of the worst cell types estimated; nevertheless, in the simulation situation, there are not any common cell types. For example, the worst cell types estimated by *linseed* are number three and five and, in *deconICA*, the worst is number 2, in which seems to be mixed by the same proportion in all samples. In GSE64385, B cells are estimated with more difficulty than the others, excluding *linseed* (method with the best *heatmap* representation according to reality) in which neutrophils are the cells with a greater error.

Visualizing *heatmaps*, it'ss clear that the names of the estimated cell types are not the same as they are called in the original dataset. This is because in this type of methods, there is not a signature matrix with cell types information. For this reason, the association between estimated cell types and observed is done by a correlation plot as seen in the graph (obtained with real dataset) shown below:



Although they are no supervised deconvolution methods and it is a biologic sample deconvolution, both methods present high percentages with associated cell types. This is also observed in the other methods except in *CIBERSORT* and *MIND* where some lymphocytes are confused.

After seeing the last graphics, it is fair to say that there are some differences between the methods but in order to know if they are significant, Friedman (non-parametric ANOVA) and Wilcoxon (paired t-test) tests are performed.

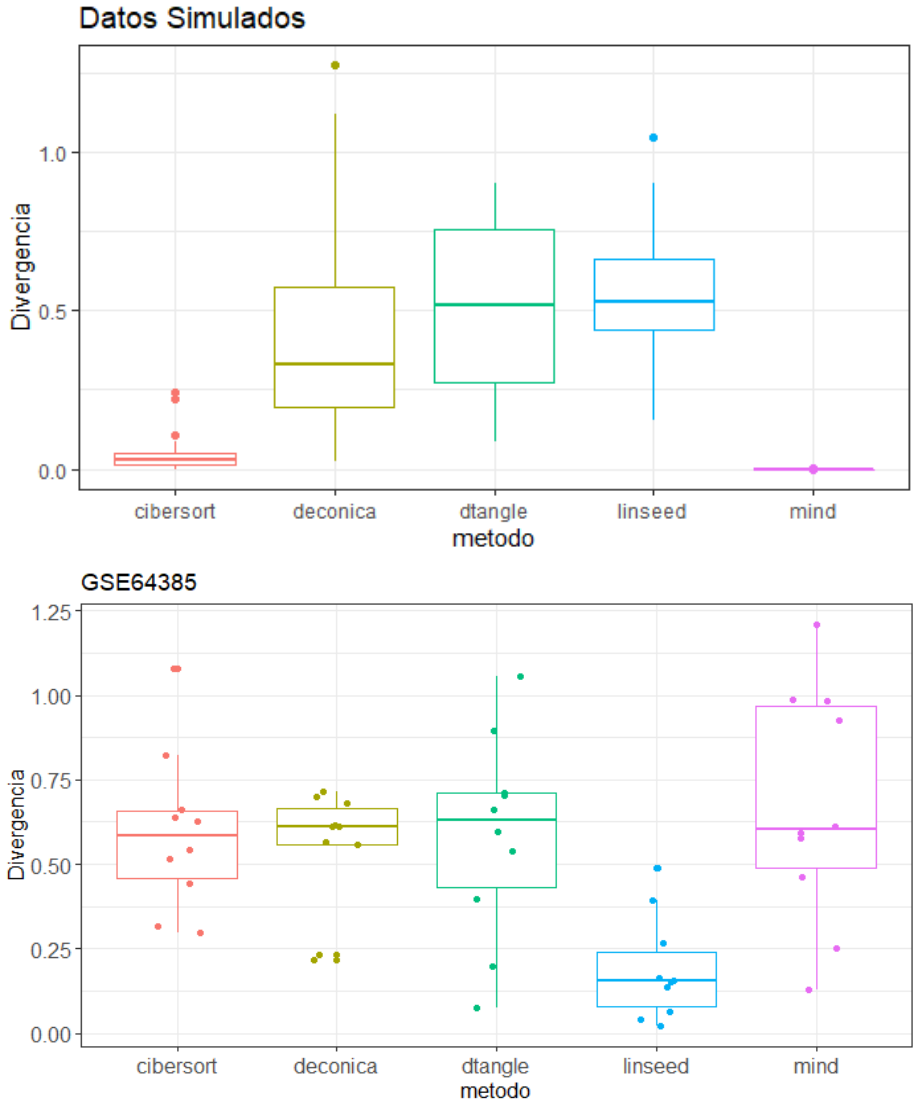
Related to error measurement, Friedman test detects significant differences in both datasets (simulation and GSE64385). In addition, Wilcoxon test leads to significant p-values for all methods on simulated data so that bearing in mind medians values listed in the following table, methods could be sorted from best to worst as: *MIND* < *CIBERSORT* < *deconICA* < *linseed* < *dtangle.*

DATASET	***CIBERSORT***	***dtangle***	***MIND***	***linseed***	***deconICA***
SIMULACIÓN	0.027	0.2047	0.0013	0.1832	0.1009
GSE64385	0.1587	0.1742	0.2001	0.0657	0.1619

In GSE64385, there are just significant differences between *CIBERSORT* and *linseed*, it can therefore be stated that *linseed* < *CIBERSORT* and thus, *linseed* would obtain a better result than *CIBERSORT* in error terms.

The same procedure is repeated for Kullback-Leibler divergence with the difference that this time *CIBERSORT* and *MIND* are the only significant ones respect to the others in the simulation. Observing its boxplot, it can be assumed that these methods have obtained a better performance in comparison with the others, without any differences between them. Moreover, this kind of graphics allows evaluating the degree of dispersion in divergence values looking at the whiskers of the boxes.

Real data reveals the same result as we have seen before with error: *CIBERSORT* and linseed are the methods in which significant differences are observed so that linseed has a superior position to *CIBERSORT* (*linseed* < *CIBERSORT*). According to boxplot, *linseed* median value is the smallest divergence and it has less dispersion, too.



Both in error and divergence, the first two samples are removed in the real data, since in this way, extreme values are avoided that could influence on benchmarking tests.

Ultimately, after studying methods comparison, it concludes that within partial deconvolution approaches, *CIBERSORT* is considered as a proper option for simulated data and, despite it is surpassed by *linseed* in the real dataset, its range of values is less dispersed than the rest no supervised methods. Viewing its drastic change from one dataset to another, *MIND* does not turn out to be reliable to be applied in a biologic sample deconvolution. The reason of this thought is because of the great variability on this type of data. *Linseed* highlights in complete deconvolution due to its minimal error and divergence and high correlation, identifying each estimated cell type with a unique real cell type. *DeconICA* does not present quite high values in terms of error and divergence, but the transformation step of convert scores in percentages might affect in loss of useful information in deconvolution process.

Lastly, no pattern has been found to determine a cell type that is better or worse estimated either a particular linage that could be inferred with more accuracy. Although, answering this question must be assessed with several datasets deconvolution composed by the same cell types. Because of that, using a signature matrix well characterized and obtaining single cell data are essential to continue on surveys on this field.