VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

Trabajo de Fin de Grado

Análisis e implementación de algoritmos de deconvolución de mezclas celulares complejas basados en expresión de genes (firmas génicas) y aplicación a muestras de tumores

Analysis and implementation of complex cell mixture deconvolution algorithms based on gene expression (gene signatures) and application to tumor samples

Autora: **Natalia Alonso Moreda**
Tutores: **Dr. José Manuel Sánchez Santos**
Dr. Javier De Las Rivas Sanz
Alberto Berral González

Facultad de Ciencias
Grado en Estadística

CiC
CENTRO DE INVESTIGACIÓN DEL CÁNCER

Julio, 2021

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

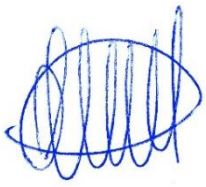
Trabajo de Fin de Grado

Análisis e implementación de algoritmos de deconvolución de mezclas celulares complejas basados en expresión de genes (firmas génicas) y aplicación a muestras de tumores

Analysis and implementation of complex cell mixture deconvolution algorithms based on gene expression (gene signatures) and application to tumor samples

Autora: **Natalia Alonso Moreda**
Tutores: **Dr. José Manuel Sánchez Santos**
Dr. Javier De Las Rivas Sanz
Alberto Berral González

Dr. José Manuel Sánchez Santos | Dr. Javier De Las Rivas Sanz | Alberto Berral González



Natalia Alonso Moreda

ÍNDICE

1. INTRODUCCIÓN

En las últimas décadas, el interés por el análisis del transcriptoma humano ha ido creciendo exponencialmente debido al gran interés científico y biomédico que conlleva conocer su funcionamiento. Permite identificar la actividad de genes específicos de cada tipo celular, así como determinar de qué manera influyen los cambios producidos en dichas células en algunos procesos de desarrollo de los organismos (morfogénesis, embriogénesis y diferenciación celular) o en enfermedades tales como el cáncer (National Human Genome Research Institute Home | NHGRI, 2021).

El estudio de este tipo de datos se basa en el análisis de la expresión de los genes marcadores de diferentes tipos de células, que se obtiene a partir de diversas técnicas diseñadas para ello. Entre las técnicas más utilizadas se encuentra la tradicionalmente denominada tecnología de microarrays, basada en la hibridación de dos cadenas de DNA complementarias. Desafortunadamente, presenta algunas complicaciones en su diseño, derivadas de la necesidad de conocer a priori las secuencias de los organismos, lo que no siempre es posible cuando dichos organismos pertenecen a especies no modelo (López de Heredia, 2016). Para solventar los problemas derivados de estas limitaciones se han desarrollado técnicas más avanzadas basadas en la ultra-secuenciación del RNA y el alineamiento de todo el número de fragmentos de secuencia obtenidos de dichos RNAs con el genoma de referencia. Esta técnica se denomina RNA-Seq (*RNA-Sequencing*) que analiza y contabiliza todos los RNA de un conjunto de células con una secuenciación complete. Otra técnica más específica, en la que se aplica la misma metodología para la secuenciación de RNA en las células individuales previamente aisladas, es conocida como scRNA-Seq (*Single Cell RNA-Sequencing*). Ambas técnicas pertenecen a las tecnologías de nueva generación (NGS: *Next-Generation-Sequencing*). Esta última resulta más eficaz en cuanto al estudio de las distintas poblaciones celulares, pero presenta una serie de limitaciones ligadas al coste de tiempo y dinero requerido para conseguir aislar las células. No obstante, a pesar de algunas ventajas frente a la primera técnica, ambas conllevan el uso de un gran volumen de datos, por lo que la complejidad computacional al utilizar este tipo de información es considerablemente superior, siendo mayor en scRNA-Seq por el aislamiento y la secuenciación de células individuales.

Los métodos de carácter experimental (in vitro) diseñados para el estudio de la heterogeneidad celular más utilizados son los siguientes:

· *Inmunohistoquímica*: Esta técnica permite identificar marcadores antígenos a partir de las reacciones antígeno-anticuerpo (Alcances de la Inmunohistoquímica en el estudio de los tejidos, 2015).

· *Citometría de flujo*: Se trata de un método que permite el análisis de un tipo celular mediante la detección de determinados anticuerpos fluorescentes que las caracterizan. Para su análisis, las células son incubadas con anticuerpos específicos conjugados con fluorocromos fluorescentes, que serán detectados por los diferentes canales del citómetro, de esta manera obtenemos características como el tamaño, complejidad o el porcentaje de las células, entre otras.

Pero dichas técnicas no están exentas de limitaciones, ya que los marcadores fenotípicos o la disgregación tisular antes de finalizar todo el procedimiento de la citometría de flujo pueden representar obstáculos en el análisis. Como solución a este problema, han ido surgiendo diversos métodos de deconvolución, los cuales proponen algoritmos capaces de extraer determinada información sobre los tipos celulares de interés desde un punto de vista computacional (*in silico*), inferir las frecuencias relativas en perfiles de expresión en una mezcla (Shen-Orr et al., 2010) e identificar biomarcadores y poblaciones celulares actualmente desconocidos (Newman et al., 2015). Además, se distinguen dos clases de deconvolución en función de los resultados obtenidos; parcial, que estima las proporciones de los tipos celulares en cada muestra o la expresión de los genes específicos en dichas poblaciones celulares, y la deconvolución completa, cuya solución abarca ambos resultados.

En cuanto a la estructura del presente trabajo, primero se definirán algunos de los conceptos básicos de la biología molecular, como el transcriptoma, el proceso de la transcripción, la importancia y los objetivos del análisis de datos transcriptómicos y las técnicas encargadas de la obtención de la expresión génica. Después, se explicará la parte matemática y estadística del estudio, definiendo y detallando el

problema de la deconvolución, además de los algoritmos matemáticos que utilizan los métodos seleccionados para el análisis. Tras este primer bloque, se expondrán los objetivos del trabajo, los datos utilizados para el mismo y los métodos de deconvolución elegidos. Se aplicarán CIBERSORT y FARDEEP en datos génicos con señal de microarrays y con señal de RNA-Seq, añadiendo la implementación de los métodos DECONICA y LINSEED al primer tipo de datos y ABIS al segundo tipo. En el último punto de este apartado, se comentarán las medidas y pruebas estadísticas utilizadas para la evaluación de los métodos y la comparación entre ellos (coeficiente de correlación de Pearson y raíz del error cuadrático medio) y, para cerrar este capítulo, se detallará el procesamiento de los datos antes de aplicar los algoritmos. A continuación, se presentarán los resultados obtenidos para las distintas bases de datos, estructurados en tres bloques. En los dos primeros, se realizarán dos comparativas. La primera se basa en la comparación de cuatro métodos (CIBERSORT, FARDEEP, DECONICA y LINSEED) utilizando datos cuya señal de expresión génica ha sido detectada mediante microarrays, aplicados a dos bases de datos de células sanguíneas (GSE64385 y GSE20300). En la segunda comparativa, la comparación será entre dos bases de datos correspondientes a una misma mezcla, una contiene datos con señal de microarrays y la otra con señal detectada mediante RNA-Seq (GSE107011). El tercer y último bloque de esta sección, estará constituido por un análisis de los marcadores celulares a través del estudio de la matriz LM22, requerida para la aplicación de los métodos supervisados (CIBERSORT y FARDEEP), y las matrices de firmas que estiman los métodos DECONICA y LINSEED. Para finalizar, se añadirá una última parte en la que se discutirán y se expondrán las conclusiones referidas a la determinación del método que mejor funciona para los diferentes tipos de datos evaluados, y se concluirá con la mención de las referencias bibliográficas utilizadas durante todo el estudio.

El objetivo principal de este trabajo es el de presentar, estudiar, comparar y utilizar herramientas para descomponer mezclas biológicas complejas de elementos mediante la aplicación de **algoritmos matemáticos de deconvolución** (*mixture deconvolution*), y así identificar las firmas celulares correspondientes a cada tipo celular y calcular las proporciones de los elementos de las mezclas.

1.1. Análisis del transcriptoma

Existen tres tipos de moléculas que contienen y transportan la información dentro de un organismo (Principios de la biología molecular, 2016). La primera de ellas es el DNA (ácido desoxirribonucleico) que contiene codificada la información genética. La estructura de esta molécula se basa en dos cadenas, la principal y la complementaria, que se entrelazan conformando una estructura de doble hélice. Cada una de las cadenas está formada por azúcares (desoxirribosa) y grupos fosfato, las cuales se enrollan debido al proceso químico conocido como hibridación, en el que las cuatro bases nitrogenadas (también llamadas nucleótidos) que conforman esta molécula, adenina (A), citosina (C), guanina (G) y timina (T), se unen dos a dos mediante enlaces de puentes de hidrógeno (A con T y C con G, perteneciendo cada una de las bases a hebras distintas), manteniendo así las dos cadenas vinculadas (Figura 1.1).

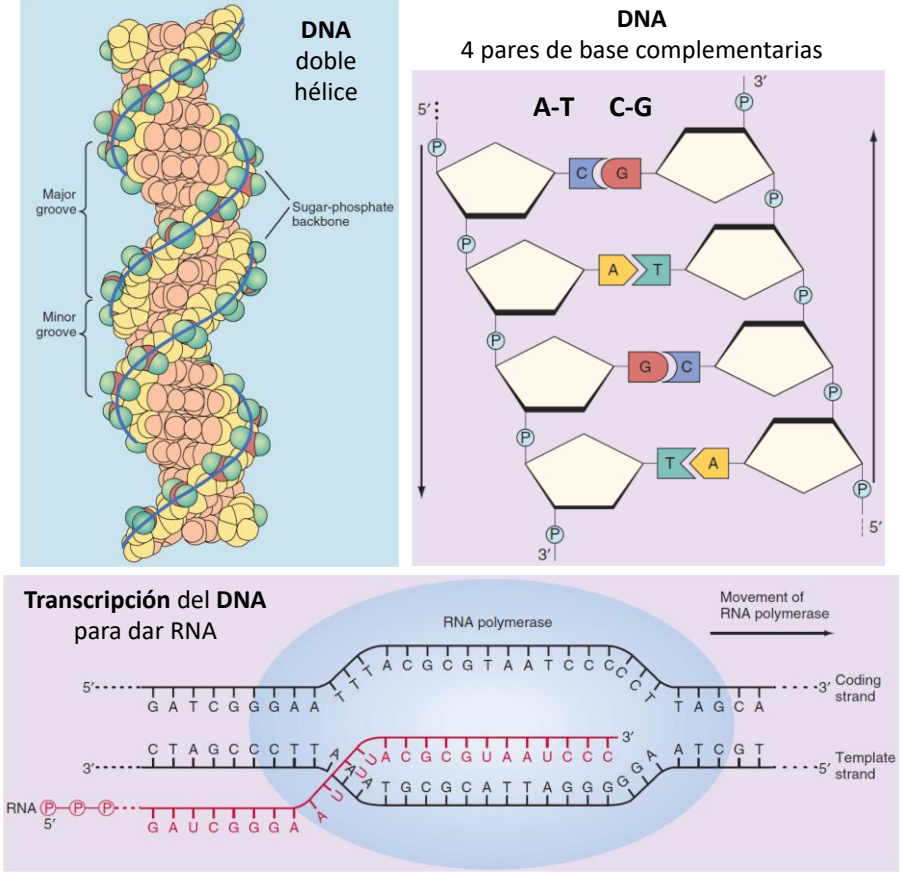


Figura 1.1. Dibujo de la estructura de doble hélice del DNA (cadena polinucleotídica), cuatro pares de base complementarias (A-T, C-G) que forman las secuencias de DNA, y esquema de la transcripción del DNA a RNA por acción de la RNA polimerasa. (Tomado de Elsevier: Principles of Medical Biochemistry, 4th Edition: Meisenberg & Simmons, 2021)

Las secuencias de las bases nitrogenadas determinan las proteínas y moléculas de RNA que se formarán (Deoxyribonucleic Acid (DNA) Fact Sheet, 2021). Cada segmento de DNA (secuencia ordenada de nucleótidos) constituye un gen, que es la unidad funcional y física de la herencia que trasciende de generación en generación (Diccionario de Cáncer Del NCI - Instituto Nacional Del Cáncer, 2021). La segunda molécula de interés es el RNA (ácido ribonucleico), encargado de transcribir la información genética que codifica el DNA. A diferencia de la anterior, ésta cuenta con una única cadena, el azúcar que lo forma es la ribosa y, en cuanto a sus bases nitrogenadas, contiene el uracilo (U) en lugar de la timina (T). Dentro de una célula, se diferencian tres tipos de RNA; RNA mensajero (mRNA), RNA ribosomal (rRNA) y RNA de transferencia (tRNA) (Ribonucleic Acid (RNA), 2021) cada uno de los cuales lleva a cabo funciones distintas dentro de dicha célula relacionadas con la síntesis proteica y la expresión génica. Por último, las proteínas son macromoléculas presentes en las células que desempeñan numerosas funciones de tipo estructural, mecánicas, bioquímicas y de señalización celular. Están formadas por otras moléculas de menor tamaño conocidas como aminoácidos (Proteína | NHGRI, 2021). Cada codón de un gen (tres nucleótidos) codifica el aminoácido que se transcribe y, como cada gen está

compuesto por múltiples codones, dicha secuencia posee las instrucciones necesarias para proporcionar la proteína final.

El DNA y el RNA se relacionan entre sí mediante el proceso de la **transcripción** (descrito en el dogma central de la biología junto con la traducción) mediante el cual se transfiere la información codificada en el **DNA** a través de su copia complementaria a una cadena de **mRNA** (RNA mensajero), y posteriormente esta información será trasladada a los ribosomas. Los ribosomas son la maquinaria celular encargada de **traducir** el mRNA para sintetizar la proteína. La secuencia de nucleótidos, determina la secuencia de aminoácidos de la proteína (Figura 1.2).

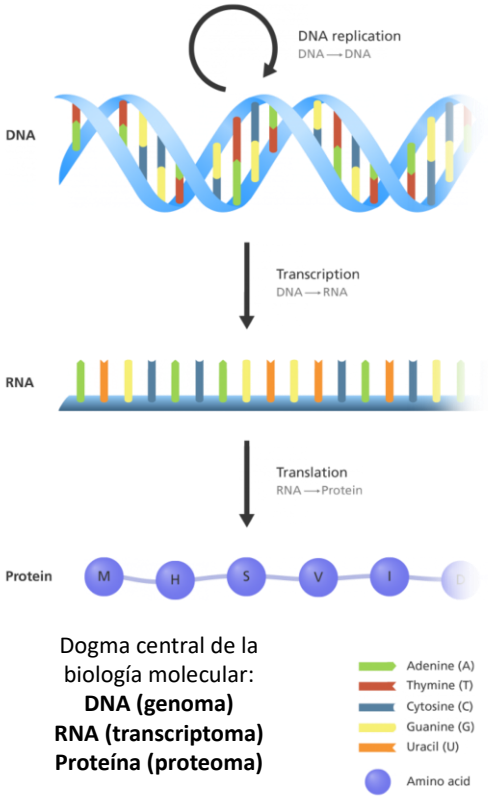


Figura 1.2. Figura representando el dogma central de la biología molecular que incluye DNA, RNA y proteína junto a los procesos de replicación (copia del DNA), transcripción (paso de DNA a RNA) y traducción (paso de RNA a proteína). También se indican las bases nitrogenadas que son la base de las secuencias de DNA (A, T, C, G) y de RNA (A, U, C, G). (Tomado de https://www.yourgenome.org/facts/what-is-the-central-dogma)

Se denomina transcriptoma al conjunto de moléculas de mRNA (también llamados transcritos) presentes en una célula o grupo de células en un momento determinado. Como se ha mencionado anteriormente, la secuencia de RNA viene dada por la del DNA, por tanto, el análisis del transcriptoma permite identificar cuándo y dónde se está transcribiendo, y por tanto está activado cada gen en las distintas células y tejidos. La disciplina encargada de este estudio se conoce como *Transcriptómica*. En los seres humanos, una gran parte de las células del organismo poseen los mismos genes, pero no la misma expresión génica, lo que determina las distintas propiedades y funciones de cada una de las células, tanto en una situación normal, como en estado de enfermedad, por lo que el estudio de su expresión puede resultar de gran interés. Por ejemplo, si un gen presenta niveles de expresión notablemente más altos en células cancerosas, podría ser que este llevara a cabo funciones relacionadas con el estado alterado que presenta la célula, tales como influir en la multiplicación celular, o si se expresa de manera considerable en células del tejido adiposo en lugar de en células de tejido muscular, podría estar desempeñando funciones relacionadas con el almacenamiento de grasa u otras implicadas con el metabolismo (National Human Genome Research Institute Home | NHGRI, 2021). En definitiva, representa un papel fundamental en la caracterización de genes marcadores de cada tipo celular y en la determinación de la importancia de estos en las mismas dependiendo de la función que desempeñan en ellas y las propiedades que las proporcionan.

Técnicas de detección de la expresión de los genes

A lo largo de los años, se han desarrollado varias técnicas para el estudio de la transcriptómica que, de la más antigua a la más reciente, se exponen a continuación:

- **Microarrays** (*Microarrays de oligos de alta densidad de expresión génica*): Se trata de una superficie sólida en la que se distribuyen secuencias cortas de oligo-nucleótidos (denominadas sondas) que son fragmentos de DNA complementarios a los transcritos expresados en una muestra estudiada de una especie en particular (en nuestro caso siempre DNA humano). Los transcritos expresados en una muestra se recogen, se fragmentan en tamaño igual a los oligos y se etiquetan con fluorescencia, para ser depositados en la superficie del microarray, dejando que hibriden con

las correspondientes secuencias de DNA. A continuación, se escanea la señal de fluorescencia para determinar los niveles de expresión de cada sonda (Romero Campero, 2019).

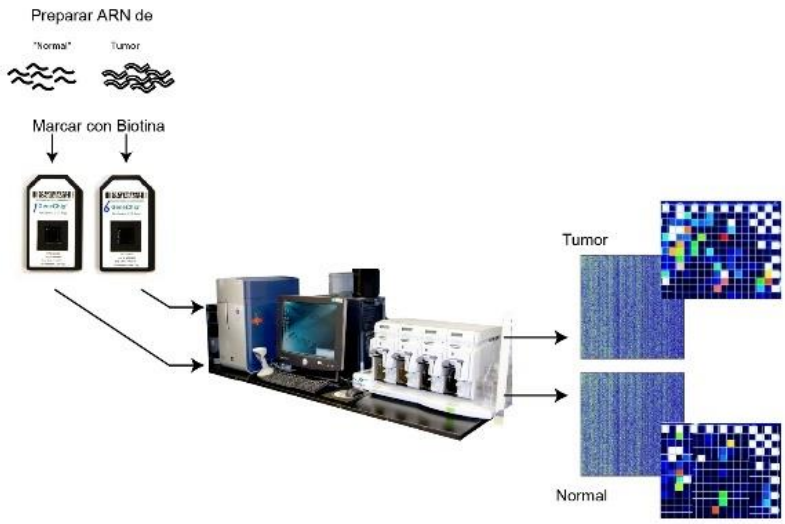


Figura 1.3. Tecnología de Microarrays.

(Tomado de Tecnología de Microarrays (Chips de ADN o RNA) | NHGRI, 2021).

➔ **RNA-Seq** (*RNA-Sequencing*): Consiste en la secuenciación del mRNA, que se romperá en fragmentos de longitud variable. La secuencia de estos fragmentos es recuperada, almacenada y alineados al genoma de referencia de la especie con la que estemos trabajando. Por último, obtenemos un conteo de fragmentos alineados para cada gen o exón.

➔ **scRNA-Seq** (*Single Cell RNA-Sequencing*): Es similar a la anterior, pero posee la particularidad de que la secuenciación de moléculas de RNA se realiza para cada célula de manera individual. Gracias a esto, es posible identificar nuevas poblaciones celulares poco comunes, hallar relaciones reguladoras entre genes y estudiar la conducta de diferentes clases de células en desarrollo (Hwang et al., 2018).

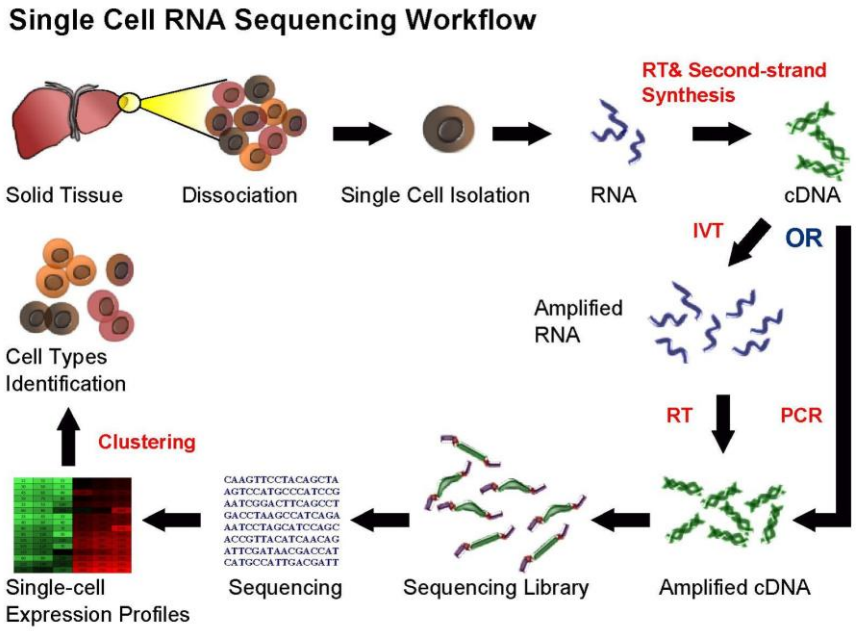


Figura 1.4. Técnica de Single Cell RNA-Sequencing. (Tomado de Vladimir Kiselev, 2019)

Como se ha mencionado anteriormente, la primera técnica que se desarrolló para estimar los niveles de expresión fue la tecnología de microarrays. El hecho de ser la primera proporciona algunas ventajas, como el gran conocimiento del protocolo a seguir en los laboratorios y los numerosos instrumentos analíticos testados. Sin embargo, presenta algunas desventajas tales como la posible aparición de problemas en la hibridación de las cadenas, la limitación a transcritos conocidos y una sensibilidad un tanto restringida debido a la saturación de las sondas (Romero Campero, 2019). Consecuentemente, los inconvenientes anteriores han creado la necesidad de diseñar nuevas técnicas basadas en la secuenciación masiva de moléculas de RNA en las que el cálculo de la expresión se obtiene por el recuento de fragmentos de RNA (*reads*) derivados de cada transcrito presente en la muestra. En primer lugar, se desarrolló RNA-Seq (*RNA-Sequencing*), que actualmente representa la técnica más utilizada para estimar los perfiles de expresión debido a diversas ventajas tales como la identificación de isoformas (distintos transcritos de un gen) y una mayor sensibilidad, con un rango dinámico de detección mucho más amplio (Romero Campero, 2019). No obstante, se trata de una técnica más costosa que la anterior, los análisis computacionales resultan más intensos y, además, se necesita mejorar y estandarizar los algoritmos de cálculo de señal.

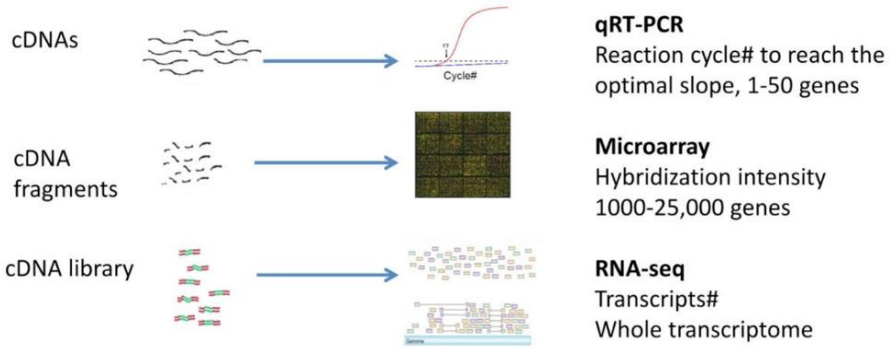


Figura 1.5. Técnicas de cuantificación de la expresión génica. (Tomado de Narrandes & Xu, 2018)

Posteriormente, apareció la técnica scRNA-Seq (*Single Cell RNA-Sequencing*), que se fundamenta en la separación de cada célula única (*single cell*) antes de proceder a la secuenciación por la misma metodología que RNA-Seq. De este modo, se consigue estudiar cada tipo celular independientemente, separando las poblaciones de distintos linajes. Esta técnica resulta más eficiente cuando el propósito del

estudio es descubrir nuevos biomarcadores y así conseguir caracterizar subgrupos celulares o identificar genes causales de enfermedades con el fin de explicar las respuestas patológicas, que pueden estar además asociadas a variaciones precisas en la expresión génica de ciertas células (Papalexi & Satija, 2018). Sin embargo, cabe destacar que los análisis computacionales llevados a cabo con este tipo de datos consumen mucho tiempo y una gran cantidad de memoria, además, la técnica tiene un alto coste económico, por lo que no siempre es posible abordar dichos estudios.

1.2. Deconvolución

1.2.1. Problema de análisis de mezclas y cálculo de componentes

Desde un punto de vista general de cálculo, se plantea a menudo el problema de analizar unas señales o valores numéricos que miden la cantidad total de varios elementos presentes en una mezcla. Habitualmente, se tienen los valores de la mezcla y se trata de calcular el número de elementos presentes, su proporción y su señal o valor individual. Para ello hay que descomponer la mezcla con un método matemático que nos permita identificar dichos elementos. Esta aproximación matemática se suele denominar deconvolución.

1.2.2. Definición del problema de deconvolución

Como se ha indicado, el análisis de mezclas se aborda matemáticamente por **deconvolución** (***deconvolution***). De este modo, el problema de la deconvolución se define como una operación matemática cuyo objetivo es la descomposición de un conjunto de señales en los elementos que la componen. Sabiendo que la convolución se refiere a la transformación o composición de dos funciones (o matrices), f y g, en una tercera, h (Arranz Rodríguez & Tabernero, 2019), se entiende como deconvolución al proceso inverso; la descomposición de una función (o matriz) de convolución en las funciones (o matrices) que se utilizaron en el proceso para separar sus efectos (Deconvolution | Definition of deconvolution by Oxford Dictionary, 2021):

$$h = f * g$$

Habitualmente, para explicar dicho problema matemático, se utiliza un ejemplo muy conocido llamado "*cocktail party problem*"(Cherry, 1953). Este experimento está ambientado en un cóctel con fondo musical donde se reúnen numerosas personas, cuyas voces son grabadas con varios micrófonos. El objetivo del estudio es disgregar las distintas voces mediante la separación ciega de fuentes. Este mismo concepto se puede aplicar a datos transcriptómicos, donde la señal de expresión de un gen es un cóctel y cada muestra un micrófono que recoge la mezcla de la señal (Czerwińska, 2018).

En relación al área de la transcriptómica, el problema de deconvolución pretende averiguar las proporciones de tipos celulares específicos y la búsqueda de marcadores celulares a partir de factores biológicos, que serán los genes activados. Los primeros en aplicar técnicas de este tipo en dicho campo fueron Peng Lu, Aleksey Nakorchevskiy y Edward M. Marcotte que, utilizando datos de expresión procedentes de señal de microarrays, estudiaron las proporciones de diferentes tipos celulares de levadura en distintas fases del ciclo celular (Lu et al., 2003).

En este trabajo, se aborda el problema de deconvolución aplicado a matrices. La matriz que se desea descomponer se corresponde con una mezcla de distintos tipos celulares en varias muestras, *T* ("matriz de mezclas"), y las matrices objetivo son la formada por la expresión génica de cada tipo celular, *C* ("matriz de firmas"), y la matriz de proporciones de estos mismos tipos celulares en las muestras mencionadas, *P* ("matriz de proporciones").

En este caso, el estudio se centrará en mezclas de distintos tipos celulares en muestras biológicas complejas. Se estudiarán mezclas de células sanguíneas en datos de microarrays y mezclas de células cerebrales en el caso de RNA-Seq (*RNA-Sequencing*).

1.2.3. Planteamiento y desarrollo del problema

Sean *n*, *m* y *c* el número de genes, muestras y tipos celulares, respectivamente. En términos matriciales, la deconvolución de datos transcriptómicos permite cuantificar las proporciones de las células en un tejido complejo (Abbas et al., 2009) y se define como una ecuación lineal de manera que:

$$T_{nxm} = C_{nxc} * P_{cxm} \quad (1)$$

Donde T_{nxm} representa la mezcla compleja que se desea descomponer formada por *n* genes y *m* muestras, que se obtiene del producto de dos matrices. La primera de ellas es C_{nxc}, la cual contiene la expresión de los *n* genes en los *c* tipos celulares estudiados, que al multiplicarse por la matriz P_{cxm}, cuyos elementos son las proporciones celulares en las muestras de la mezcla, se obtiene el primer objeto. Para que el proceso de deconvolución pueda llevarse a cabo, la matriz *P* debe cumplir dos requisitos:

i) Las columnas de la matriz (las muestras) deben sumar uno:

$$\sum_{j=1}^{m} P_{kj} = 1, \quad \forall\, k \in [1, \dots, c] \quad (2)$$

ii) Cada elemento de la matriz de proporciones debe valer mayor o igual que cero y como máximo uno:

$$P_{kj} \in [0,1] \qquad \forall\, k \in [1, \dots, c], \forall\, j \in [1, \dots, m] \quad (3)$$

Utilizamos la siguiente notación para plantear el problema:

g_{ij} = señal del gen *i* en la muestra *j* (*i*=1...*n*, *j*=1...*m*)

e_{ik} = señal del gen *i* en el tipo celular *k* (*i*=1...*n*, *k*=1...*c*)

p_{kj} = proporción del tipo celular *k* en la muestra *j* (*k*=1...*c*, *j*=1...*m*)

Con esta notación, el problema de deconvolución de datos de expresión se basa en el establecimiento de un conjunto de ecuaciones, una por gen para cada una de las muestras (en total $n \times m$) donde el valor g_{ij} es una combinación lineal del nivel de expresión e_{ik} del gen *i* en el tipo celular *k*, ponderado por la proporción p_{kj} del tipo celular *k* en la muestra *j* (Lu et al., 2003). Por lo tanto, para cada muestra fija *j* (1...*m*), el modelo se plantea de la siguiente forma:

$$\begin{cases} g_{1j} = p_{1j}e_{11} + p_{2j}e_{12} + \cdots + p_{kj}e_{1c} \\ g_{2j} = p_{1j}e_{11} + p_{2j}e_{12} + \cdots + p_{kj}e_{1c} \\ \qquad\qquad . \\ \qquad\qquad . \\ \qquad\qquad . \\ g_{nj} = p_{1j}e_{11} + p_{2j}e_{12} + \cdots + p_{kj}e_{1c} \end{cases}$$

En el anterior sistema de ecuaciones, se puede observar que las proporciones dependen únicamente del tipo celular y de la muestra, por lo que es constante para todos los genes de una misma muestra. Este mismo concepto se aplica a la expresión génica, es decir, sólo varía en función del gen y del tipo celular, por lo que este valor permanece estático independientemente de la muestra.

1.1.4. Tipos de deconvolución

Existen dos clases de deconvolución en función de los datos requeridos para su aplicación:

a) *Deconvolución parcial*: En este tipo, se necesita la matriz de mezclas *T* y una de las otras dos matrices que, dependiendo de cuál de las dos se facilite, se distinguen dos casos:
I) Si se proporciona la matriz *C* (matriz de expresión génica), entonces el resultado obtenido serán las proporciones de las diversas poblaciones celulares en las muestras consideradas en la mezcla.
II) Si la matriz que se introduce es la correspondiente a las proporciones celulares (*P*), el objetivo será estimar los perfiles de expresión de los genes en los tipos de células del tejido complejo.

b) *Deconvolución completa:* Únicamente se parte de la matriz de mezclas *T* y se obtendrán las otras dos matrices restantes (*C* y *P*). Por lo tanto, los métodos que son capaces de resolver este tipo de deconvolución permiten inferir tanto los perfiles de expresión como las proporciones de las poblaciones celulares presentes en la mezcla de estudio.

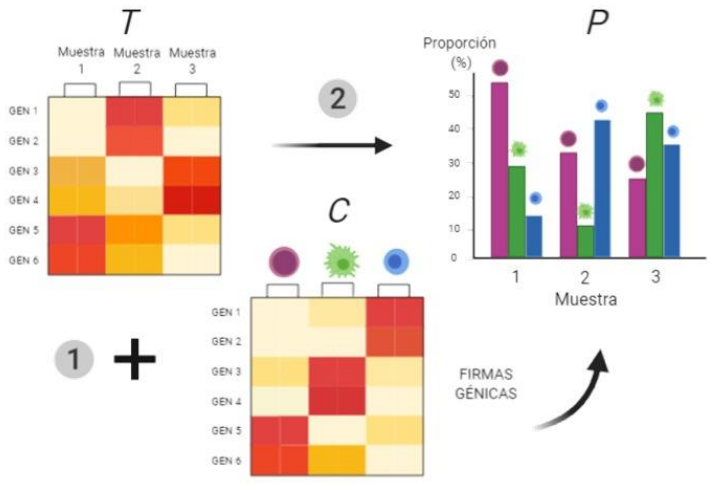


Figura 1.6. Tipo de deconvolución. 1) Deconvolución parcial 2) Deconvolución completa (Tomado Gutiérrez et al., 2020**)**

En el caso de este estudio, se utilizarán tanto métodos de deconvolución parcial (CIBERSORT, FARDEEP y ABIS), cuyo fin es estimar las proporciones celulares, como técnicas de deconvolución completa (DECONICA y LINSEED).

2. OBJETIVOS

Principalmente, mediante este trabajo se pretende realizar un análisis comparativo entre distintos métodos de deconvolución aplicados a datos transcriptómicos desde diferentes perspectivas. En dichos métodos, se desea evaluar la capacidad de calcular las proporciones de las diferentes poblaciones celulares y la identificación de los factores biológicos o características que mejor permitan realizar la separación en la mezcla considerada. Los factores buscados serán los genes activados, por ser los factores biológicos que se miden en estudios genómicos y transcriptómicos sobre las muestras analizadas. Para ello, se aplicarán un total de cinco métodos a datos de diferente naturaleza para responder a otros objetivos secundarios:

1. Exponer y plantear el problema matemático-estadístico de la **deconvolución**, entendida como la resolución de una mezcla de elementos en número y proporciones no conocidas que hay que desvelar/estimar.
2. Evaluar la precisión de los cinco **métodos de deconvolución** seleccionados en cuanto a la estimación de las proporciones de **diferentes tipos celulares** presentes en las **mezclas**, todo ellos basados en datos de expresión global de genes medidos en dichas mezclas.
3. Determinar el **método de deconvolución** de mezclas celulares que **mejor** funciona, utilizando como tests varias **series de datos de expresión** obtenidos con tecnologías de microarrays y de RNA-Seq, en los que se conocen las proporciones.

4. Estudiar la exactitud/precisión de las distintas técnicas utilizadas, en datos con diferente nivel de complejidad.
5. Analizar las **firmas de genes** que permiten identificar **cada tipo celular** (es decir, cada elemento de las mezclas), así como comparar y ver la intersección entre las firmas que infiere cada método o que se proporcionan previamente (en el caso de los métodos supervisados).

3. MATERIAL Y MÉTODOS

3.1. Datos

En este trabajo, aplicaremos los métodos seleccionados a cuatro mezclas distintas (*T*), que contienen tipos celulares sanguíneos (GSE64385, GSE20300, GSE106898 y GSE107011). Por otro lado, la señal de expresión génica de las tres primeras ha sido obtenida mediante la tecnología de microarrays, mientras que los niveles de expresión del último conjunto de datos, se han cuantificado mediante la tecnología de nueva generación RNA-Seq (*RNA-Sequencing*).

Es importante mencionar la dificultad en la búsqueda de mezclas en las que se pueden aplicar las técnicas diseñadas para la deconvolución, ya que es necesario tener las proporciones reales de los tipos celulares estudiados contenidos en la mezcla, lo que no siempre es posible. Esto se debe a que, sin esta información, resulta imposible analizar la precisión de cada método en cuanto al cálculo de las proporciones estimadas.

En el caso de los métodos supervisados (CIBERSORT, FARDEEP y ABIS), también se necesitará la matriz de firmas (*C*). Para los dos primeros conjuntos de datos, se utilizará la matriz LM22 proporcionada por CIBERSORT (Newman et al., 2015) y, para la descomposición de las otras dos mezclas, se implementarán las matrices de firmas dadas por el método ABIS (Monaco et al., 2019), una para datos de microarrays y otra para los de RNA-Seq (*RNA-Sequencing*).

A continuación, se presentarán cada uno de los conjuntos de mezclas que se utilizarán en este trabajo, todos ellos diseñados como conjuntos de datos para probar los diferentes métodos de deconvolución:

GSE64385: Se trata de una mezcla de células del sistema inmunológico y células tumorales. En este experimento (Becht et al., 2016) se seleccionaron cinco tipos de células inmunitarias de sangre periférica (**linfocitos T**, **linfocitos B**, **células NK**, **neutrófilos** y **monocitos**) mezcladas con otras correspondientes a la línea celular de cáncer de colon (**HCT116**) en un total de 12 muestras humanas, en las que se miden 54675 genes.

GSE20300: Esta mezcla está compuesta por cuatro tipos celulares procedentes de sangre periférica (**neutrófilos**, **linfocitos**, **monocitos** y **eosinófilos**) analizados en 24 muestras de pacientes pediátricos con trasplante renal estable y rechazo agudo (Shen-Orr et al., 2010), en cada una de las cuales se miden 54675 genes.

La expresión génica de células sanguíneas para los pacientes de los dos conjuntos de datos anteriores es medida en matrices de expresión del genoma completo (secuencia de DNA contenida en una célula) a través de la tecnología de microarrays (en concreto con la plataforma de *Affymetrix*). En cuanto a la fuente de procedencia, ambas mezclas se han obtenido de la página oficial del NCBI (*National Center for Biotechnology Information*), en concreto, de la plataforma GEO (*Gene Expression Omnibus*) mediante los códigos “GSE64385” y “GSE20300”.

GSE107019: En este experimento (Monaco et al., 2019), se estudiaron perfiles de expresión de microarrays y RNA-Seq de células mononucleares de sangre periférica (PBMC) de 13 individuos. Para ello, dichas muestras se analizaron utilizando plataformas de microarrays (*Affymetrix*) y de RNA-Seq (*Illumina*). Para estudiar la abundancia relativa, se consideraron diferentes números de tipos celulares para ambas plataformas. En el caso de la expresión obtenida mediante microarrays (mezcla GSE106898), la mezcla está formada por 11 tipos celulares, y para los datos de RNA-Seq (GSE107011), por 17. No obstante, en ambos experimentos se analizaron 17487 genes. Estos datos han sido obtenidos

a través del repositorio de GitHub del paquete ABIS https://github.com/giannimonaco/ABIS/tree/master/data (archivo "TPMPBMC.txt").

Una vez definidas las mezclas que se estudiarán en el trabajo, se expondrán las matrices de firmas (*C*) que se requieren para la descomposición de las anteriores en los métodos supervisados (CIBERSORT, FARDEEP y ABIS). La primera de ellas es la matriz **LM22**, formada por 22 subtipos celulares sanguíneos humanos. Los valores de expresión de esta matriz definen firmas únicas para cada subconjunto celular estudiado, que fueron generados utilizando la plataforma *Affymetrix* para datos de señal de microarrays (Chen et al., 2018). Las filas de esta matriz son los genes y las columnas se corresponden con los subtipos celulares. Es proporcionada por la página web de CIBERSORTx (https://cibersortx.stanford.edu/) y sólo se utilizará para la descomposición de las dos primeras mezclas (GSE64385 y GSE20300) ya que se requiere que los datos de expresión de la matriz de firmas y de la matriz de mezclas hayan sido analizados mediante la misma plataforma. Para la deconvolución de la mezcla restante se necesitan dos matrices de firmas distintas, una para la descomposición de la misma con la expresión génica procesada mediante microarrays y otra para la procesada con RNA-Seq (*RNA-Sequencing*). Ambos ficheros se obtuvieron del mismo repositorio de GitHub que el archivo de la mezcla (mencionado anteriormente) y los nombres son **"sigmatrixMicro.txt"** para las firmas de microarrays y **"sigmatrixRNAseq.txt"** para las de RNA-Seq (*RNA-Sequencing*). La primera matriz está constituida por 819 genes (filas) y 11 tipos celulares (columnas), mientras que la otra contiene la expresión de 1296 genes (filas) en 17 tipos celulares (columnas).

Se han recogido en una tabla las características principales de los conjuntos de datos que se han utilizado a lo largo de este trabajo. En ella, se mencionarán los códigos para poder encontrar las bases de datos en la página web del NCBI, las plataformas mediante las que se han analizado los datos de expresión génica, el número de muestras y el número de genes que forman la matriz, la procedencia de los tipos celulares y el número de los mismos y, por último, las referencias bibliográficas de los artículos científicos de los que se han obtenido:

Tabla 1. Resumen de los conjuntos de datos utilizados en el trabajo.

Nº acceso	Plataforma de expresión génica	Nº muestras	Nº genes	Fuente biológica	Tipos celulares	Referencia
GSE64385	Microarray HGU133 Plus 2.0 - *Affymetrix*	12	54675	PBMCs y celulas HCT116	5	(Becht et al., 2016)
GSE20300	Microarray HGU133 Plus 2.0 - *Affymetrix*	24	54675	Sangre periférica	4	(Shen-Orr et al., 2010)
GSE107011	RNA-Seq HiSeq 2000 - *Illumina*	13	17487	PBMCs	17	(Monaco et al., 2019)
GSE106898	Microarray Human HT-12 V4.0 - *Illumina*	13	17487	PBMCs	11	(Monaco et al., 2019)

3.2. Algoritmos utilizados en los métodos

Los métodos de deconvolución se basan en determinados algoritmos para el cálculo del resultado, que cambia en función del tipo de problema que resuelvan (deconvolución parcial o completa). En esta sección, se explicarán cada uno de ellos:

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

Técnica multivariante encargada de resolver problemas de clasificación o separación ciega de fuentes de señal ("*cocktail party problem*") (Czerwińska, 2018). Es semejante al análisis de componentes principales (PCA: *Principal Component Analysis*) salvo por algunos aspectos:

1. El objetivo es encontrar variables latentes (componentes) independientes entre sí, que es una condición mucho más fuerte que la búsqueda de componentes sin correlación, ya que la independencia implica esta otra propiedad.
2. Las variables latentes no presentan una distribución gaussiana.

ICA construye las variables latentes como combinaciones de las variables observables y asocia a cada conjunto de variables latentes una función objetivo (función de la asimetría o de la curtosis), que alcanza su valor máximo cuando se consigue la independencia estadística entre ellas, es decir, cuando

el producto de sus distribuciones marginales es lo más parecido posible a la distribución conjunta. En la separación ciega de fuentes, las variables latentes son los tipos celulares, y las observables están representadas por los genes. Entonces, si hay *n* variables observables (genes), *m* muestras, y *k* componentes en los que se quieren descomponer los datos, se propone un modelo de la forma:

$$T_{mxn} = A_{mxk} C_{kxn} \quad (4)$$

Donde tratamos de obtener una matriz *A* cuyos valores maximicen los estadísticos muestrales de asimetría o de curtosis de los tipos celulares.

Simplex

Desde su creación en 1947, el método del simplex se ha estado utilizando para resolver problemas de programación lineal (Dantzig, 1990). El objetivo de este algoritmo es optimizar (maximizar o minimizar) una función *z* teniendo en cuenta una serie de restricciones. Por lo tanto, siendo *A* una matriz de *m* × *n* dimensiones y *b* un vector formado por *m* componentes, el problema matemático se plantea de la siguiente manera:

$$\text{Min (max) } z = f(x)$$

$$Ax=b$$

En nuestro caso, la selección de los genes mutuamente relacionados y la estimación de los tipos celulares (eligiendo previamente el número de componentes k que se desea calcular, siendo como máximo el número de muestras menos uno, es decir, *m*-1), determinan el espacio simplex. La función que se desea optimizar para cada una de las muestras, es la equivalencia entre la expresión de los genes marcadores de cada tipo celular y el producto de las frecuencias relativas de dichas células, normalizadas por filas (por muestras) y multiplicadas por un coeficiente no negativo (α), que sume uno por cada muestra. Por otro lado, las restricciones planteadas están relacionadas con las propiedades fundamentales de la deconvolución: los coeficientes α deben valer mayor o igual que cero y la suma de este valor para cada muestra debe sumar uno. Por lo tanto, sea X la expresión de los genes seleccionados, H las proporciones celulares normalizadas por filas, y α los coeficientes no negativos que suman uno, dichos elementos forman un espacio *k*-dimensional:

$$\text{Max (min) } z: \ X = \alpha H$$

$$\alpha \geq 0 \ y \sum_{j}^{k} \alpha_j^i, \forall i = 1, \dots, m$$

Desde otra perspectiva, también se puede resolver dicho problema gráficamente. Para ello, se establece un sistema de coordenadas cartesianas (cada eje es una variable de decisión) en el que se dibujan las restricciones. Finalmente, los vértices del área correspondiente a la intersección de todas las regiones representan las posibles soluciones del problema. En el caso de datos transcriptómicos, los vértices (puntos óptimos) se corresponden con los genes marcadores y los tipos celulares estimados.

Modelo Lineal Robusto (RLM: *Robust Linear Model*)

Para *k* variables, un modelo lineal (LM: *Linear Model*) se define como:

$$y = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots \hat{\beta}_k x_k + \varepsilon \quad (5)$$

En el área de la transcriptómica, $\boldsymbol{y}$ representa la expresión de un gen en un conjunto de muestras, $\boldsymbol{x_1, x_2, \dots, x_k}$ representan los valores de la expresión de un mismo gen en distintas muestras, $\boldsymbol{\beta_1, \beta_2, \dots, \beta_k}$ son los coeficientes que determinan el cambio de *y* respecto de *x* (proporción del tipo celular $i \in [1, \dots, k]$) (Monaco et al., 2019) y $\boldsymbol{\varepsilon}$ el error cometido en la estimación.

Ahora bien, en el modelo lineal robusto (RLM: *Robust Linear Model*) el parámetro $\hat{\beta}$ está formado por el producto dos. En nuestro caso, estos parámetros son la proporción celular ($\hat{p}$) y la abundancia de mRNA ($\hat{\alpha}$), que deben ser siempre positivos:

$$\hat{\beta}_i = \hat{p}_i \hat{\alpha}_i$$

Por lo tanto, un modelo RLM se define de la siguiente manera:

$$y = \hat{p}_1 \hat{\alpha}_1 x_1 + \hat{p}_2 \hat{\alpha}_2 x_2 + \cdots \hat{p}_k \hat{\alpha}_k x_k + \varepsilon \quad (6)$$

Modelo Lineal Adaptado Recortado (aLTS: *Adaptative Least Timmed Squares*)

Se trata de un modelo lineal diseñado para la detección y eliminación de valores atípicos (*outliers*), cuya base parte del modelo lineal recortado (LTS: *Least Trimmed Squares*) propuesto por Qianqian Xu (Xu et al., 2017) con el objetivo de explorar dichos valores. Partiendo del planteamiento del modelo lineal (5) y considerando las mismas interpretaciones de los parámetros, el modelo lineal recortado adaptado incorpora valores atípicos:

$$y = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots \hat{\beta}_k x_k + \tau + \varepsilon \quad (7)$$

Donde $\tau = (\tau_1 + \tau_2 + \cdots + \tau_n)^t$ indica que el gen i-ésimo ($i \epsilon [1, \ldots, n]$) es un valor atípico. Sabiendo que $\hat{\beta}_{OLS} = (X^T X)^T X^T$ es la estimación de Mínimos Cuadrados Ordinarios (OLS: *Ordinary Least Square*) y $H = X(X^T X)^{-1} X^T$ es la matriz de proyección, los residuos se definen como:

$$r = y - X\hat{\beta}_{OLS} = (I - H)\tau + (I - H)\varepsilon \quad (8)$$

con media $(I - H)\tau$ y varianza $\sigma^2(1 - H)$.

Posteriormente, se establece el valor E como el conjunto de valores atípicos encontrados:

$$E = \{i : |r_i| > k + r_{med}\} \quad (9)$$

Donde k representa el parámetro de sensibilidad y r_{med} la mediana de los residuos.

El procedimiento de este algoritmo (aLTS) comienza con la estimación del parámetro E ($\hat{E}$), a través de la ecuación (9). Denotando como N y $\overline{N}$ el número total de elementos de E y $\hat{E}$, respectivamente, y Ñ representa una sobreestimación de N, lo que se puede utilizar para obtener una subestimación de la siguiente forma:

$$\underline{N} = \alpha_1 \overline{N}, \quad con\ \alpha_1\ \epsilon\ (0,1) \quad (10)$$

Utilizando la ecuación (10) se consigue ajustar el modelo de Mínimos Cuadrados Ordinarios eliminando los $\underline{N}$ valores atípicos y, consecuentemente, modificando el valor de E y de la sobreestimación de N ($\overline{N}$). Este proceso se repite iterativamente actualizando la subestimación del número de valores atípicos ($\underline{N}$) mediante el coeficiente α_2 (con $\alpha_2 > 1$) con el objetivo de alcanzar la convergencia entre el valor de la subestimación ($\underline{N}$) y de la sobreestimación ($\overline{N}$) de los valores atípicos (Hao et al., 2019). El algoritmo se inicia como una estimación de Mínimos Cuadrados Ordinarios:

$$\begin{aligned} \hat{\beta}^{(0)} &= (X^T X)^{-1} X^T y \\ r^{(0)} &= y - X\hat{\beta}^{(0)} \end{aligned} \quad (11)$$

Para la j-ésima iteración (con j ≥ 1) se actualiza el valor de $\overline{N}$:

$$\overline{N}^{(j)} = \begin{cases} \left| \left\{ i : \left| r_i^{(j-1)} \right| > r_{med}^{(j-1)} \right\} \right|, & j = 1 \\ \min\left(\left| \left\{ i : \left| r_i^{(j-1)} \right| > k \cdot r_{med}^{(j-1)} \right\} \right|, \overline{N}^{(j-1)} \right), & j \geq 2 \end{cases} \quad (12)$$

Como se muestra en la ecuación (12), se escoge el mínimo entre el número de valores atípicos sobreestimado calculado en la j-ésima iteración y el estimado en el paso anterior, lo que impide que $\overline{N}^{(j)}$ aumente. Posteriormente, se actualiza $\underline{N}^{(j)}$ utilizando la igualdad (10):

$$\underline{N}^{(j)} = \begin{cases} [\alpha_1 \overline{N}^{(j)}], & j = 1 \\ \min\left([\alpha_2 \underline{N}^{(j-1)}], \underline{N}^{(j)}\right), & j \geq 2 \end{cases} \quad (13)$$

Donde el operador min $(\cdot,\cdot)$ garantiza que el valor $\underline{N}^{(j-1)}$ no sobrepase $\underline{N}^{(j)}$. Posteriormente, se actualizan los valores utilizados en el modelo ($\hat{\beta}^{(j)}$ y $r^{(j)}$):

$$\hat{\beta}^{(j)} = \left(X^{(j)T} X^{(j)}\right)^{-1} X^{(j)T} y^{(j)}$$
$$r^{(j)} = y - X\hat{\beta}^{(j)} \quad (14)$$

Se repite el proceso hasta que $\overline{N}^{(j)}$ y $\underline{N}^{(j)}$ converjan hacia una misma solución.

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

Las máquinas de soporte vectorial (SVM: *Support Vector Machine*) se encargan de resolver problemas de optimización, tratando de encontrar el error máximo que permita clasificar correctamente el mayor número de puntos en un hiperplano (Awad & Khanna, 2015). A partir de esta técnica surge la regresión con máquinas de soporte vectorial (SVR: *Support Vector Regression*), que se define como un algoritmo de aprendizaje supervisado cuyo objetivo es representar este mismo hiperplano y estimar valores continuos mediante un modelo de regresión.

En un modelo de regresión lineal simple, el hiperplano está definido por la ecuación de una recta ($y = b + wx$) y la función que se pretende minimizar es la suma de cuadrados. Tomando como referencia un modelo de mínimos cuadrados ordinarios (OLS: *Ordinary Least Squares*), la función que se desea optimizar (minimizar) se muestra a continuación:

$$MIN \sum_{i=1}^{n} (y_i - w_i x_i)^2$$

Donde y_i es la variable dependiente y w_i son los coeficientes que determinan el grado de variación de la primera en función de x_i (variable independiente o predictora). A diferencia del modelo anterior, SVR presenta la característica de poder seleccionar el error (ε) que permita un mejor ajuste de los datos a partir de la búsqueda del hiperplano que resulte más adecuado (ε-*SVR*) o seleccionar los vectores de soporte (*v-SVR*) que a su vez definen los límites determinados por ε. Dicho error se utiliza para construir dos líneas, una a cada lado del hiperplano, denominadas bandas de soporte. Los puntos representados dentro de la región delimitada por estas líneas se ignoran, mientras que los valores atípicos (*outliers*) que se encuentran próximos a dichas bandas se conocen como vectores de soporte (*v*) y serán los puntos utilizados en la construcción de la función lineal. En este caso, la función que se desea minimizar es:

$$MIN\left(\frac{1}{2}||w||^2 + C \sum_{i=1}^{n} |\xi_i|\right) \quad (15)$$

Donde *C* es una constante (siempre mayor que cero) que permite controlar el error de manera que a medida que aumenta su valor también lo hace la tolerancia para los puntos fuera de ε (fuera del área delimitado por las bandas de soporte). Por último, ξ_i es el parámetro encargado de controlar el error cometido en la aproximación de las bandas de soporte, calculando la distancia entre los puntos representados fuera de las mismas y los límites de la región de aceptación.

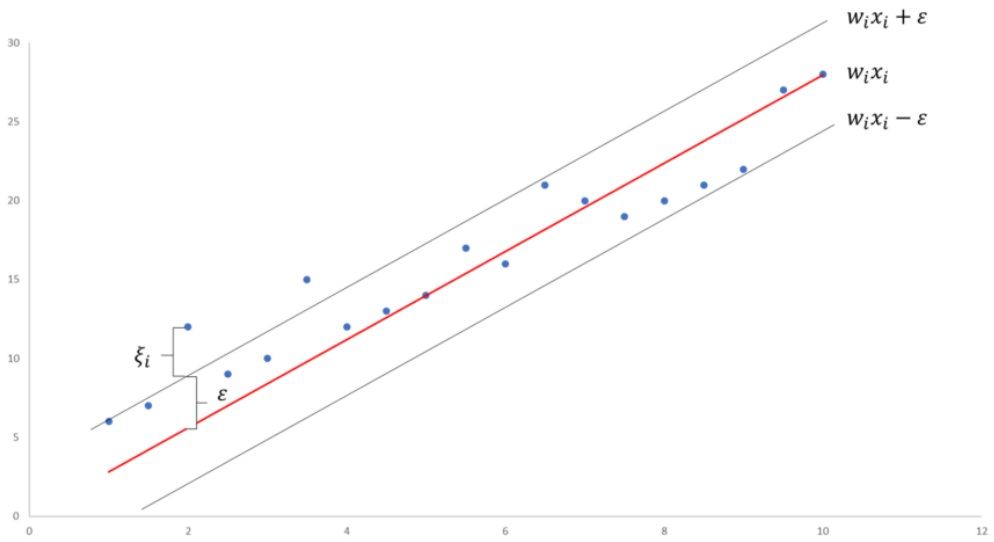


Figura 3.1. SVR: Support Vector Regression. (Tomado de Sharp, 2020)

En el caso de datos no lineales el procedimiento funciona exactamente igual, pero requiere una previa implementación de un núcleo (*kernel*) para obtener la linealidad de los datos y poder aplicar el mismo protocolo.

En el caso de este trabajo, el método de deconvolución en el que se implementa este algoritmo (CIBERSORT) se basa en el tipo de regresión *v-SVR*, seleccionando como vectores de soporte los genes incluidos en la matriz de firmas (firmas celulares).

3.3. Métodos de deconvolución

En este apartado, se explicarán los métodos seleccionados para la resolución del problema de deconvolución. Cada método se basa en alguno de los algoritmos descritos en el anterior apartado y están implementados en el software R (R: lenguaje de computación estadística, 2021). En este trabajo se ha modificado parte del código de cada método para adaptarlo al análisis de nuestras bases de datos.

DECONICA (*Deconvolution of transcriptome through Immune Component Analysis*)

Se trata de un método diseñado para resolver un problema de deconvolución completa (método no supervisado). Para ello, utiliza el algoritmo matemático ICA (*Independent Component Analysis*) para el cálculo de la expresión génica y las proporciones celulares. Su desarrollo se basa en el método fasTICA, diseñado para la descomposición de componentes independientes (o separación ciega de fuentes) en datos que contienen ruido gaussiano (Hyvarinen, 1999).

Con objeto de inferir unas variables latentes (tipos celulares) tan estadísticamente independientes entre sí como sea posible o, equivalentemente, que contengan distribuciones no Gaussianas, este método establece la ecuación (5) pretendiendo de este modo encontrar los genes marcadores a partir del cálculo de la matriz de firmas (*S*) para poder realizar la estimación de las proporciones celulares. Sea *mixture* (*n* × *m*) la matriz de mezclas *T*, considerando *n* genes, *m* muestras y *k* tipos celulares y seleccionando 'R' como el programa en el que se desarrollará el proceso, se aplica la función encargada de la deconvolución de los datos (`run_fastica()`), que se encuentra en el paquete 'deconica' (GitHub - UrszulaCzerwinska / DeconICA 2021), cuyo nombre corresponde a "deconvolución del transcriptoma mediante análisis de componentes inmunitarios":

```
library(deconica)

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

La salida (*output*) de esta función está formada por un conjunto de objetos:

X ($n \times m$): Matriz de mezclas de los datos preprocesados.

K ($m \times k$): Matriz que proyecta los componentes principales.

W ($k \times k$): Matriz de separación invertible.

A ($k \times m$): Matriz de las contribuciones de cada tipo celular en cada muestra.

S ($n \times k$): Matriz que conforma los pesos de cada gen (frecuencias absolutas) en los tipos celulares.

names (n): Vector que contiene el nombre de los genes.

samples (m): Vector que contiene el nombre de las muestras.

log.counts ($n \times m$): Matriz de mezclas en escala logarítmica (log2(mixture+1)).

Para calcular la matriz de firmas, se seleccionan los genes marcadores (los x genes que más caracterizan los k tipos celulares) y se calculan las frecuencias absolutas:

```
deconica_markers <- generate_markers(deconica, x,
                    return = "gene.ranked")
deconica_scores <- get_scores(deconica$log.counts, deconica_markers)
```

Para estimar las frecuencias relativas se dividen los valores obtenidos entre la suma total de la expresión génica:

```
deconica_scores2 <- deconica_scores/rowSums(deconica_scores)
```

LINSEED (*Linear Subspace identification for gene Expression Deconvolution*)

Este método fue diseñado en 2019 por Konstantin Zaitsev con objeto de resolver una deconvolución completa de perfiles transcripcionales, por lo que es capaz de identificar subpoblaciones celulares mediante la descomposición de mezclas biológicas complejas sin conocimiento a priori de los tipos celulares presentes en la misma. Además, se caracteriza por su robustez tanto al ruido técnico como al biológico. Para la resolución del problema, Linseed se basa en la propiedad de la mutua linealidad entre los genes específicos debido a que, en una situación ideal, la expresión de los estos genes (normalizados por filas) se comporta de manera lineal con las proporciones de los tipos celulares en las muestras (se cumple que $y = kp$, siendo y la expresión de los genes normalizada y p las proporciones de los tipos celulares en las muestras). Previamente, se realiza un filtrado de los genes no colineales y un análisis de descomposición en valores singulares (SVD: *Singular Value Descomposition*) para después proyectar los tipos celulares estimados (valores descompuestos mediante SVD) en un subespacio lineal (simplex) generado por las frecuencias de los tipos celulares dentro de las muestras (Zaitsev et al., 2019).

A continuación, se presentarán los comandos utilizados para la ejecución de este método en R. Primero se crea el objeto complejo de tipo linseed, utilizando el paquete 'linseed' (GitHub - Ctlab/LinSeed: Linseed (*LINear Subspace Identification for Gene Expresion Deconvolution*), 2021):

```
library(linseed)

lo <- LinseedObject$new(T)
```

En el siguiente paso se calcula la correlación entre los genes utilizando la correlación de *Spearman* y se seleccionan los *x* más correlacionados entre sí, de acuerdo con un cierto nivel de significación (α):

```
lo$calculatePairwiseLinearity()
lo$calculateSpearmanCorrelation()
lo$calculateSignificanceLevel(x)
lo$significancePlot(α)
lo$filterDatasetByPval(α)
```

Después, se estiman *k* tipos celulares en función de la representación del SVD:

```
lo$svdPlot()
lo$setCellTypeNumber(k)
```

Tras seleccionar los genes mutuamente relacionados e inferir los tipos celulares, se proyectan en un subespacio lineal (simplex) y se realiza la deconvolución:

```
lo$project("full") # projecting full dataset
lo$projectionPlot(color = "filtered")
lo$project("filtered")
lo$smartSearchCorners(dataset = "filtered", error = "norm")
lo$deconvolveByEndpoints()
```

Visualizamos las proporciones estimadas trasponiendo la matriz para que la matriz real y estimada tengan las mismas dimensiones:

```
plotProportions(lo$proportions)
dotPlotPropotions(round(lo$proportions, digits = 2), t(P),
guess = TRUE)
```

Por último, se accede a las proporciones estimadas a través del objeto linseed calculado inicialmente. Se debe trasponer la matriz para que la matriz inferida presente las muestras por filas y los tipos celulares por columnas, obteniendo un objeto con las mismas dimensiones que la matriz de proporciones original:

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

A pesar de su robustez al ruido, presenta dos inconvenientes importantes:

1) Cuando se considera un número muy grande de genes, resulta muy difícil calcular las correlaciones y, consecuentemente, la probabilidad de que el error sea mayor aumenta.
2) El número máximo de tipos celulares que se pueden calcular es igual al número de muestras menos uno (*m*-1), que se corresponde con el número máximo de restricciones que acepta el algoritmo simplex (algoritmo utilizado por LINSEED para la aproximación matemática).

ABIS (*ABsolute Inmune Signal deconvolution*)

Aunque la deconvolución de datos biológicos ha ido creciendo desde la primera técnica propuesta por Peng Lu en 2003, todavía quedan muchos aspectos por mejorar e investigar (Monaco et al., 2019). Uno de ellos es la aplicación de los mismos implementando datos de RNA-Seq (*RNA-Sequencing*), mucho más útiles en cuanto a la revelación de nuevas subpoblaciones celulares. Por ello, Gianni Mónaco propuso el método ABIS (*Absolute Inmune Signal deconvolution*), diseñado para la resolución de problemas de deconvolución parcial de datos transcriptómicos, por lo que se trata de un método supervisado. Para el cálculo de las proporciones, se basa en la normalización de los datos mediante la composición de mRNA (ácido ribonucleico mensajero). Además, cabe destacar que este método es capaz de abordar el problema tanto para datos de microarrays como para datos de RNA-Seq (*RNA-Sequencing*).

Antes de realizar la deconvolución, se debe normalizar la matriz de firmas multiplicando cada tipo celular por un factor de escala ($\hat{\alpha}$), que se calcula aplicando un modelo RLM (*Robust Linear Model*) utilizando la ecuación (16), de acuerdo a la igualdad $\hat{\beta}_i = \hat{p}_i\hat{\alpha}_i$ y tomando las mismas interpretaciones de los parámetros. Para la estimación de un $\hat{\alpha}$ óptimo, se desea minimizar la función del error cuadrático medio (RMSE: *Root Mean Square Error*) entre el valor de la proporción estimado ($\hat{p}$) y el real (p) para k tipos celulares:

$$\min_{\hat{\alpha}\epsilon(l,u)} \sqrt{\sum_{i=1}^{k} \widehat{(p_i} - p_i)^2}$$

Donde l y u son los límites inferior y superior del parámetro $\hat{\alpha}$, respectivamente. Para el cálculo del factor de escala en R, se aplica la función `optimize()` para cada uno de los tipos celulares, obteniendo de esta manera tantos $\hat{\alpha}$ como tipos celulares haya.

Una vez normalizados los datos, se realiza la deconvolución en el programa mencionado, para lo cual se necesita el paquete MASS (CRAN - Package MASS, 2021), encargado de estimar el modelo RLM:

```
library(MASS)

genes <- intersect(rownames(mixture),
   rownames(SignatureMatrixNormalized))

P_est <- apply(mixture[genes,], 2,

    function(x) coef(rlm(as.matrix(SignatureMatrixNormalized[genes,]),
    x, maxit =100))) *100

P_est <- round(P_est,3)
```

donde 'SignatureMatrixNormalized' es la matriz de firmas normalizada, 'mix' representa la matriz de mezclas *T* y en la última línea se selecciona el número de decimales que se desea tener en la matriz de proporciones.

La matriz de salida debe escalarse debido a que contiene valores negativos, además de no sumar 100 los valores de las proporciones en cada una de las muestras:

```
P_scale <- if(min(P_est)<0) {P_est+abs(min(P_est))} else {P_est}

P_scale <- apply(P_scale, 2, function(x) (x/sum(x))*100)
```

FARDEEP (*Fast And Robust Deconvolution of Expression Profiles*)

Un problema común en la resolución de problemas de deconvolución es la presencia de valores atípicos (*outliers*) en los datos. Por ello, Yuning Hao ha propuesto un método de deconvolución parcial denominado fardeep (*Fast And Robust Deconvolution of Expression Profiles*) diseñado para reducir el ruido de datos de expresión de genes tumorales detectando y eliminando valores atípicos (Hao et al., 2019). Esta herramienta se trata de una técnica supervisada, que permite calcular las composiciones celulares dentro de una muestra conociendo la expresión génica de las firmas correspondientes a los tipos celulares presentes en la mezcla.

Para la resolución del problema, FARDEEP se basa en el algoritmo aLTS (*Adaptative Least Trimmed Squares*). Como estamos trabajando con proporciones, los valores deben ser siempre mayores que cero, por lo que el modelo lineal que se plantea es una regresión de mínimos cuadrados no negativa (NNLS: *Non-Negative Least Squares*):

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} ||y - X\beta||_2^2, \text{ donde } \beta \geq 0$$

El algoritmo inicializa como se muestra en la ecuación (11), reemplazando el modelo MCO por NNLS:

$$\hat{\beta}^{(0)} = \underset{\beta}{\operatorname{argmin}} ||y - X\beta||_2^2, \quad \text{donde } \beta \geq 0;$$

$$r^{(0)} = y - X\hat{\beta}^{(0)}$$

Después se continúa con el procedimiento descrito en aLTS, calculando y actualizando los valores $\underline{N}$, $\overline{N}$, $\hat{\beta}^{(j)}$ y $r^{(j)}$en cada iteración. El proceso se repite hasta que $\underline{N}$ y $\overline{N}$ converjan hacia una misma solución ($\underline{N} = \overline{N}$).

En R se debe cargar primero la librería 'fardeep', perteneciente al paquete con este mismo nombre (GitHub - Cran/FARDEEP: CRAN R Package Repository., 2021) que contiene la función encargada de resolver el problema mediante este método:

```
library(fardeep)

result = fardeep(SignatureMatrix, mixture)
```

El objeto de salida muestra frecuencias absolutas, por lo que se debe dividir este resultado por la suma total de cada fila (muestra):

```
Scores_estF <- result$abs.beta

P_estF <- Scores_estF/rowSums(Scores_estF)
```

CIBERSORT

En 2015, Aaron M Newman propuso un método supervisado de deconvolución parcial capaz de estimar las frecuencias relativas de subpoblaciones celulares presentes en una mezcla a partir de sus biomarcadores (Newman et al., 2015).

Como se trata de un método supervisado, cibersort requiere una matriz de firmas con los biomarcadores de los subtipos celulares presentes en la mezcla. Como solución a esta cuestión, se pone a disposición la página web 'CIBERSORTx' (https://cibersortx.stanford.edu/) donde se proporciona una matriz de firmas validada para datos de microarrays de células inmunes (LM22), además de tutoriales que sirven como herramienta para construir una matriz de firmas para otros tipos de células que se deseen analizar utilizando datos de microarrays, RNA-Seq (*RNA-Sequencing*) o scRNA-Seq (*Single Cell RNA-Sequencing*).

El procedimiento de este algoritmo se basa en la aplicación de una regresión con máquinas de soporte vectorial (SVR: *Support Vector Regression*), un algoritmo de aprendizaje automático que pretende resolver el problema planteado en la ecuación (15). En concreto, el objetivo es resolver el tipo de regresión *v*-SVR (*v-Support Vector Regression*) seleccionando como vectores de soporte los genes marcadores presentes en la matriz de firmas que se introduce como entrada. En la función desarrollada en R, se aplica un núcleo (*kernel*) lineal para resolver la ecuación mencionada (15) y se introducen tres valores distintos para *v* (*v*={0.25,0.50,0.75}), pero únicamente se mantendrá aquel que proporcione un menor valor del error (RMSE: *Root Mean Square Error*). El código utilizado para la aplicación de dicha función se muestra a continuación:

```
res <- CIBERSORT(SignatureMatrix, mixture, perm = 100); res
```

donde 'SignatureMatrix' es la matriz de firmas, 'mixture' se refiere a la mezcla y el argumento 'perm' indica el número de permutaciones que se ejecutarán en el algoritmo (se recomienda un mínimo de 100), en cada una de las cuales se genera una muestra. Siendo *k* el número de tipos celulares presentes en la matriz de firmas, el resultado obtenido es una matriz con los siguientes elementos:

a) Desde la columna 1 a la columna *k*: Frecuencias relativas de los tipos celulares presentes en la matriz de firmas.
b) Columna *k*+1: P-valor obtenido tras la aplicación del método de Montecarlo a cada una de las muestras, considerando como hipótesis nula:

H_0: *La matriz T no presenta ningún tipo celular contenido en C*

c) Columna *k*+2: Valor de la correlación entre la proporción esperada y observada para cada muestra.
d) Columna *k*+3: Valor del error cometido en la estimación de frecuencia relativa en cada muestra.

Por lo tanto, únicamente se deben seleccionar las *k* primeras columnas correspondientes a las composiciones celulares:

```
P_est <- res[, 1:22]
```

Tabla 2. Resumen de las características principales de los métodos de deconvolución.

Método	Algoritmo de deconvolución	Supervisado	Selección recursiva de las variables	Software	Referencia
DECONICA	ICA	No	Sí	R (GitHub)	(Czerwińska, 2018)
LINSEED	Simplex	No	Sí	R (GitHub)	(Zaitsev et al., 2019)
ABIS	RLM	Sí	No	R (CRAN)	(Monaco et al., 2019)
FARDEEP	aLTS	Sí	No	R (CRAN)	(Hao et al., 2019)
CIBERSORT	*v*-SVR	Sí	No	R (cibersortX)	(Newman et al., 2015)

3.4. Medidas utilizadas para la evaluación y comparación de métodos

Tras la aplicación de los métodos, se aplicarán distintas herramientas estadísticas con el fin de medir la precisión en las estimaciones calculadas por los métodos y comparar estos últimos entre sí. A continuación, se expondrán cada una de dichas herramientas:

Coeficiente de correlación de Pearson

Se define como una prueba estadística cuyo objetivo es medir la dependencia lineal entre dos variables continuas. En nuestro caso, se desea medir la relación entre las proporciones reales y las estimadas. Para dos variables, X e Y, se calcula utilizando la siguiente ecuación:

$$r_{xy} = \frac{S_{xy}}{S_X S_Y}, \qquad -1 \leq r \leq 1$$

donde S_{xy} representa la covarianza entre ambas variables y S_X y S_Y son las desviaciones de cada una de ellas. Si el valor obtenido es negativo, indica una relación inversa de las variables (cuando una aumenta la otra disminuye) y si el coeficiente es positivo, entonces la relación entre ambas variables es directa (los valores de ambas variables aumentan o disminuyen simultáneamente).

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

El RMSE mide la diferencia entre dos conjuntos de datos o vectores de la misma dimensión. En concreto, calcula las diferencias entre los valores esperados o estimados y los reales u observados. En nuestro trabajo, utilizaremos esta medida para calcular las diferencias entre las proporciones estimadas y las originales para cada uno de los tipos celulares presentes en la mezcla. Para k tipos celulares y *n* muestras, el RMSE se calcula de la siguiente forma:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(p_i - \hat{p}_i)^2}$$

donde p_i es la proporción real del tipo celular i y $\hat{p}_i$ representa la proporción estimada del mismo tipo celular, para un determinado método.

En nuestro caso, como el trabajo se centra en el estudio de proporciones celulares, para determinar que un método funciona correctamente la correlación entre las proporciones calculadas y originales debe ser alta y positiva, indicando una relación directa, y el RMSE lo más pequeño posible, lo que supone que, cuanto más pequeño sea este valor, más se aproximan las frecuencias estimadas a las reales.

3.5 Preprocesamiento de los datos

El análisis y tratamiento de los datos se realiza a través del software estadístico R (R: lenguaje de computación estadística, 2021), en concreto, utilizando la versión 3.6.3. Este programa nos permite instalar los paquetes requeridos para la aplicación de los métodos y las medidas para su evaluación.

Preprocesamiento de datos formados por células de sangre periférica

Las dos bases de datos de células sanguíneas han sido generadas con tecnología de microarrays. Para su descarga, utilizamos las librerías *Biobase* (Huber et al., 2015) y *GEOquery* (Sean & Meltzer, 2007) y la función *getGEO()*:

```
library(Biobase)

library(GEOquery)

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

Donde en "numberAccession" debe ir el número de acceso de la base de datos que se desea descargar (en nuestro caso GSE64385 o GSE20300), el argumento *GSEMatrix* indica si se desean mantener las columnas de la matriz original (*GSEMatrix=TRUE*) y *AnnotGPL* se refiere a la conservación de la información actualizada sobre la anotación de los genes (*AnnotGPL=TRUE*). La ejecución de este comando muestra como salida un objeto complejo de tipo *ExpressionSet* con toda la información sobre la matriz de expresión de los genes, que contiene tanto el identificador de las sondas de estos últimos como su símbolo HUGO (nomenclatura que utilizaremos para el tratamiento de datos genéticos). Además de contener la información de la matriz de mezclas (*T*), este objeto complejo también nos proporcionará la información sobre la matriz de proporciones original (*P*).

a) **Matriz de mezclas *(T).*** Sea *mix2* la matriz de mezclas *T*:

```
mix2 <- exprs(numberAccession)
```

La matriz de expresión obtenida contiene como filas los nombres de los genes, y como columnas las muestras. Sin embargo, los nombres de los genes se corresponden con las sondas y, como se ha mencionado anteriormente, se utilizarán como identificadores los símbolos HUGO (*HGNC Database, HUGO Gene Nomenclature Committee (HGNC)*, 2021), por lo que se deben sustituir por estos últimos:

```
Gene_Symbol <- GSE64385@featureData@data[["Gene symbol"]]

row.names(mix2) <- Gene_Symbol
```

b) **Matriz de proporciones *(P)***

Primero, se crea un objeto con las propiedades y características de las muestras (nombre, número de acceso, tipo de células que contiene, la cantidad de mRNA en nanogramos y el estado del paciente del que se ha extraído la muestra sanguínea):

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

Después introducimos las proporciones reales proporcionadas por la página NCBI. Por ejemplo, para el primer tipo celular de la mezcla con número de acceso "GSE64385":

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

Por último, se transforman los valores absolutos en relativos:

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

c) Matriz de firmas *(C)*

Para las mezclas de células sanguíneas, se utilizará como matriz de firmas la matriz LM22 (definida en la sección 3.1). En R, se lee como un fichero de texto, se asignan como filas los nombres de los genes con nomenclatura HUGO localizados en la primera columna y, posteriormente, se elimina dicha columna:

```
LM22 <- read.table("LM22.txt", header=TRUE,  sep = "\t")
C <- LM22
row.names(C) <- C[,1]
C <- C[,-1]
```

Preprocesamiento de datos de sangre periférica mononuclear (PBMC)

Los objetos necesarios para la aplicación y evaluación de los métodos, tanto para los datos con señal de microarrays como para los de RNA-Seq (*RNA-Sequencing*), son proporcionados por Gianni Monaco como archivos de texto en su GitHub (https://github.com/giannimonaco/ABIS). Para ambos tipos de datos, se utilizan las mismas matrices de mezclas y de proporciones ("TPMPBMC.txt" y "Proportions.xlsx"). No obstante, como ambas mezclas no están formadas por el mismo número de tipos celulares (11 tipos en datos de microarrays y 17 tipos con señal de RNA-Seq), la matriz de frecuencias relativas se debe transformar de modo que se conserven únicamente los tipos celulares presentes en cada mezcla:

a) Proporciones de los datos con señal de microarrays:

```
# Seleccionamos 11 tipos celulares en la matriz de proporciones

CellTypes_11 <- c("B.Naive", "B.Memory", "Plasmablasts", "T.Naive",
                  "T.Memory", "NK", "pDCs", "mDCs", "Monocytes",
                  "Neutrophils.LD", "Basophils.LD")

# Creamos la matriz con los tipos celulares de interés

index <- match(CellTypes_11, colnames(P))

P_11 <- P[,index]
```

b) Proporciones de los datos con señal de RNA-Seq (*RNA-Sequencing*):

```
# Seleccionamos 17 tipos celulares en la matriz de proporciones

cellTypes_17 <- c("B.Naive", "B.Memory", "Plasmablasts", "T.CD4.Naive",
"T.CD8.Naive", "T.CD4.Memory", "T.CD8.Memory", "T.gd.Vd2","T.gd.non-
Vd2", "MAIT", "NK", "pDCs", "mDCs", "Monocytes.C", "Monocytes.NC+I",
"Neutrophils.LD", "Basophils.LD")

# Creamos la matriz con los tipos celulares de interés

index <- match(cellTypes_17, colnames(P))

P_17 <- P[,index]
```

En cuanto a la matriz de firmas, los archivos descargados del GitHub son “sigmatrixMicro.txt” para las firmas de microarrays y “sigmatrixRNAseq.txt” para las de RNA-Seq (*RNA-Sequencing*), en los que los nombres de los genes (con símbolos HUGO) se disponen en las filas y las muestras en las columnas.

4. RESULTADOS

En esta sección, se presentarán las soluciones obtenidas tras la implementación de los diferentes métodos de deconvolución seleccionados y escogiendo, posteriormente, el método que sea capaz de optimizar las proporciones calculadas desde cada uno de los tres enfoques analíticos:

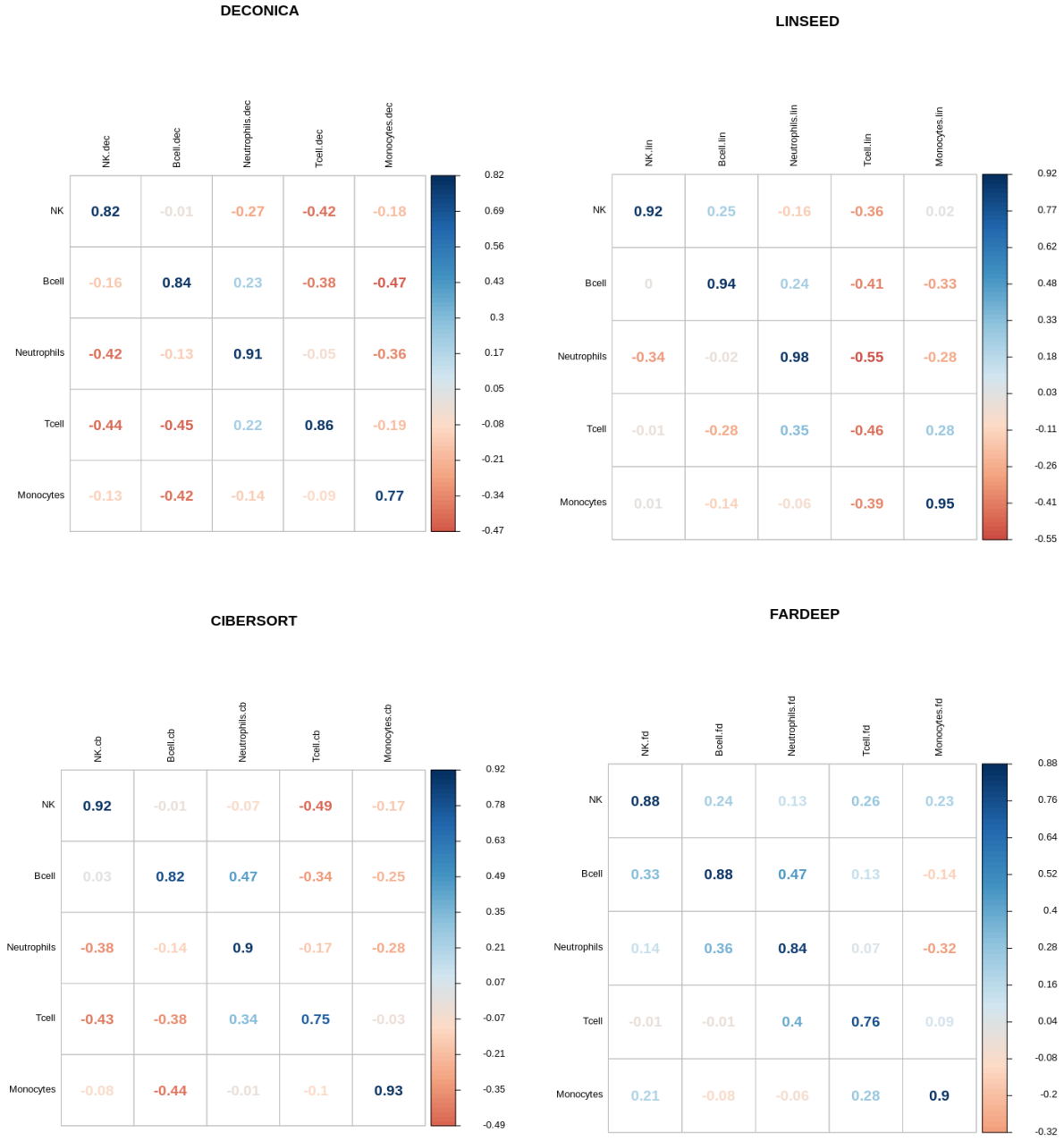
4.1. Comparación de métodos en datos con señal de microarray

En primer lugar, se han aplicado los cuatro métodos considerados en esta comparativa (DECONICA, LINSEED, CIBERSORT Y FARDEEP) en dos conjuntos de datos de células obtenidas de la sangre (correspondientes a los conjuntos de datos GSE64385 y GSE20300), cuya señal de expresión génica ha sido analizada mediante la tecnología de microarrays. Los resultados del análisis comparativo de los métodos se presentan mediante una serie de gráficos (varios de ellos diseñados de modo expreso en este trabajo para ilustrar y representar mejor los datos), que tienen siempre por objetivo contrastar en cada caso las proporciones estimadas de los distintos tipos celulares con las proporciones conocidas determinadas experimentalmente:

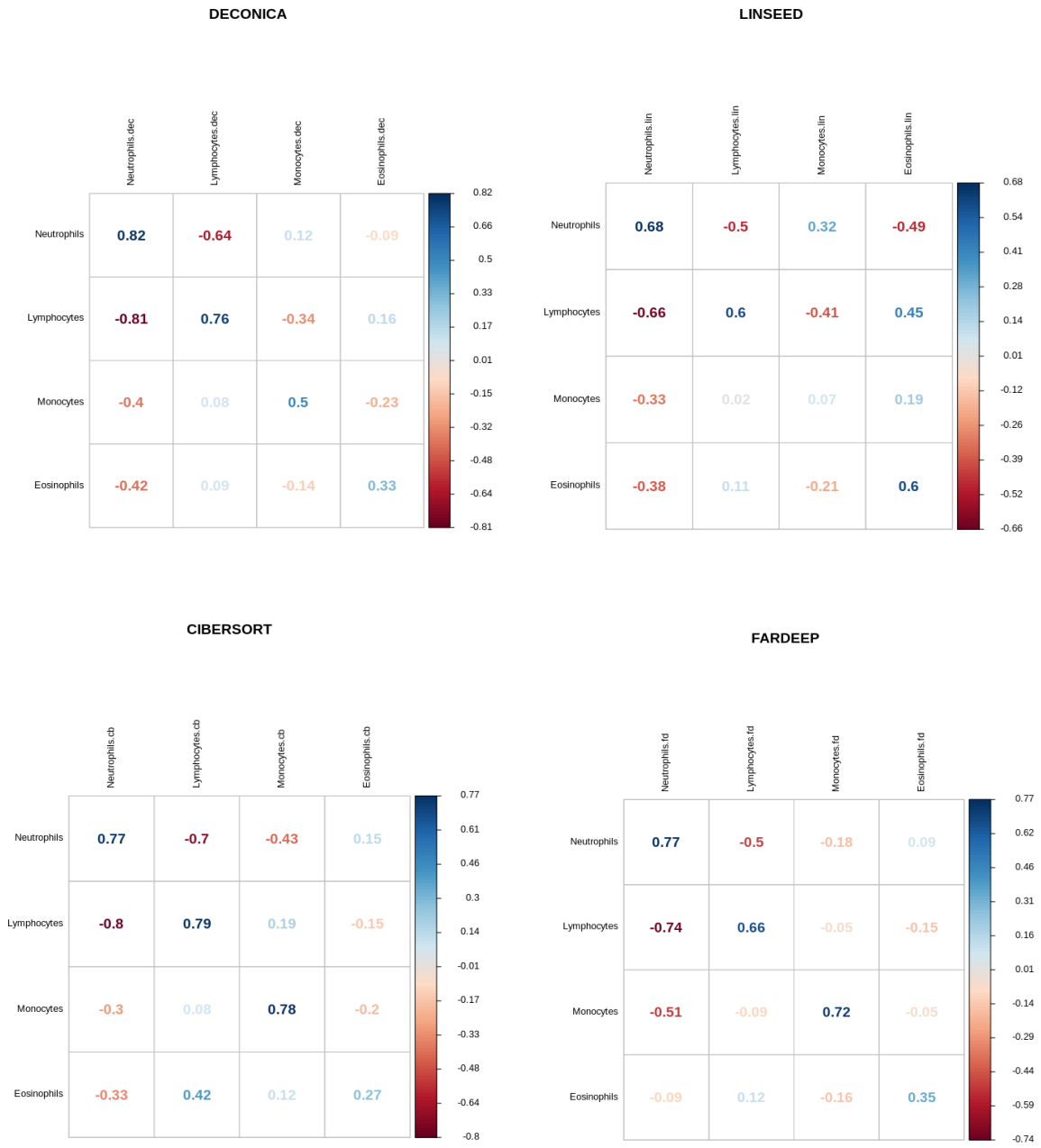
1. **Gráfico de correlaciones** (*corrplot*): que representa el coeficiente de correlación (*r, Pearson*) entre las proporciones estimadas y las conocidas para cada tipo celular incluido en las mezclas por cada método.
2. **Mapas de calor** (*heatmap*): que representan con la intensidad de color la abundancia relativa de cada tipo celular presente en las muestras comparando los datos conocidos originales con los estimados (en dos gráficos adjuntos) para cada uno de los métodos.
3. **Gráficos de firmas** (*cell signature plot*): que presentan para cada tipo celular por separado los valores relativos de abundancia (en porcentaje, %) en la serie de muestras estudiada, presentando dos trayectorias, para los datos reales y para los datos estimados, por un método. Estos gráficos incluyen además el cálculo de la correlación de *Pearson* y el valor de RMSE (raíz del error cuadrático medio) entre dichas trayectorias.
4. **Gráficos de barras apiladas** (*bar mixture plot*): que presentan en dos barras comparadas (de los datos reales y de datos estimados), para cada muestra, las abundancias relativas (proporcionales) de cada tipo celular apiladas (representadas con colores específicos).

Primero, se expondrán los gráficos de correlaciones para ambos conjuntos de datos, en los que se calculan los valores de la correlación de Pearson entre las proporciones estimadas por los métodos y las proporciones originales existentes en la mezcla:

GSE64385


DECONICA
NK.dec
Bcell.dec
Neutrophils.dec
Tcell.dec
Monocytes.dec
NK 0.82 -0.01 -0.27 -0.42 -0.18
Bcell -0.16 0.84 0.23 -0.38 -0.47
Neutrophils -0.42 -0.13 0.91 -0.05 -0.36
Tcell -0.44 -0.45 0.22 0.86 -0.19
Monocytes -0.13 -0.42 -0.14 -0.09 0.77
0.82 0.69 0.56 0.43 0.3 0.17 0.05 -0.08 -0.21 -0.34 -0.47
LINSEED
NK.lin
Bcell.lin
Neutrophils.lin
Tcell.lin
Monocytes.lin
NK 0.92 0.25 -0.16 -0.36 0.02
Bcell 0 0.94 0.24 -0.41 -0.33
Neutrophils -0.34 -0.02 0.98 -0.55 -0.28
Tcell -0.01 -0.28 0.35 -0.46 0.28
Monocytes 0.01 -0.14 -0.06 -0.39 0.95
0.92 0.77 0.62 0.48 0.33 0.18 0.03 -0.11 -0.26 -0.41 -0.55
CIBERSORT
NK.cb
Bcell.cb
Neutrophils.cb
Tcell.cb
Monocytes.cb
NK 0.92 -0.01 -0.07 -0.49 -0.17
Bcell 0.03 0.82 0.47 -0.34 -0.25
Neutrophils -0.38 -0.14 0.9 -0.17 -0.28
Tcell -0.43 -0.38 0.34 0.75 -0.03
Monocytes -0.08 -0.44 -0.01 -0.1 0.93
0.92 0.78 0.63 0.49 0.35 0.21 0.07 -0.07 -0.21 -0.35 -0.49
FARDEEP
NK.fd
Bcell.fd
Neutrophils.fd
Tcell.fd
Monocytes.fd
NK 0.88 0.24 0.13 0.26 0.23
Bcell 0.33 0.88 0.47 0.13 -0.14
Neutrophils 0.14 0.36 0.84 0.07 -0.32
Tcell -0.01 -0.01 0.4 0.76 0.09
Monocytes 0.21 -0.08 -0.06 0.28 0.9
0.88 0.76 0.64 0.52 0.4 0.28 0.16 0.04 -0.08 -0.2 -0.32

GSE20300



En el primer conjunto de datos, se observa que las correlaciones calculadas son ligeramente más altas en los métodos supervisados (CIBERSORT y FARDEEP). No obstante, el método DECONICA también presenta buenos resultados considerando esta medida. Por otro lado, LINSEED no es capaz de reconocer los linfocitos T, pero en el resto de tipos celulares se obtienen buenas correlaciones entre las frecuencias observadas y estimadas. En general, las correlaciones calculadas con los datos de la segunda mezcla de células sanguíneas (GSE20300) son más bajas. De hecho, ningún método es capaz de identificar los cuatro tipos de células, pero como ocurre en la descomposición de la primera mezcla, se obtienen mejores resultados utilizando los métodos supervisados (CIBERSORT y FARDEEP).

Para apreciar mejor los valores de las proporciones, y no su tendencia, se ha realizado un mapa de calor (*heatmap*) para cada método y conjunto de datos estudiado, en el que cuanto mayor es la proporción de cada tipo celular en una muestra, más oscuro se muestra el color del gráfico. A continuación, se mostrará el caso de CIBERSORT en la primera mezcla de células sanguíneas (GSE64385), mientras que el resto de gráficos se pueden encontrar en los datos suplementarios (**gráfico S11**):

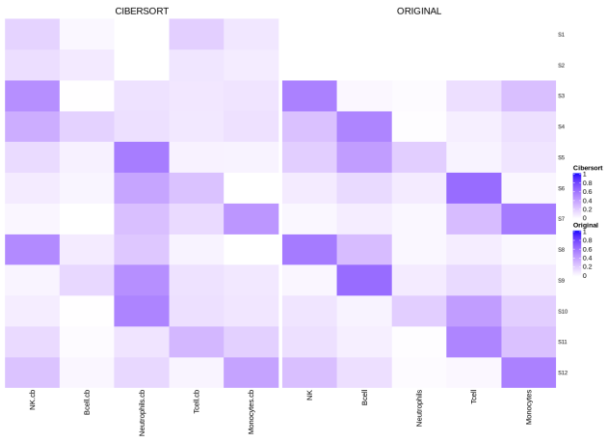


Figura 4.1. Mapas de calor (heatmaps) presentando la abundancia relativa de los distintos tipos celulares (columnas) en la serie de 12 muestras (filas) estimada por CIBERSORT, frente a la abundancia relativa ORIGINAL conocida para dichos tipos celulares.

En el *heatmap* anterior se observa que, aunque se habían obtenido valores de correlación altos utilizando las proporciones estimadas por dicho método, CIBERSORT no es capaz de calcular correctamente las proporciones de las muestras puras de células cancerígenas. Esto se debe a que, por tratarse de un método supervisado, requiere una matriz de firmas con todos los tipos celulares que se desean identificar en la mezcla. Sin embargo, este tipo de gráficos no desvelan mucha información contenida en los datos, por lo que se ha decidido diseñar nuevas formas para la evaluación de los métodos. Dichas técnicas de evaluación se tratan de dos tipos de gráficos en los que se aprecian con más detalle los resultados obtenidos: un gráfico de firmas (*cell signature plot*) en el cual aparecen representadas mediante puntos las proporciones estimadas (puntos azules) y las proporciones reales (puntos rojos) para cada tipo celular, y otro de barras apiladas (*bar mixture plot*), en el que las frecuencias relativas de cada tipo celular contenidas en cada muestra se engloban en un mismo gráfico. Como ejemplo, los siguientes gráficos de firmas representan la precisión de los métodos no supervisados (DECONICA y LINSEED), introduciendo en los algoritmos la primera mezcla de células sanguíneas (GSE64385). Dichos gráficos, que aparecen en la parte inferior, desvelan mucha más información intrínseca contenida en los resultados. En los gráficos de correlación (*corrplots*), los valores para DECONICA eran elevados. No obstante, estos gráficos nos muestran que, aunque la tendencia en las frecuencias relativas estimadas y reales es la misma, los valores de las proporciones calculadas por el método se distribuyen en torno a un cierto valor medio, por lo que no hay variación entre los calculados para cada muestra y tipo celular. Los gráficos realizados para el resto de métodos y para el conjunto de datos GSE20300, se exponen en los datos suplementarios (**gráficos S15** y **S16**):

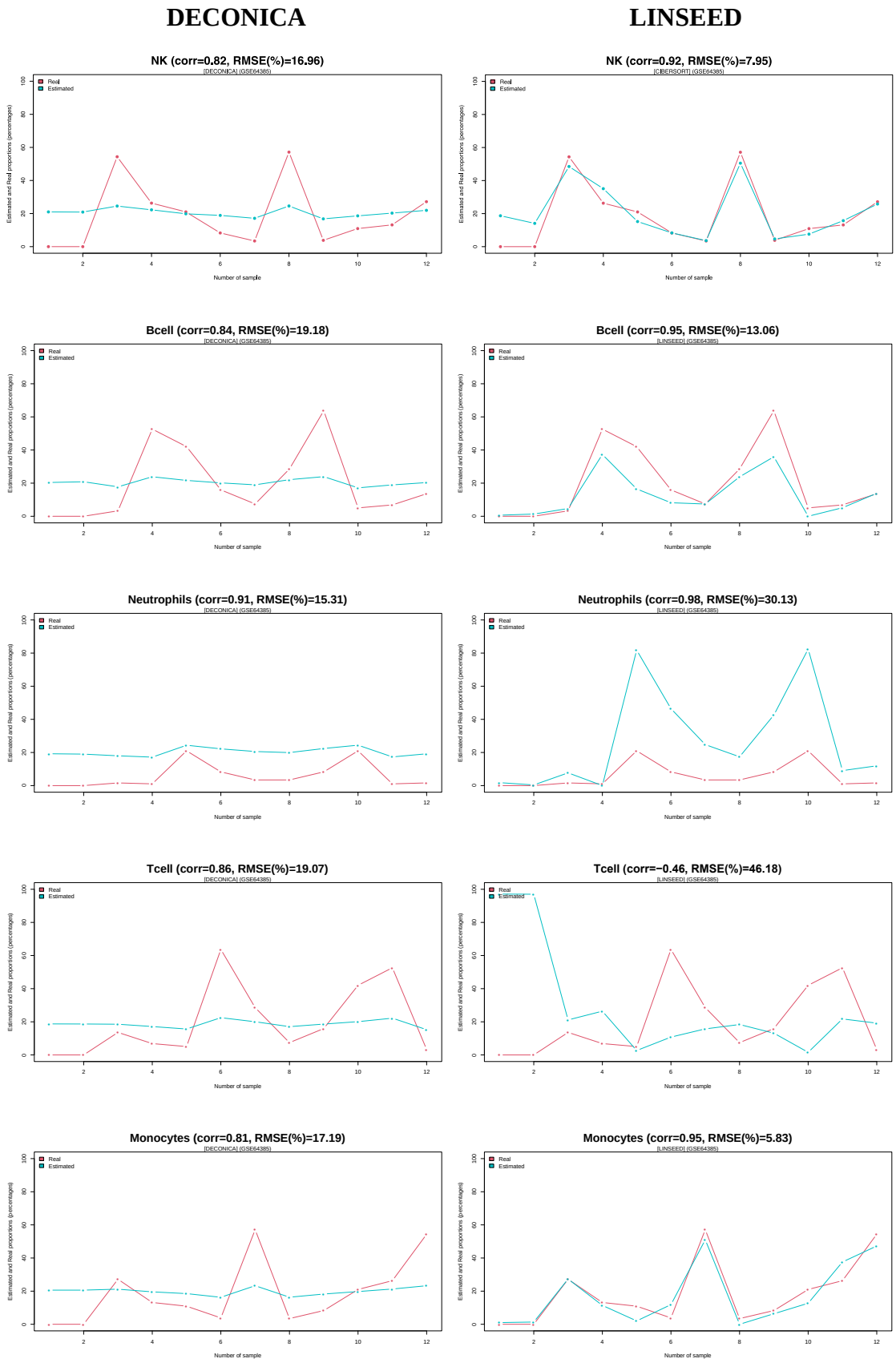


Figura 4.2. Comparación de las proporciones estimadas por LINSEED y DECONICA (en la mezcla GSE64385) y las proporciones reales, a través de un cell signature plot.

En los gráficos anteriores, se observa que los errores (RMSE) cometidos en la estimación utilizando DECONICA, se aproximan a un 20%, que en general son mayores que los cometidos por LINSEED, por lo que se puede decir que las distribuciones de los tipos celulares estudiados, proporcionadas por

este último, se parecen más a las distribuciones originales. Sin embargo, como se muestra en el gráfico de correlación, LINSEED no es capaz de identificar las células T, deducción que se vuelve a concluir mediante estos gráficos, cometiendo un error del 46.18% en la estimación de dichas células. En los correspondientes a los métodos supervisados (CIBERSORT y FARDEEP), **gráfico S15** de los datos suplementarios, se observan correlaciones más altas y errores más pequeños en ambos métodos, siendo el RMSE más pequeño en FARDEEP, ya que se trata de un método diseñado con el objetivo de eliminar los *outliers* antes de realizar la deconvolución de la mezcla (en nuestro caso el ruido está representado por las células cancerígenas, presentes en las dos primeras muestras). En cuanto al análisis de los resultados obtenidos tras la descomposición de la segunda mezcla (GSE20300), se extrae la misma conclusión para DECONICA (no hay variación significativa entre las proporciones de los distintos tipos celulares y las diferentes muestras), LINSEED tampoco proporciona buenos resultados, ya que la distribución de los valores estimados no se parece demasiado a la de los valores originales. En este último conjunto de datos, se observa una clara mejoraría en la estimación con la implementación de los algoritmos supervisados (CIBERSORT y FARDEEP) que, al contrario que en los resultados obtenidos en la descomposición de la otra mezcla, las correlaciones son más altas y los errores más pequeños en los resultados proporcionados por CIBERSORT, por lo que en este caso se escogería este método como el más preciso. No obstante, cabe destacar que la última mezcla no contiene ruido en los datos, es decir, no está formada por otro tipo de células que no sean las células sanguíneas, lo que supone una condición importante en la elección de uno de los dos algoritmos para la descomposición de una mezcla determinada. Para extraer información más precisa, también se ha decidido analizar los resultados mediante gráficos de barras (*bar mixture plot*). A continuación, se muestran los gráficos correspondientes a los métodos supervisados CIBERSORT y FARDEEP, tras descomponer la primera mezcla (GSE64385):

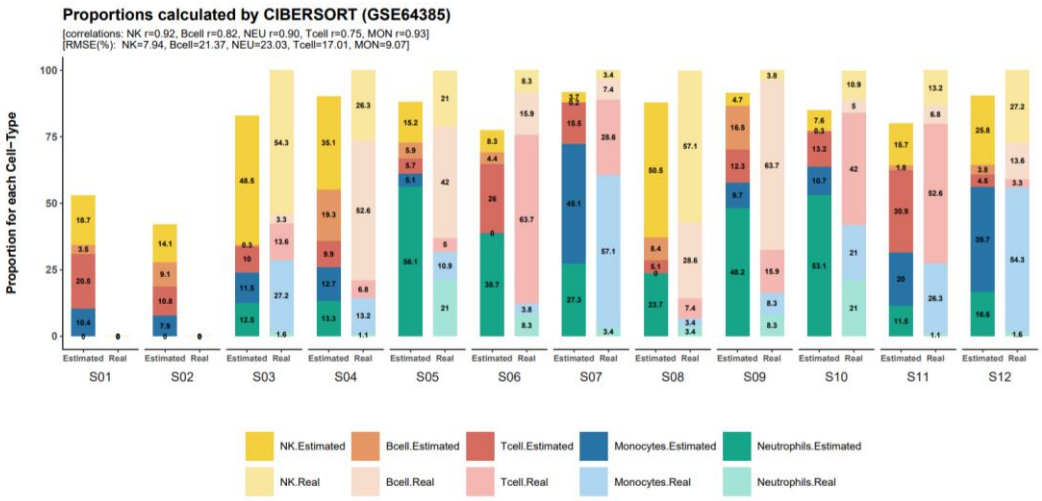


Figura 4.3. Comparación de las proporciones estimadas por CIBERSORT (en la mezcla GSE64385) y las proporciones reales a través de un *bar mixture plot.*

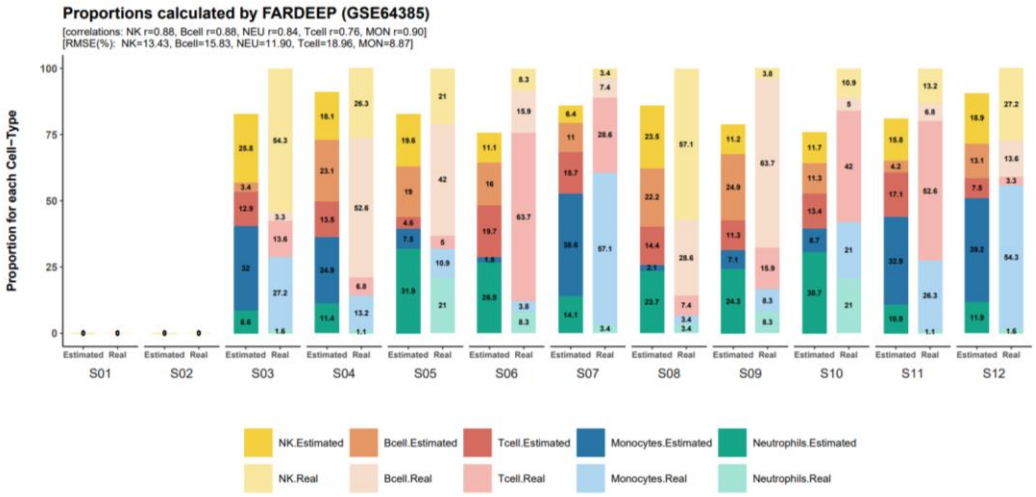


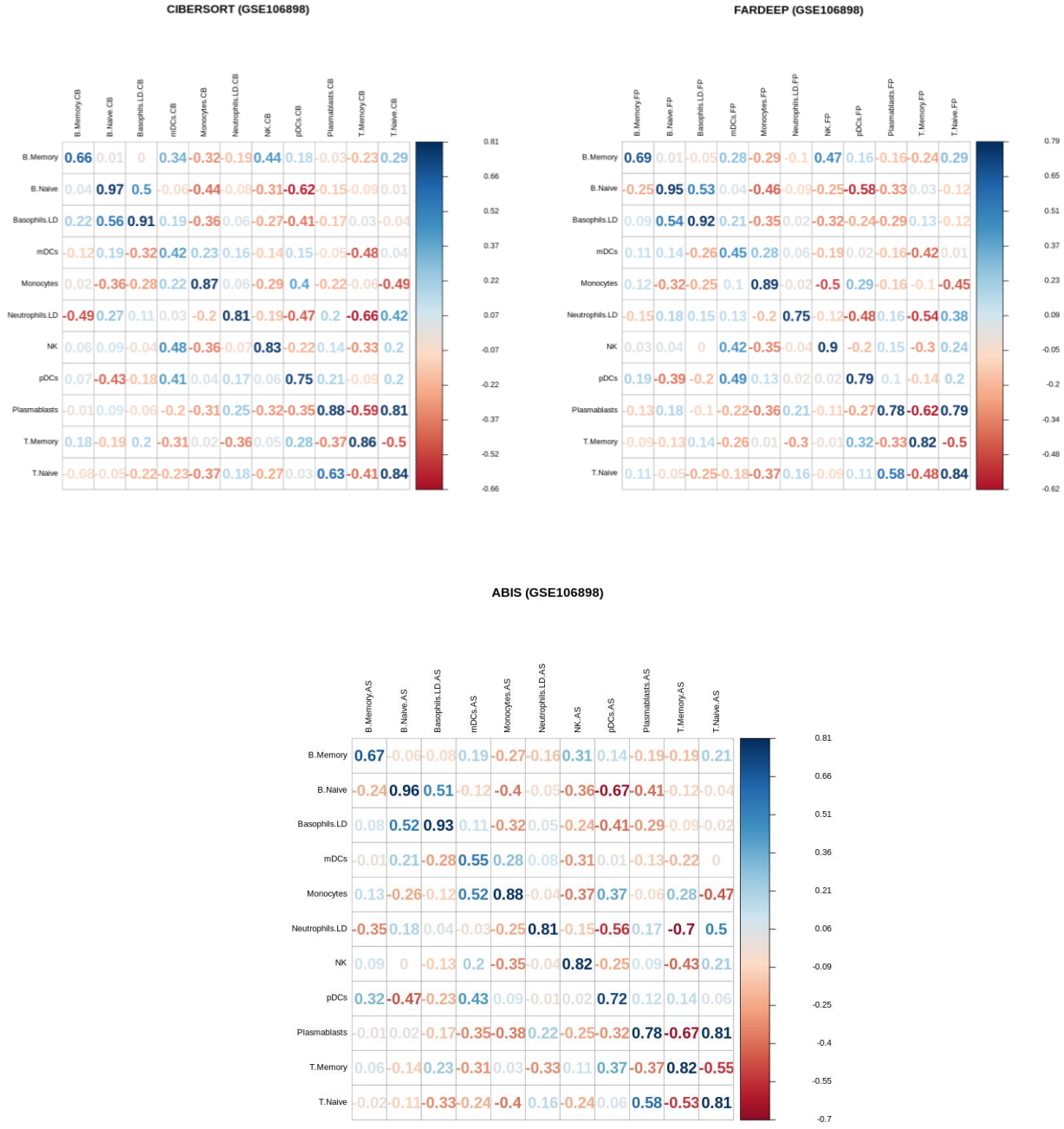
Figura 4.4. Comparación de las proporciones estimadas por FARDEEP (en la mezcla GSE64385) y las proporciones reales a través de un *bar mixture plot*.

Este tipo de gráficos refleja detalladamente la robustez del método ante la presencia de ruido en los datos. Aunque ambos se tratan de métodos supervisados, en los que se ha introducido la misma matriz de firmas (LM22), la principal diferencia entre ellos, como se ha mencionado en el párrafo anterior, es su precisión en el cálculo de las frecuencias relativas en presencia de valores atípicos (*outliers*). Dicha precisión es mayor cuando se utiliza FARDEEP, ya que, como se ha explicado en la sección 3 de este trabajo, su principal característica es la capacidad de eliminar el ruido en los datos y poder descomponer la mezcla sin ningún elemento que pueda alterar los resultados. Los gráficos correspondientes a los métodos no supervisados (DECONICA y LINSEED) y los referidos a la segunda mezcla (GSE20300) están disponibles en los datos suplementarios (**gráficos S17** y **S18**). En la representación de los resultados obtenidos tras la implementación de los métodos no supervisados, referidos a la misma mezcla que los presentes (**gráfico S17** de los datos suplementarios), destaca la monotonía en las proporciones calculadas por DECONICA a través del tamaño de las cajas en las que se plasman dichos valores, ya que éste es igual para cada muestra y tipo celular, mientras que en LINSEED se observa una clara confusión en la estimación de las células T. Los gráficos de barras (*bar mixture plot*) que representan las proporciones de la otra mezcla (GSE20300) muestran, en general, una subestimación en las proporciones. Además, se observa una clara similitud entre los resultados obtenidos por CIBERSORT y los calculados por FARDEEP, lo que no ocurría en el caso anterior por la presencia de ruido en los datos.

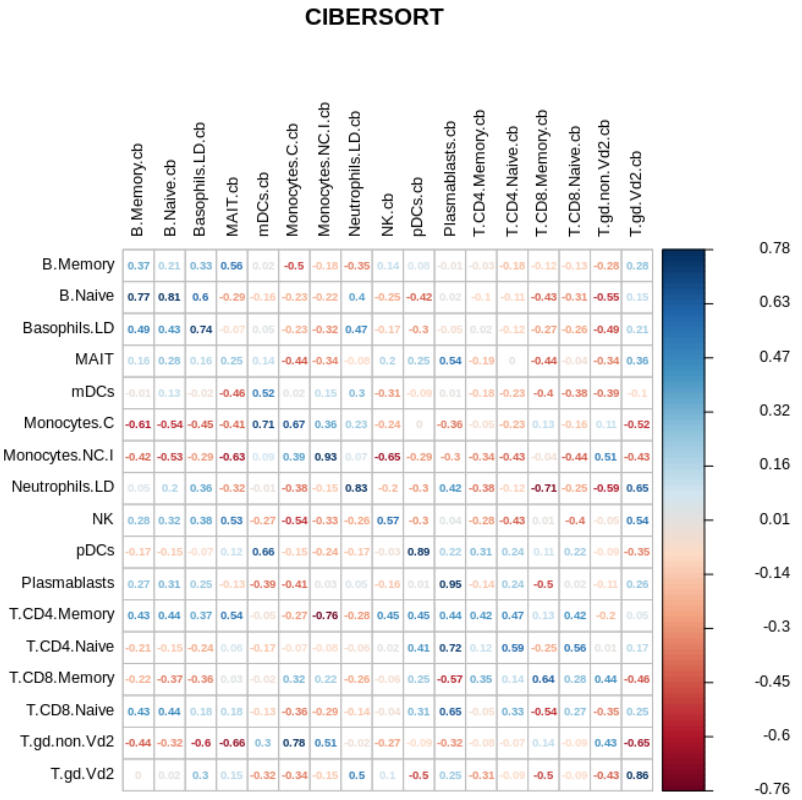
4.2. Comparación de métodos entre datos con señal de expresión de microarray y datos con señal de expresión de RNA-Seq

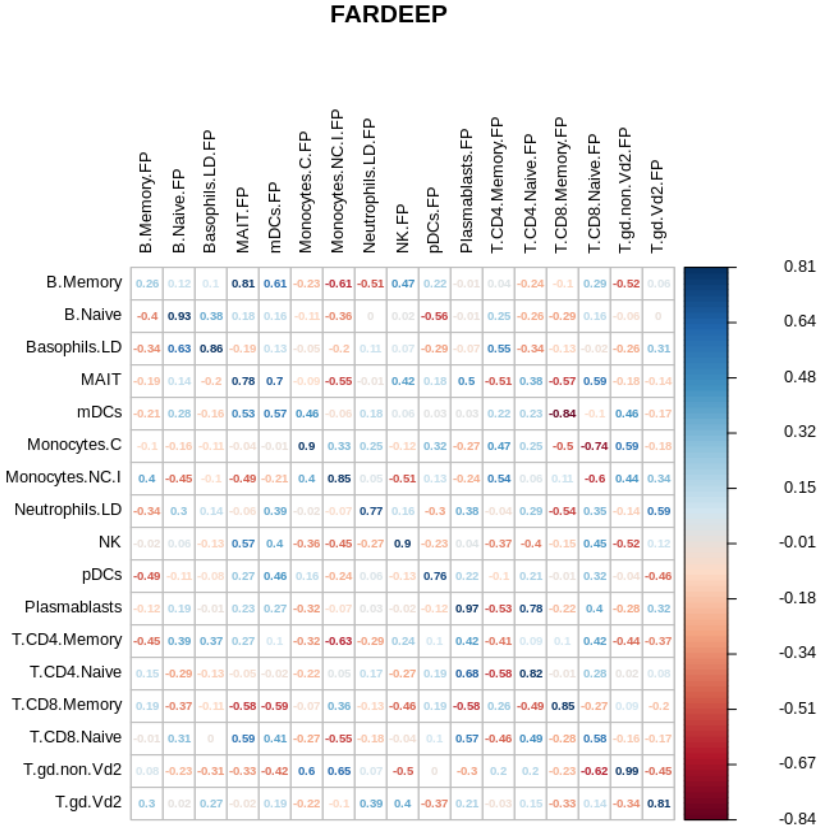
Una vez realizada la comparativa entre métodos descomponiendo mezclas con señal de expresión de microarrays, se ha decidido medir la habilidad de tres métodos supervisados (CIBERSORT, FARDEEP y ABIS), habiendo introducido en ellos tanto el mismo tipo de datos de expresión, como otros más complejos, cuyas muestras fueron analizadas mediante una de las tecnologías de nueva generación: RNA-Seq (*RNA-Sequencing*). Por lo tanto, a diferencia del apartado anterior, el objetivo de esta comparativa es determinar qué método es mejor en función del tipo de señal que contengan los datos. Además, solamente se han implementado tres métodos, todos ellos diseñados para resolver un problema de deconvolución parcial (métodos supervisados), ya que, como el segundo tipo de datos es más complejo, los algoritmos no supervisados no son capaces de converger hacia una solución factible. Por otro lado, también se ha decidido introducir un nuevo método (ABIS), desarrollado con el fin de realizar la deconvolución utilizando datos con señal de expresión de RNA-Seq. No obstante, dicho método también puede utilizarse introduciendo datos cuya señal de expresión fue detectada mediante microarrays. En el primer enfoque analítico, se han realizado gráficos de correlaciones (*corrplot*) para cada conjunto de datos, uno por cada método de deconvolución utilizado:

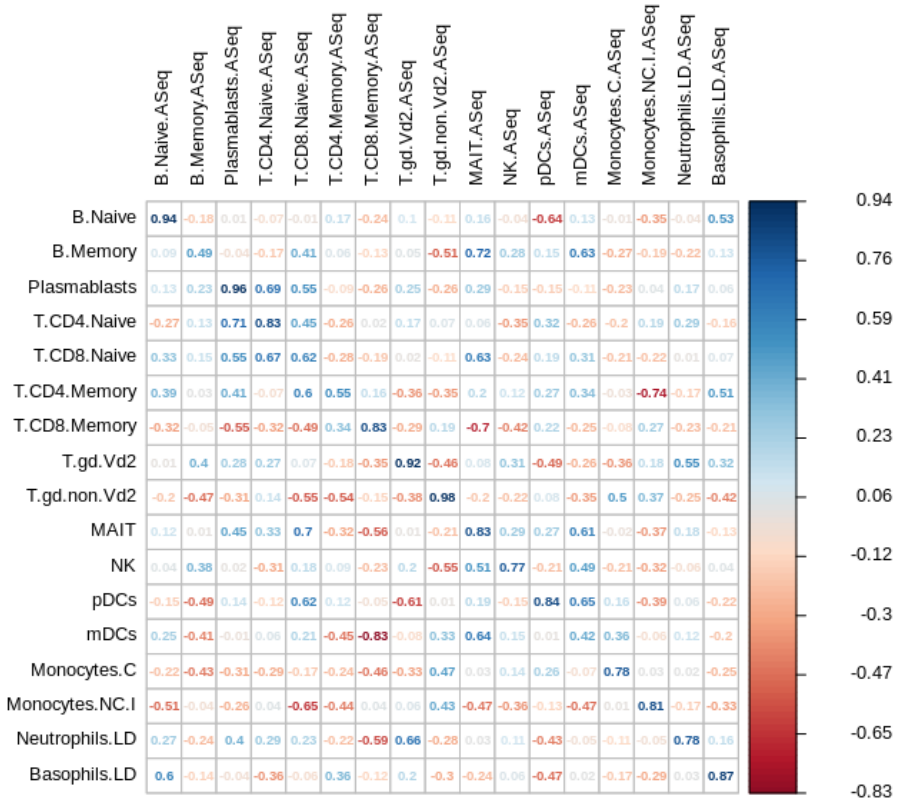
Datos con señal de expresión detectada mediante microarrays



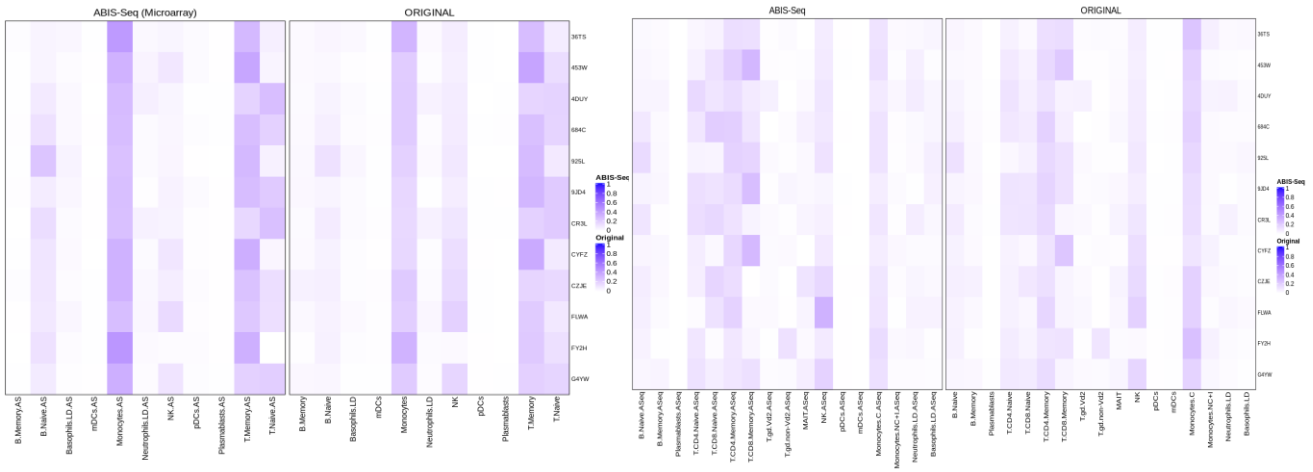
Datos con señal de expresión detectada mediante RNA-Seq







En el primer caso, los resultados obtenidos tras la aplicación de los métodos son muy similares entre sí, especialmente entre CIBERSORT y FARDEEP. Sin embargo, cuando se han utilizado datos con señal de expresión génica de RNA-Seq, se ha producido un decrecimiento en la precisión del cálculo de las proporciones celulares estimadas por CIBERSORT. Habitualmente, este tipo de datos contiene más ruido que el que podría presentar otro con señal de expresión detectada a través de microarrays, lo que explica la diferencia que se observa entre las correlaciones calculadas utilizando las frecuencias estimadas por CIBERSORT y las calculadas con las proporcionadas por FARDEEP, siendo más altas para este último, lo que no ocurría en el primer caso. Además, aunque con FARDEEP también se obtienen buenos resultados, son ligeramente mejores los obtenidos tras la implementación de ABIS. Para complementar la evaluación de los métodos de una manera más visual, también se han reflejado los resultados a través de *heatmaps*, uno para cada método y tipo de dato. En este documento, se mostrarán como ejemplo los gráficos correspondientes al nuevo método implementado, tanto para datos con señal de microarrays como para los que contienen expresión obtenida mediante la técnica de RNA-Seq, y el resto se pueden ver en los datos suplementarios (**gráficos S21** y **S22**):



La intensidad del color de cada celda, que muestran los anteriores gráficos, indica la cantidad de cada tipo celular que contiene una muestra (cuanto más oscuro sea el azul, mayor será la cantidad del tipo celular existente en la mezcla). En este caso, cabe destacar la baja frecuencia de las células en las muestras, en especial en el segundo *heatmap*, donde se aprecian unos tonos azules bastante claros. Sin embargo, como ocurre en la primera comparativa, estos gráficos no revelan demasiada información. Por lo tanto, para un análisis más exhaustivo, se ha realizado un gráfico de firmas (*cell signature plot*) para cada tipo celular, de la misma forma que los expuestos en la sección 4.1 de este trabajo. A continuación, se presentarán los correspondientes a los linfocitos B primitivos (*naive)* para ambos tipos de datos de expresión, aplicados a cada uno de los tres métodos, mientras que el resto se encuentran disponibles en los datos suplementarios (**gráficos S23-S28**):

Datos con señal de expresión detectada mediante microarrays

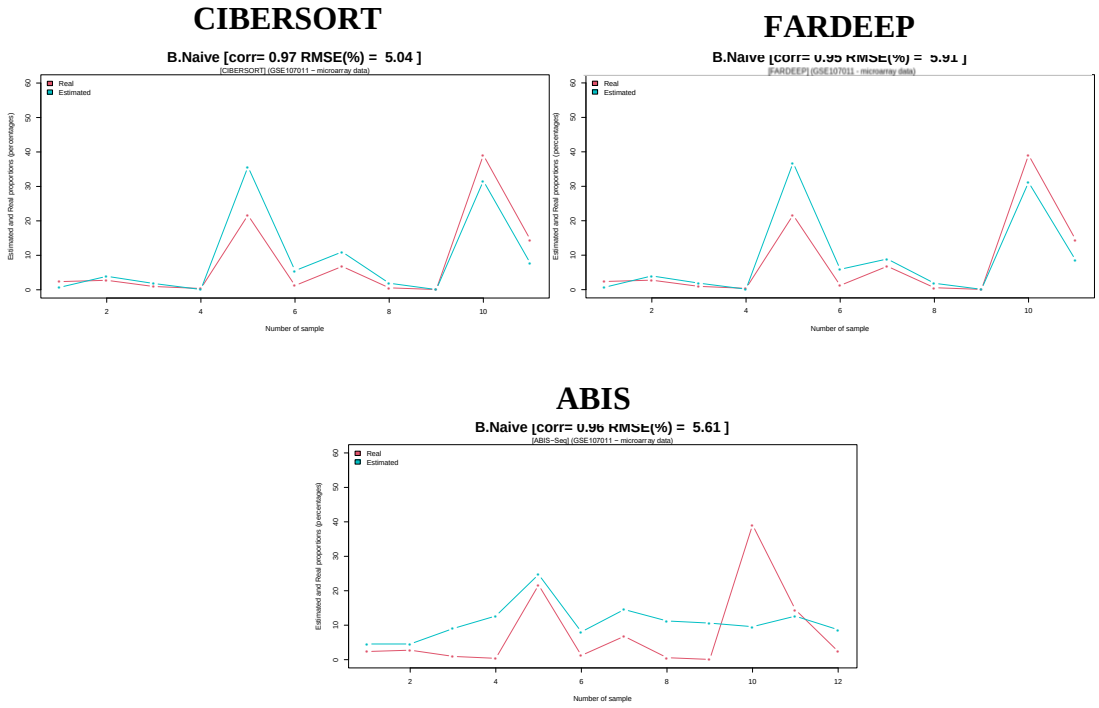


Figura 4.5. Comparación de las proporciones estimadas y las proporciones reales, a través de un cell signature plot. Las frecuencias estimadas fueron calculadas por CIBERSORT, FARDEEP y ABIS, implementando datos con señal de expresión de microarrays.

Datos con señal de expresión detectada mediante RNA-Seq

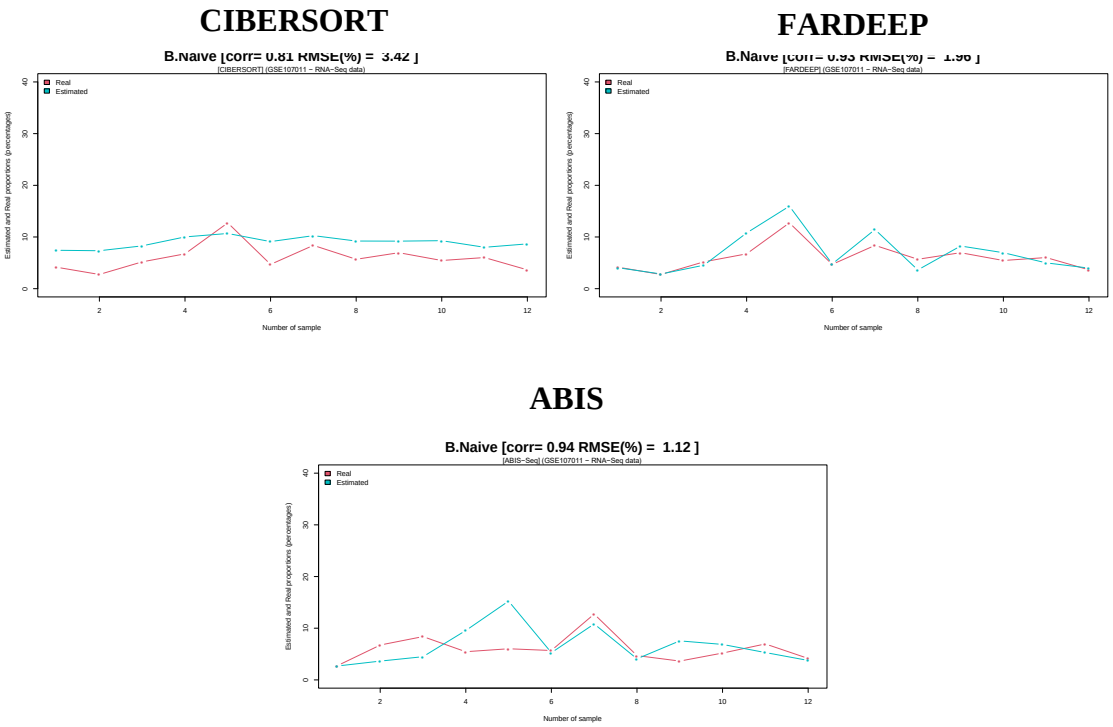


Figura 4.6. Comparación de las proporciones estimadas y las proporciones reales, a través de un cell signature plot. Las frecuencias estimadas fueron calculadas por CIBERSORT, FARDEEP y ABIS, implementando datos con señal de expresión de RNA-Seq.

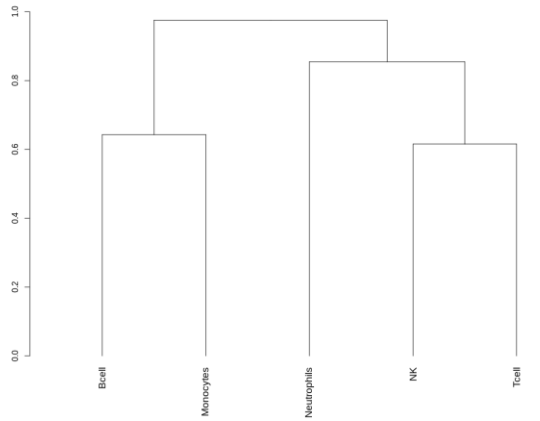
Como se explicó en la anterior sección, estos gráficos muestran las proporciones estimadas y observadas mediante puntos (azules y rojos, respectivamente). En ellos, se observan los valores de las proporciones, tanto los estimados como los reales. En general, tras la aplicación de los métodos en esta mezcla, se obtienen valores de correlación altos y errores pequeños, que es lo que se busca en este estudio. Sin embargo, las frecuencias relativas originales son mucho más bajas que en las mezclas utilizadas anteriormente, por lo que, desde un punto de vista matemático, se esperan errores más pequeños, ya que es lógico pensar que la diferencia entre los valores estimados y reales será más pequeña, y no por ello se debe determinar que la estimación es más precisa que en el caso anterior. Por ejemplo, en el caso de los linfocitos B (**gráficos S21** y **S22** de los datos suplementarios), se obtiene una correlación de 0.66 y un error (RMSE) del 2.18% tras la aplicación de CIBERSORT, utilizando datos con señal de expresión detectada mediante microarrays, mientras que implementando este mismo método con datos de expresión de RNA-Seq, el valor de la correlación disminuye hasta 0.37 y el error tan solo aumenta un 1%, aproximadamente, tomando un valor de 3.76%. Por lo tanto, en el caso de frecuencias relativas pequeñas, es conveniente comprobar las distribuciones de cada tipo celular (estimadas y reales) a través de un gráfico similar a éste, ya que los valores de las correlaciones o de los errores pueden crear confusión si no se interpretan correctamente.

4.3. Análisis de marcadores celulares a través de las matrices de firmas

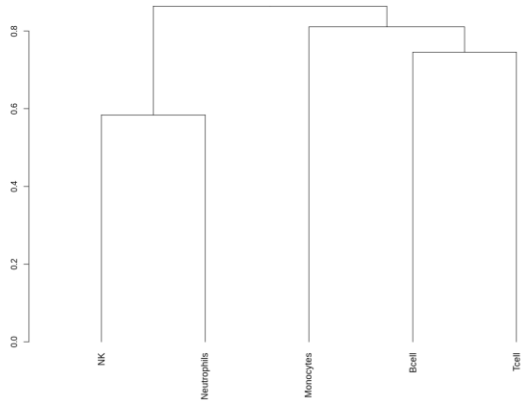
En este tercer y último bloque de resultados, se pretenden estudiar las firmas celulares que sirven para identificar cinco células de sangre periférica (**linfocitos T**, **linfocitos B**, **células NK**, **neutrófilos** y **monocitos**), contenidos en la primera mezcla (descrita en el punto 3.1). En concreto, se desean comparar las matrices de firmas propuestas por los cuatro métodos utilizados en la primera comparativa de este trabajo (CIBERSORT, FARDEEP, DECONICA y LINSEED). Los dos primeros utilizan como matriz de firmas la matriz LM22, una matriz de expresión génica diseñada para caracterizar 22 subtipos de células sanguíneas. Por otro lado, LINSEED Y DECONICA, al tratarse de métodos no supervisados, se encargan de realizar una deconvolución completa, por lo que estiman su propia matriz de firmas

utilizando un único parámetro de entrada: la matriz de mezclas. Para esta parte del estudio, se ha optado por agrupar los datos mediante análisis de *clustering*, considerando cada grupo un tipo celular, utilizando la función `hclust()` y seleccionando como técnica de agrupamiento el método de la mínima varianza 'Ward.D2' (Métodos Jerárquicos de Análisis. Cluster., 2021). La forma en la que se mostrarán los resultados será mediante un diagrama en forma de árbol (**dendrograma**), en el que cada hoja representa un tipo celular. De esta manera, se puede observar la clasificación de los tipos celulares presentes en cada matriz de firmas:

MATRIZ DE FIRMAS ESTIMADA POR DECONICA



MATRIZ DE FIRMAS ESTIMADA POR LINSEED



MATRIZ DE FIRMAS LM22

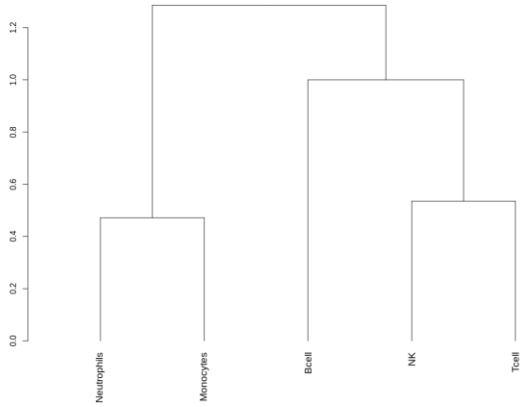


Figura 4.7. Dendrograma que representa la clasificación obtenida de las matrices de firmas estimadas por DECONICA y LINSEED y la matriz LM22. Esta última ha sido implementada en los algoritmos supervisados (CIBERSORT y FARDEEP) cuando se trabaja con datos de expresión con señal de microarrays.

Biológicamente hablando, las células hematopoyéticas se clasifican en dos grandes ramas: la rama mieloide y la rama linfoide. En este caso, dos de los cinco tipos celulares pertenecen a la rama mieloide (**neutrófilos** y **monocitos**) y las tres restantes se clasifican como células de la rama linfoide (**linfocitos B**, **linfocitos T** y **células NK**). Por lo tanto, se espera que la proximidad entre los tipos celulares observada en el agrupamiento se corresponda con la clasificación de dichas células en los anteriores grandes linajes. Sin embargo, esto sólo se cumple en el caso de la matriz LM22, que, como se ha mencionado anteriormente, es una matriz estudiada y analizada con el fin de clasificar poblaciones celulares contenidas en una mezcla de células hematopoyéticas. Una vez observado el agrupamiento de estos cinco tipos celulares, se ha realizado un ***heatmap***, que representa el grado de expresión de los genes mercadores para un determinado tipo celular, escalando los datos de expresión, de manera que los valores se distribuyen en un intervalo de -1 a 1, de modo que los colores varían de tonos azules (para

valores cercanos a -1) hasta tonos rojizos (cuando el valor de la expresión se aproxima a 1). En el caso de ausencia de relación entre ambas variables (gen y tipo celular), la casilla correspondiente en el *heatmap* aparecerá de color blanco:

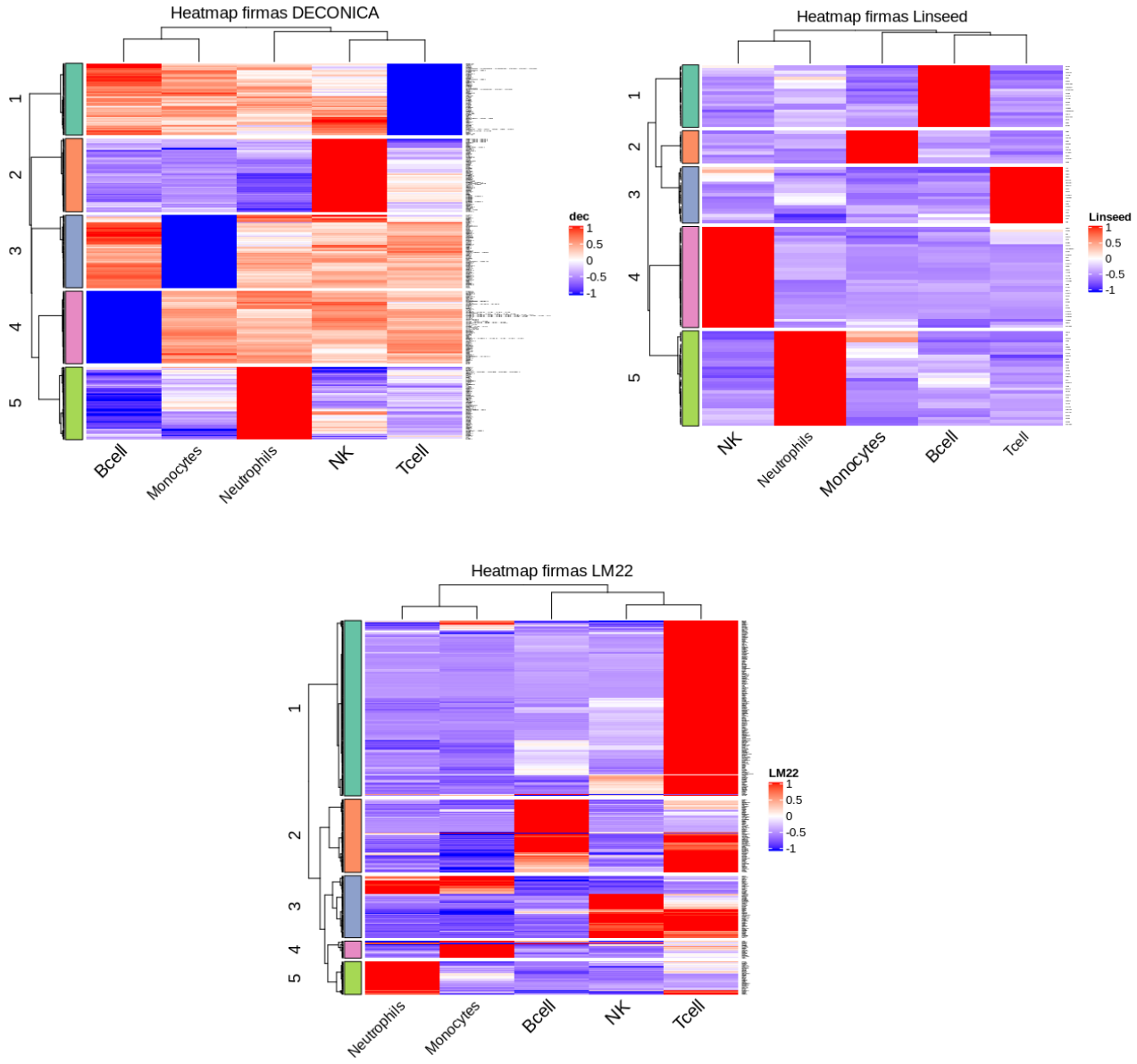


Figura 4.8. Representación de la expresión de los genes marcadores de cada tipo celular mediante un heatmap.

Un detalle importante que destaca en el primer *heatmap* respecto de los otros dos, es que DECONICA considera como genes marcadores aquellos cuya intensidad de expresión es muy baja, es decir, a la ausencia de ese grupo de genes en un determinado tipo celular. Esto se debe a que, para estimar los distintos tipos celulares, se basa en la búsqueda de componentes independientes (o no-gaussianos), o mejor dicho, tipos celulares que presenten distribuciones lo más asimétricas posible (que contengan coeficientes de asimetría alejados de cero), sin tener en cuenta el sentido de dicha asimetría. Por otro lado, DECONICA estima el mismo número de firmas celulares para cada tipo celular (se selecciona dicho número antes de ejecutar el método), lo que puede crear confusiones debido a que, en una mezcla, no es difícil que no contenga el mismo número de genes marcadores para cada tipo celular que se desea estimar. Por otro lado, tanto la matriz de firmas obtenida por LINSEED como la matriz LM22, contienen diferentes números de genes marcadores para cada tipo celular. La diferencia más destacable entre ambas matrices es el número de firmas que caracterizan

los linfocitos T, que coincide con el tipo celular que no es capaz de reconocer el método LINSEED. Por último, se ha realizado una comprobación para saber el número de genes que coinciden entre las tres matrices de firmas, de la que se han obtenido los siguientes resultados:

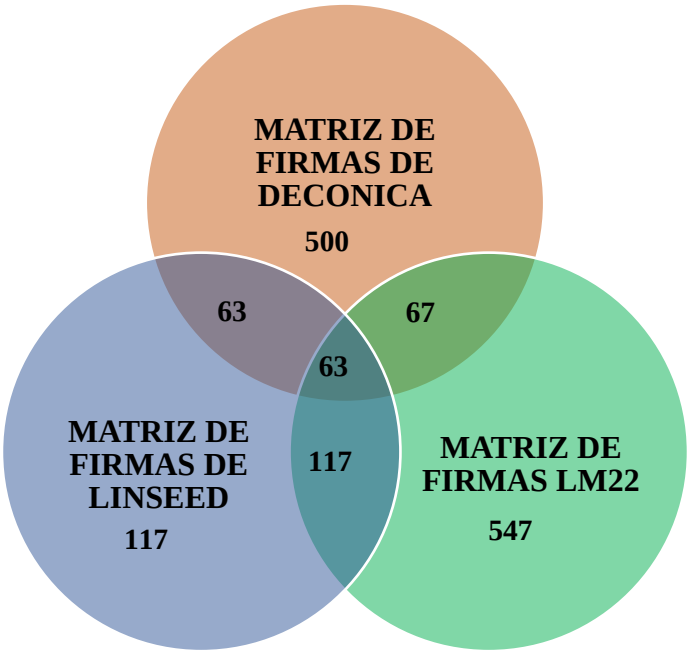


Figura 4.8. Representación del número de genes coincidentes entre las tres matrices estudiadas.

La figura anterior muestra el número de genes que contiene cada matriz de firmas y los que coinciden entre los tres conjuntos de datos estudiados. En primer lugar, la matriz LM22 contiene 117 genes marcadores en común con la estimada por LINSEED, y coincide en 67 firmas celulares con la matriz estimada por DECONICA, mientras que las otras dos matrices (la matriz estimada por DECONICA y la calculada por LINSEED) coinciden en 63 firmas celulares. Un detalle importante es el número de genes que contiene cada matriz; la matriz estimada por DECONICA está formada por 500 genes (100 genes marcadores por cada tipo celular) y la calculada por LINSEED por sólo 117 marcadores celulares. No obstante, esta última coindice en el doble número de genes con la matriz LM22, que se considera como la matriz de referencia, ya que está validada para su uso en métodos de deconvolución (Newman et al., 2015). De hecho, el número de genes en común representa el 100% de las firmas celulares que contiene la matriz de LINSEED, lo que indica que, aunque quizá se necesiten más marcadores celulares para caracterizar correctamente los componentes de la mezcla, LINSEED es más robusto en la selección de los mismos. Además, el número de genes que están contenidos en las tres matrices es 63, que coincide con la intersección entre la matriz de DECONICA y la de LINSEED, debido a que, como es lógico, el número de genes de la matriz de firmas proporcionada por LINSEED es menor que el de la LM22 y, además, todos los genes de la primera matriz aparecen en la segunda, por lo que la intersección de los genes marcadores entre LINSEED y DECONICA es la misma que entre las tres matrices.

5. DISCUSIÓN Y CONCLUSIONES

El estudio de la heterogeneidad celular conlleva diversos beneficios en el conocimiento biológico, como la identificación de genes marcadores, o la determinación de la forma en la influyen los cambios producidos en el organismo en ciertos procesos biológicos o en enfermedades tales como el cáncer. Sin embargo, las técnicas experimentales (*in vitro*) presentan algunas desventajas como la limitación a marcadores fenotípicos conocidos. Por ello, se han desarrollado técnicas de carácter computacional (*in silico*) encargados de realizar la deconvolución, que son capaces de descomponer mezclas celulares e identificar biomarcadores desconocidos. En estas técnicas, se distinguen dos tipos; los no supervisados (realizan la deconvolución completa) y los métodos supervisados (realizan la deconvolución parcial) que, como ya se explicó en el punto 1.1.4 de este trabajo, los algoritmos supervisados necesitan como

parámetro de entrada una matriz de firmas para su ejecución, en la que aparece la expresión de los genes marcadores en los tipos celulares de estudio. Para la obtención de dicha matriz de expresión, se utilizan técnicas encargadas de la detección de la expresión génica, tales como la tecnología de microarrays o una de las dos tecnologías basadas en la secuenciación de moléculas de RNA: RNA-Seq (*RNA-Sequencing*) y scRNA-Seq (*Single Cell RNA-Sequencing*). La técnica de microarrays es la más antigua, y por ello la más conocidas y estudiada, pero como menciona Gianni Monaco (Monaco et al., 2019), presenta inconvenientes relacionados con la resolución de la señal de expresión génica. Actualmente, la técnica más utilizada es RNA-Seq, que es más potente que la tecnología de microarrays en cuanto a la identificación de nuevas isoformas y el conocimiento de nuevas características sin la necesidad de un conocimiento previo (Métodos y Flujos de Trabajo de RNA-Seq, 2021), aunque los análisis computacionales resulten más complejos por el gran volumen de datos. Por otro lado, la técnica scRNA-Seq se basa en la secuenciación por la misma metodología que RNA-Seq, pero con un previo aislamiento de las células, y es la más eficiente en cuanto a la caracterización de nuevos biomarcadores y la detección de la expresión génica, pero el coste económico requerido para el aislamiento de las células individuales es muy alto, además de un aumento considerable en la complejidad de los análisis utilizando datos de este tipo, ya que cuantifica mediante conteos de transcritos la expresión génica en células únicas. Para este trabajo, se intentó implementar un nuevo método diseñado para caracterizar los tipos celulares mediante un conjunto de datos de expresión de referencia analizado con scRNA-Seq, denominado *CellR* (Doostparast Torshizi et al., 2021), pero no fue posible incluirlo en ningún punto del estudio ya que no se encontraron los datos de las proporciones celulares reales, para poder así validar la precisión del método. Por ello, durante este trabajo únicamente se han utilizado datos de expresión génica detectados mediante microarrays y RNA-Seq, la cual presenta cada vez más datos de células inmunes para la validación de los métodos (Monaco et al., 2019). Durante este Trabajo de Fin de Grado, se han analizado diversos métodos de deconvolución a través del *software* estadístico R, utilizando datos de expresión de microarrays y datos de expresión de RNA-Seq. En la primera comparativa, se aplicaron cuatro métodos, dos de ellos supervisados (**CIBERSORT** y **FARDEEP**) y otros dos no supervisados (**DECONICA** y **LINSEED**), para dos matrices de mezclas de células sanguíneas (**GSE64385** y **GSE20300**). Los *corrplots* obtenidos muestran que se obtienen **mejores estimaciones** para la primera mezcla, siendo mejores **en los métodos supervisados** (CIBERSORT y FARDEEP), aunque en principio, también se obtuvieron buenos resultados para DECONICA, mientras que el gráfico correspondiente al método LINSEED muestra correlaciones muy altas para cuatro tipos celulares pero, en cambio, no es capaz de reconocer los linfocitos T, en la primera mezcla, ni los monocitos en la segunda, aunque en el último caso todos los métodos se alejan de las proporciones reales en alguno de los tipos celulares. Posteriormente, se realizó un *heatmap* para cada método en cada una de las dos mezclas. En los que representan las proporciones del primer conjunto de datos, se observa con claridad que FARDEEP es más robusto al ruido (en nuestro caso el ruido está representado por las células tumorales), ya que las dos primeras muestras sólo contienen dos tipos de células cancerígenas mezcladas (muestras puras), por lo que las frecuencias relativas de las células sanguíneas en dichas muestras deberían ser nulas, lo que sólo se consigue con FARDEEP. Por otro lado, los gráficos diseñados y conocimos como *Cell Signature Plot* y *Bar Mixture Plot* revelan más información contenida en los datos. Por una parte, los valores de RMSE (*Root Mean Square Error*) más pequeños se corresponden con el método LINSEED, aunque en la primera mezcla el valor se dispara para el caso de los linfocitos T. Además, las conclusiones sobre el buen funcionamiento de DECONICA cambian radicalmente, debido a que, aunque los errores no difieran mucho respecto a los calculados para CIBERSORT y FARDEEP, presenta una monotonía en las frecuencias relativas estimadas. En la segunda comparativa, se han aplicado tres métodos supervisados (**CIBERSORT**, **FARDEEP** y **ABIS**) en una misma mezcla pero en dos conjuntos de datos distintos: uno con señal de expresión de microarrays (**GSE106898**) y otro de RNA-Seq (**GSE107011**). En los *corrplots* del primer caso (con señal de expresión de microarrays) se tienen resultados parecidos en los tres conjuntos de datos. Sin embargo, cuando se aplican los mismos algoritmos en datos de expresión de RNA-Seq, disminuyen las correlaciones entre los valores estimados por CIBERSORT y las proporciones reales. Los datos de expresión de RNA-Seq presentan más complejidad, por el gran volumen de datos, que los datos de expresión de microarrays, lo que podría explicar este decrecimiento en la precisión de dicho método. Un detalle importante es la obtención de valores de RMSE (raíz del error cuadrático medio) muy pequeños en comparación con los calculados tras deconvolucionar las dos mezclas utilizadas en la primera comparativa. Este hecho puede llegar a crear confusión en cuanto a la precisión del método, ya que la obtención de valores pequeños en esta medida de error no conlleva una mejora en la precisión de la estimación de las frecuencias relativas, sino que este resultado está ligado a

la presencia de proporciones más pequeñas y, por lo tanto, matemáticamente, la diferencia entre valores más pequeños, será menor. Un ejemplo muy claro de esto es el caso de los eosinófilos, cuyas proporciones reales se distribuyen en un intervalo de 1-3% y un 55-70%, respectivamente. Ahora bien, las correlaciones calculadas para estos tipos celulares en CIBERSORT, FARDEEP y DECONICA tienen aproximadamente un valor de 0.3 para los eosinófilos, mientras que el RMSE es menor que en otros tipos celulares con correlaciones mayores de 0.7 (como por ejemplo los neutrófilos, cuyas frecuencias relativas reales toman valores entre 55% y 70%), lo que explica que no siempre se pueden comparar los valores del RMSE entre distintos conjuntos de datos y/o tipos celulares. En cuanto al cálculo de proporciones, se han obtenido las siguientes conclusiones sobre los métodos de deconvolución:

1. Los métodos **más eficaces** en la deconvolución de muestras transcriptómicas son **CIBERSORT** y **FARDEEP**, ya que son con los que se obtienen correlaciones altas, valores de RMSE pequeños y, además, las distribuciones de los tipos celulares se asemejan a las de los tipos celulares conocidos. Sin embargo, presentan un inconveniente importante a la hora de ejecutar el método correctamente. Dicha desventaja está relacionada con la plataforma con la que se han analizado las muestras, porque para poder considerar un resultado válido, la expresión génica de las muestras en la mezcla y la expresión de los genes marcadores en la matriz de firmas, deben haber sido procesadas mediante la misma plataforma (por ejemplo, en nuestro primer caso se analizó con *Affymetrix HGU133A* y la ampliación de esta, *HGU133 Plus 2.0*).

2. Dentro los dos mejores métodos, se ha comprobado que **FARDEEP** es más **robusto** a la presencia de **ruido** en los datos, de manera que en aquellos conjuntos de datos que contengan elementos que puedan alterar los resultados (como en el primer caso células tumorales o datos de expresión complejos como los obtenidos por métodos de secuenciación), FARDEEP es más preciso en el cálculo de las proporciones celulares que CIBERSORT. En el caso en el que no existan elementos de este tipo, ambos métodos proporcionan resultados similares.

3. El cálculo de la correlación de Pearson y del RMSE utilizando las frecuencias relativas calculadas por **DECONICA** puede crear confusiones sobre la precisión de dicho método. Los coeficientes de la correlación obtenidos son valores altos y, además, los valores del RMSE no se alejan de los obtenidos para los métodos CIBERSORT y FARDEEP. No obstante, las **distribuciones** de los **tipos celulares se distribuyen en torno a un valor medio**, de manera que **no** presenta **variación** entre los valores correspondientes a cada muestra y tipo celular, por lo que no se aconseja el uso de éste para la deconvolución de datos biológicos.

4. **LINSEED** es un método no supervisado, es decir, únicamente necesita una matriz de mezclas para su ejecución, y es **robusto** a la presencia de **ruido** en los datos. Por ambas razones, este podría ser una buena elección para deconvolucionar una mezcla transcriptómica cuando no se tiene una matriz de firmas. Sin embargo, un **gran número de genes puede influir** en la **precisión** del mismo, ya que, como se basa en el cálculo de correlaciones entre genes de acuerdo a una cierta significación, el número de correlaciones calculadas es muy grande y los análisis pueden resultar muy costosos.

5. **Para** la **implementación** del método **ABIS**, se **necesitan** conocer las **proporciones reales** presentes en la mezcla que se desea descomponer, debido a que en el cálculo del factor de escala se utilizan estos valores. Por lo tanto, si el objetivo es simplemente la aplicación del método y no su validación, no siempre se pueden conocer las frecuencias originales y por lo tanto no se podría utilizar este método.

6. Para una **validación** correcta de la **precisión** de los métodos de deconvolución se deben tener en cuenta varios aspectos. Primero, el valor de la **correlación** o una medida similar a ésta, que permita medir la similitud en la tendencia entre los valores estimados y reales. En segundo lugar, una medida de error (en nuestro caso **RMSE**), que sea capaz de calcular la diferencia entre los mismos valores. Después, desde un punto de vista visual, **comprobar** la **distribución** de las **puntuaciones estimadas y conocidas** y, por último, tener en cuenta los propios **valores** de las **proporciones**, ya que es lógico pensar que la distancia entre valores más pequeños será menor, y no por ello indica una estimación más precisa.

Finalmente, se ha realizado un estudio sobre los genes marcadores de tres matrices de firmas: la matriz LM22 y las dos matrices estimadas por los métodos no supervisados (DECONICA y LINSEED). Entre estas tres matrices, la **LM22** se considera la matriz de **referencia** ya que, como se ha mencionado anteriormente, ha sido estudiada y analizada para su uso en métodos de deconvolución supervisados, con el fin de caracterizar poblaciones celulares conocidas. Para comparar las matrices estimadas por los métodos no supervisados, se ha estudiado la intersección entre los genes marcadores de ambas matrices con dicha matriz de referencia, y también los genes coincidentes entre ellas. En este análisis, se ha observado que todos los genes marcadores contenidos en la matriz de firmas calculada por LINSEED, se encuentran en la matriz LM22 (117) genes, mientras que la matriz de firmas proporcionada por el método DECONICA sólo coinciden en 67 genes con la misma matriz de referencia. Además, al calcular la intersección entre las tres matrices, se obtiene un número total de 63 genes, que coincide con la intersección entre las matrices de LINSEED y DECONICA, por contener la matriz LM22 el 100% de los genes marcadores considerados en la matriz calculada por LINSEED. Este concepto, nos indica una mayor robustez en la selección de las firmas celulares por el método LINSEED respecto a la elección de DECONICA, pese a tener un mayor número de genes marcadores para poder caracterizar los tipos celulares, cuya razón podría estar relacionada con la búsqueda de distribuciones asimétricas (ya sean positivas o negativas), debido a que, desde un punto de vista biológico, el hecho de caracterizar un tipo celular por la ausencia de un conjunto de genes carece de sentido.

Como conclusión general, ninguno de los métodos estudiados se adapta a cualquier circunstancia, por lo que, en función del estado de los datos, se debe seleccionar uno u otro. Además, en un futuro, conviene diseñar nuevos métodos de deconvolución que mejoren en alguna de las carencias que presentan los cinco métodos implementados durante este trabajo. Igualmente, también podrían desarrollarse otros algoritmos de deconvolución con el fin mejorar la descomposición de datos de expresión de las tecnologías de nueva generación (RNA-Seq y scRNA-Seq), ya que, debido a la complejidad que supone la gran profundidad de estos datos, los análisis son mucho más costosos y, por lo tanto, los resultados tienden a alejarse de la realidad.

6. BIBLIOGRAFÍA (Artículos y Libros)

Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., & Clark, H. F. (2009). **Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus**. *PLoS ONE*, *4*(7). https://doi.org/10.1371/journal.pone.0006098

Arranz Rodríguez, A. A., & Tabernero, A. (2019). **Reducción del movimiento de la cámara en una fotografía**.

Awad, M., & Khanna, R. (2015). **Support Vector Regression. In Efficient Learning Machines** (pp. 67–80). *Apress*. https://doi.org/10.1007/978-1-4302-5990-9_4

Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W. H., & 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

Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M., & Alizadeh, A. A. (2018). **Profiling tumor infiltrating immune cells with CIBERSORT**. In Methods in Molecular Biology (Vol. 1711, pp. 243–259). *Humana Press Inc*. https://doi.org/10.1007/978-1-4939-7493-1_12

Cherry, E. C. (1953). **Some Experiments on the Recognition of Speech, with One and with Two Ears**. *Journal of the Acoustical Society of America*, 25(5), 975–979. https://doi.org/10.1121/1.1907229

Czerwińska, U. (2018). **Unsupervised deconvolution of bulk omics profiles: Methodology and application to characterize the immune landscape in tumors**. Doctoral Thesis. University of Paris

Dantzig, G. B. (1990). **Origins of the simplex method. In A history of scientific computing** (pp. 141–151). *ACM*. https://doi.org/10.1145/87252.88081

Doostparast Torshizi, A., Duan, J., & Wang, K. (2021). **A computational method for direct imputation of cell type-specific expression profiles and cellular compositions from bulk-tissue RNA-Seq in brain disorders**. *NAR Genomics and Bioinformatics*, 3(2), lqab056, https://doi.org/10.1093/nargab/lqab056

Gutiérrez, L. (2020) **Statistic analysis of large-scale complex transcriptomic data of human samples and use of a deconvolution method to identify specific cell types**. Tutores: J.M. Sánchez Santos & J. De Las Rivas.

Hao, Y., Yan, M., Heath, B. R., Lei, Y. L., & Xie, Y. (2019). **Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares**. *PLoS Computational Biology*, 15(5). https://doi.org/10.1371/journal.pcbi.1006976

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MaCdonald, J., Obenchain, V., Oleš, A. K., … Morgan, M. (2015). **Orchestrating high-throughput genomic analysis with Bioconductor**. *Nature Methods*, 12(2), 115–121. https://doi.org/10.1038/nmeth.3252

Hwang, B., Lee, J. H., & 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

Hyvarinen, A. (1999). **Fast ICA for noisy data using gaussian moments**. *Proceedings - IEEE International Symposium on Circuits and Systems*, *5*. https://doi.org/10.1109/iscas.1999.777510

López de Heredia, U. (2016). **Las técnicas de secuenciación masiva en el estudio de la diversidad biológica**. *Munibe Ciencias Naturales*, *64*. https://doi.org/10.21630/mcn.2016.64.07

Lu, P., Nakorchevskiy, A., & Marcotte, E. M. (2003). **Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations**. *Proceedings of the National Academy of Sciences of the U.S.A.*, 100(18), 10370–10375. https://doi.org/10.1073/pnas.1832361100

Meisenberg, G. & Simmons, W.H. (2016). **Principles of Medical Biochemistry**. *Elsevier, 4th Edition*. https://www.elsevier.ca/ca/product.jsp?isbn=9780323296168

Métodos Jerárquicos de Análisis. Cluster. (v. 2021). https://www.ugr.es/~gallardo/pdf/cluster-3.pdf

Métodos y flujos de trabajo de RNA-Seq. (2021). https://emea.illumina.com/techniques/sequencing/rna-sequencing.html

Monaco, G., Lee, B., Xu, W., Mustafah, S., Hwang, Y. Y., Carré, C., Burdin, N., Visan, L., Ceccarelli, M., Poidinger, M., Zippelius, A., Pedro de Magalhães, J., & Larbi, A. (2019). **RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types.** *Cell Reports*, *26*(6), 1627-1640.e7. https://doi.org/10.1016/j.celrep.2019.01.041

Narrandes, S., & Xu, W. (2018). **Gene expression detection assay for cancer clinical use.** *Journal of Cancer*, 9(13), 2249-65 https://doi.org/10.7150/jca.24744

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., & Alizadeh, A. A. (2015). **Robust enumeration of cell subsets from tissue expression profiles**. *Nature Methods*, *12*(5), 453–457. https://doi.org/10.1038/nmeth.3337

Papalexi, E., & Satija, R. (2018). **Single-cell RNA sequencing to explore immune cell heterogeneity**. *Nature Reviews Immunology*, 18(1), 35–45. https://doi.org/10.1038/nri.2017.76

Romero Campero, F. J. (2019a). **Estudios Transcriptómicos Masivos: Análisis de Microarrays Técnicas Ómicas y Bioinformática** 3º Biomedicina Básica y Experimental. http://www.cs.us.es/~fran/

Romero Campero, F. J. (2019b). **Estudios Transcriptómicos Masivos basados en Secuenciación de Altas Prestaciones: RNA-seq**. http://www.cs.us.es/~fran/

Salazar Montes, AM. (2016). **Principios de la biología molecular** https://www.mheducation.es/principios-de-biologia-molecular-9786071513663-spain

Sean, D., & Meltzer, P. S. (2007). **GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor.** *Bioinformatics*, 23(14), 1846–1847. https://doi.org/10.1093/bioinformatics/btm254

Sharp, T. (2020). **An Introduction to Support Vector Regression (SVR)**. https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2

Shen-Orr, S.S., Tibshirani, R., Khatri, P., Bodian, D.L., Staedtler, F., Perry, N.M., Hastie, T., Sarwal, M.M., Davis, M.M., & Butte, A. J. (2010). **Cell type-specific gene expression differences in complex tissues**. *Nature Methods*, *7*(4), 287-9. https://doi.org/10.1038/nmeth.1439

Tecnología de microarrays (chips de ADN o ARN) | NHGRI. (2021). https://www.genome.gov/es/genetics-glossary/Tecnologia-de-microarrays

Vladimir Kiselev. (2019). **Introduction to single-cell RNA-seq | Analysis of single cell RNA-seq data**. https://scrnaseq-course.cog.sanger.ac.uk/website/introduction-to-single-cell-rna-seq.html

Xu, Q., Yan, M., Huang, C., Xiong, J., Huang, Q., & Yao, Y. (2017). **Exploring Outliers in Crowdsourced Ranking for QoE**. Proceedings of the 25th ACM International Conference on Multimedi*a*, *17*. https://doi.org/10.1145/3123266.3123267

Zaitsev, K., Bambouskova, M., Swain, A., & Artyomov, M. N. (2019). **Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures**. *Nature Communications*, *10*(1), 1–16. https://doi.org/10.1038/s41467-019-09990-5

7. OTRAS REFERENCIAS (URLS)

Alcances de la Inmunohistoquímica en el estudio de los tejidos. (2015).
https://www.grupogamma.com/alcanes-inmunohistoquimica/

Deconvolution | Definition of deconvolution by Oxford Dictionary. (v. 2021).
https://www.lexico.com/definition/deconvolution

Deoxyribonucleic Acid (DNA) Fact Sheet. (v. 2021).
https://www.genome.gov/about-genomics/fact-sheets/Deoxyribonucleic-Acid-Fact-Sheet

Diccionario de cáncer del NCI - Instituto Nacional del Cáncer. (v. 2021).
https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer

HGNC Database, HUGO Gene Nomenclature Committee (HGNC). (v. 2021).
https://www.genenames.org/

National Human Genome Research Institute Home | NHGRI. (v. 2021).
https://www.genome.gov/

Proteína | NHGRI. (v. 2021).
https://www.genome.gov/es/genetics-glossary/Proteina

R: El Proyecto R para Computación Estadística. (2021).
https://www.r-project.org/

Ribonucleic Acid (RNA). (2021).
https://www.genome.gov/genetics-glossary/RNA-Ribonucleic-Acid

Package - MASS in CRAN (v. 2021).
https://cran.r-project.org/web/packages/MASS/index.html

GitHub - FARDEEP: Fast and Robust Deconvolution using Least Trimmed Squares (v. 2021).
https://github.com/cran/FARDEEP

GitHub - LinSeed: LINear Subspace identification for gene Expresion Deconvolution (v. 2021).
https://github.com/ctlab/LinSeed

GitHub - DeconICA: Deconvolución del transcriptoma mediante análisis de componentes inmunitarios. (v. 2021).
https://github.com/UrszulaCzerwinska/DeconICA

8. SUMMARY

The analysis of the transcriptome represents a fundamental part of the study of cell heterogeneity, due to the advantages it provides. Specifically, it makes it possible to identify how cells change according to certain processes produced in the organism or due to certain diseases. This type of study is based on the analysis of gene expression, which is obtained from different techniques that detect gene expression using different methodologies. In this work, we will use expression data obtained from microarray technology, based on the hybridization of DNA fragments, and gene expression obtained from another technique based on the sequencing of RNA molecules, called RNA-Seq (RNA-Sequencing).

Experimental methodologies designed to study cell variability (immunohistochemistry and flow cytometry) are limited to known phenotypic markers, in addition to other problems that can occur in the procedures carried out by these techniques. For all these reasons, deconvolution methods have emerged, which are defined as computational tools able to decomposing a mixture of different cell types into its constituent elements, calculating their proportion and, in some cases, also the expression signal of the factors that allow the data set to be decomposed. Let *n*, *m* and *c* be the number of genes, samples and cell types, respectively, transcriptomic data deconvolution is defined as:

$$T_{nxm} = C_{nxc} * P_{cxm}$$

Where T_{nxm} is the mixture matrix, C_{nxc} is the signature matrix (matrix of gene markers expression) and P_{cxm} is the proportion matrix, which contain the relative frequencies of cell types in the mixed samples. For the deconvolution process to be successful, the P matrix has to fulfil two properties: the columns (samples) must sum to one $\sum_{j=1}^{m} P_{kj} = 1, \quad \forall\, k \in [1, \dots, c]$ and each element of the matrix P must take a value between zero and one $\sum_{j=1}^{m} P_{kj} = 1, \quad \forall\, k \in [1, \dots, c]\ \forall$ j∈[1,...,m]). There are two types of deconvolution depending on the elements to be estimated. If the aim of the method is to estimate one of the two matrices (C or P), then we are talking about partial deconvolution (supervised method) and requires, in addition to a mixture matrix, another remaining matrix (C or P). However, if the method is able to estimate both, then it is a complete deconvolution, and only requires a mixture matrix (unsupervised method).

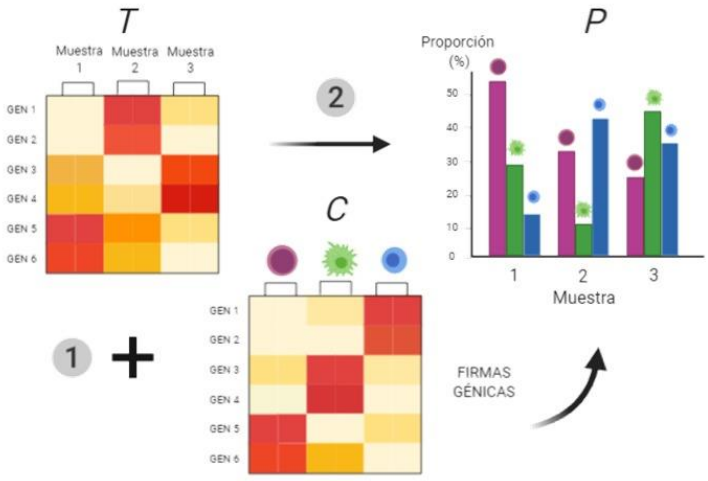


Figure 1. Type of deconvolution. 1) Partial deconvolution 2) Complete deconvolution

DATA

The following table contain the key characteristics of the mixture datasets used in this work:

Table 1. Summary of dataset used in this work.

GEO Accession	Platform of gene expression	Number of samples	Number of genes	Biological Source	Cell Types	References
GSE64385	Microarray HGU133 Plus 2.0 - *Affymetrix*	12	54675	PBMCs and HCT116 cells	5	(Becht et al., 2016)
GSE20300	Microarray HGU133 Plus 2.0 - *Affymetrix*	24	54675	Peripheral blood	4	(Shen-Orr et al., 2010)
GSE107011	RNA-Seq HiSeq 2000- *Illumina*	13	17487	PBMCs	17	(Monaco et al., 2019)
GSE106898	Microarray HumanHT-12 V4.0 - *Illumina*	13	17487	PBMCs	11	(Monaco et al., 2019)

In supervised methods, are required a mixture matrix and signature matrix for execution. The signature matrices used are the LM22 expression matrix, consisting of 22 peripheral blood cell subtypes (columns) defined by unique signatures (rows), whose values were generated using the *Affymetrix HGU133A* platform for microarray signal data, and the expression matrices "sigmatrixMicro.txt" and "sigmatrixRNAseq.txt", which also serve to characterise peripheral blood cell types, but have been generated by other platforms of gene expression: *HumanHT-12 V4.0 - Illumina* for the microarray gene expression matrix and *RNA-Seq HiSeq 2000 Illumina* for the RNA-Seq gene expression signal matrix.

DECONCOLUTION METHODS

DECONICA (Deconvolution of transcriptome through Immune Component Analysis)

This is an unsupervised deconvolution method that solves a complete deconvolution problem and therefore only requires the mixture matrix. For the estimation of cell types, it is based on FastICA, which uses a multivariate technique (ICA: Independent Component Analysis) whose objective is find uncorrelated latent variables, which present a non-Gaussian distribution (the skewness and kurtosis coefficients must be far from zero). Hence, let *n* be the number of observable variables (genes), *m* the number of samples, and *k* the components into which we want to decompose the data, we propose a model of the form:

$$T_{mxn} = A_{mxk} C_{kxn}$$

The aim is to obtain a matrix *A* whose values maximize the skewness and kurtosis statistics. When running this algorithm in R, the number of marker genes for each cell type must be selected previously, and the result obtained is a score matrix, whose values must be transformed to relative numeric values.

LINSEED (Linear Subspace identification for gene Expression Deconvolution)

This method, as the previous one as solves a complete deconvolution (is an unsupervised method). To perform the deconvolution, it is based on the simplex algorithm, in which the vertices (corners) are the marker genes and cell types, which represent the optimal points. In linear programming, the problem is proposed as follows:

Max (min) z: X= αH

$$\alpha \geq 0 \;\; y \sum_{j}^{k} \alpha_{j}^{i} = 1, \qquad \forall i = 1, \ldots, m$$

Where *X* is the expression of the genes in each cell type, *H* the cell type proportions row-normalised, and α is a non-negative coefficient, which must sum to one per sample. During the procedure of this technique in R, the following steps are carried out:

1. The 'linseed' type object is created.

2. The *x* genes most mutually correlated are selected, using Spearman's correlation.

3. *k* cell types are estimated based on a SVD (Singular Value Decomposition) representation.

4. Project the *x* genes selected in step 2 and the *k* cell types estimated lied within a simplex subspace.

5. Deconvolution is performed.

ABIS (ABsolute Inmune Signal deconvolution)

It is a supervised method, designed to decompose data with both microarray and RNA-Seq signal. In addition, before deconvolution, this method requires normalisation by mRNA abundance, calculating the optimal α coefficient for each cell type by calculating the difference between the estimated and real values:

$$\min_{\widehat{\alpha} \epsilon (l,u)} \sqrt{\sum_{i=1}^{k} \widehat{(p_i} - p_i)^2}$$

Subsequently, the expression of the signature matrix (per cell type) is multiplied by this value, and the deconvolution is performed. For this, ABIS is based on a robust linear model (RLM), which for each sample, is defined as:

$$y = \hat{p}_1 \hat{\alpha}_1 x_1 + \hat{p}_2 \hat{\alpha}_2 x_2 + \cdots \hat{p}_k \hat{\alpha}_k x_k + \varepsilon$$

Where *k* represents the number of cell types to be estimated, *y* is the expression of the gene, $\hat{p}_i$ is the cell type ratio, $\hat{\alpha}_i$ is the mRNA abundance and x_i is the expression of the gene in the corresponding cell type.

FARDEEP (Fast And Robust Deconvolution of Expression Profiles)

Supervised method designed to solve partial deconvolution problems, previously eliminating outliers that may disrupt the results. For this purpose, FARDEEP is based on the aLTS (Adaptive Least Trimmed Squares) algorithm, which, as we are working with proportions (positive values), is proposed as an iterative NNLS (Non Negative Least Squares) model that initialises as follows:

$$\hat{\beta}^{(0)} = \underset{\beta}{\operatorname{argmin}} \, ||y - X\beta||_2^2, \quad \text{donde } \beta \geq 0;$$

$$r^{(0)} = y - X\hat{\beta}^{(0)}$$

Where the coefficient β represents the proportions to be estimated, *X* the gene expression, *y* the expression of the genes in the samples and *r* the residuals, with the superscript indicating the iteration

number. With $\overline{N}$and $\underline{N}$ being an overestimate and an underestimate of the number of outliers found, for the j iteration (with j ≥ 1) the value of $\underline{N}$ is updated:

$$\overline{N}^{(j)} = \begin{cases} \left|\left\{i : \left|r_i^{(j-1)}\right| > r_{med}^{(j-1)}\right\}\right|, & j = 1 \\ \min\left(\left|\left\{i : \left|r_i^{(j-1)}\right| > k \cdot r_{med}^{(j-1)}\right\}\right|, \overline{N}^{(j-1)}\right), & j \geq 2 \end{cases}$$

And the process is repeated until $\overline{N}^{(j)}$ and $\underline{N}^{(j)}$ converge to the same solution. When running the algorithm in R, it obtained absolute frequencies values, so we must transform these counts to relative values by dividing by the total amount contained in each sample.

CIBERSORT

Supervised method that solves a partial deconvolution, consequently, a mixture matrix and a signature matrix are needed as parameters. To perform the deconvolution, it is based on the machine learning algorithm known as Support Vector Regression (SVR), which is a feature of Support Vector Machine (SVM). This algorithm represents the regression model that best fits the data on a hyperplane, selecting support vectors (in our case the support vectors are the marker genes) that define the limits of the error (ε) that the model is able to tolerate. The hyperplane is defined by the following equation:

$$MIN(\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}|\xi_i|)$$

Where *w* represents the proportions of the cell types to be estimated, *C* is a positive constant that allows controlling the error, so if this value increases, then the tolerance for points outside ε will increase too.

Finally, ξ_i is the parameter in charge of controlling the error committed in the approximation of the support bands (defined by the support vectors), calculating the distance between the points represented outside them and the limits of the acceptance region.

RESULTS

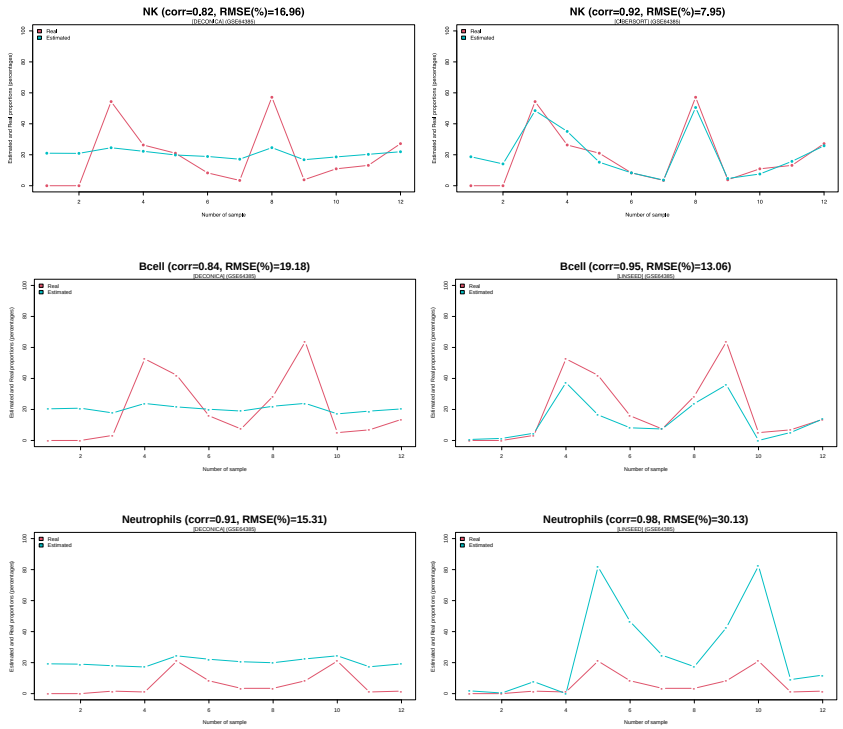
Comparison of methods using data with microarray gene expression

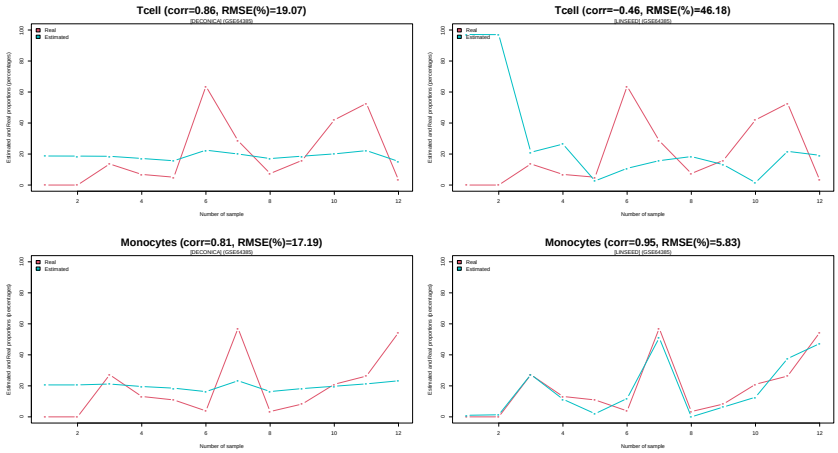
First, it was decided to analyse the results obtained after the implementation of DECONICA, LINSEED, CIBERSORT and FARDEEP, using the GSE64385 and GSE20300 databases. The corrplots, showing the values of the Pearson correlation coefficients for the GSE34685 data, are shown below:

GSE64385

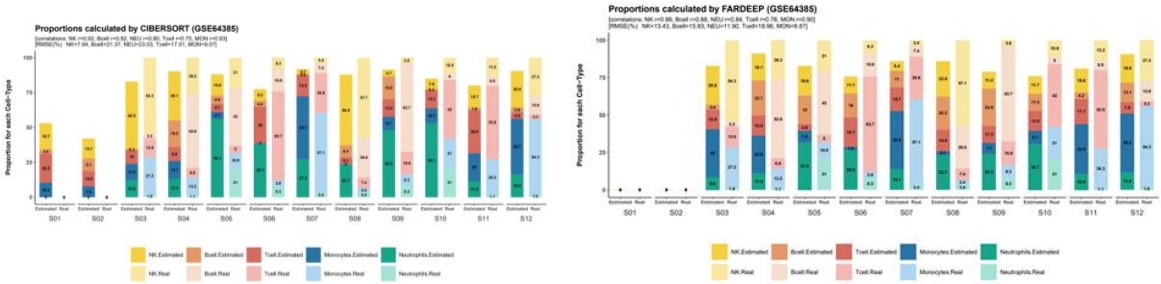
The correlation values of the supervised methods (CIBERSORT and FARDEEP) are slightly higher, although DECONICA also has fairly good correlation coefficients. To observe the proportions of each cell type in the samples, we have been made *heatmaps*, but they did not reveal too much information, so we created other types of plots (*cell signature plots* and *bar mixture plots*). The *cell signature plots* show the relative abundance values (in percentage, %) for each cell type separately in the series of samples studied, presenting two trajectories, one for the real data and another for the estimated data, for each method. These graphs also include the calculation of Pearson's correlation and the RMSE (root mean square error) value between these trajectories. As an example, those obtained from the DECONICA and LINSEED methods using GSE64385 data are shown below:

DECONICA **LINSEED**





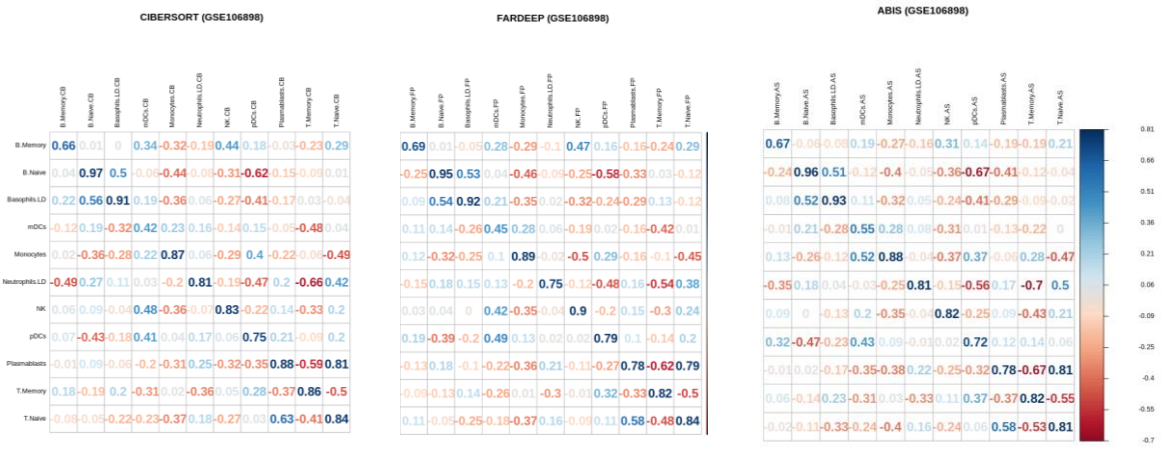
This type of plots reveals much more information contained into the data. The corrplots showed high values of Pearson's correlation coefficients for DECONICA. However, these plots show that the values of the estimated proportions show little variation between different cell types and samples, all of them being distributed around a mean value. On the other hand, the *bar mixture plots* obtained after the application of CIBERSORT and FARDEEP for the same data are shown below:



The previous figures show a clear difference between the two methods: the robustness to the presence of noise in the data. In this case, the noise is represented by tumour cells, present in the first two samples (pure cancer cell samples), so the proportions of blood cell types should be zero, as shown in the FARDEEP plot.

Comparison of methods between microarray expression signal data and RNA-Seq expression signal data

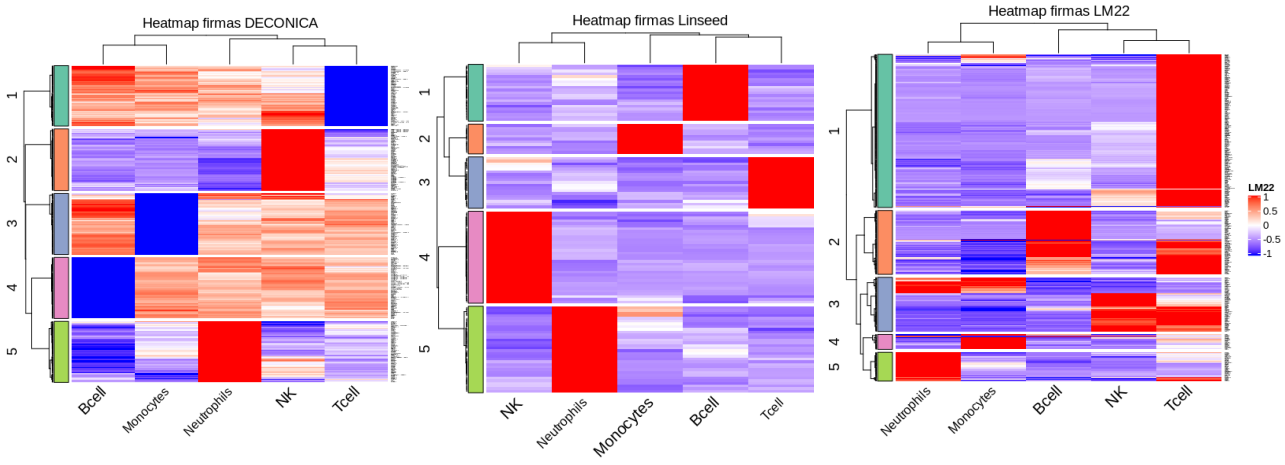
In this section, the supervised methods (CIBERSORT, FARDEEP and ABIS) have been implemented in two databases, one with microarray gene expression signal (GSE106898) and another with RNA-Seq gene expression signal (GSE107011). The corrplots obtained for the microarray data are shown below:



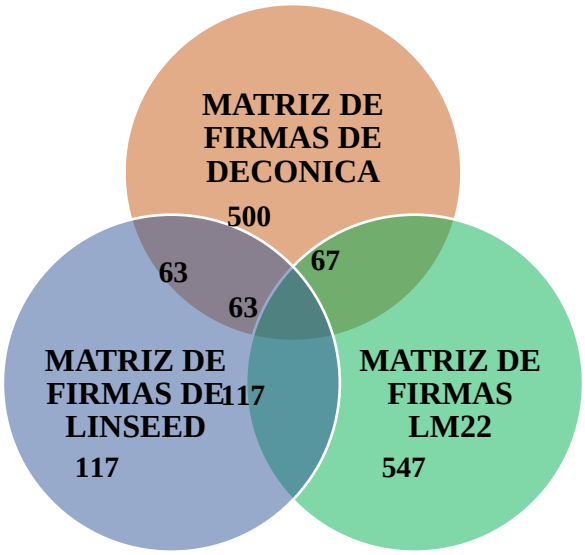
The correlations were seen in these plots are similar between the three methods studied, but when applying the methods using data with RNA-Seq signal, the correlations for CIBERSORT are slightly lower than these values for the rest methods. In addition, the *cell signature plots* for this case show too smaller RMSE values than the previous comparison, which does not indicate a better estimate, so it is logical to think that the difference will also be smaller.

Analysis of cell markers through signature matrices

Finally, an analysis was performed on the cell signatures provided by the LM22 matrix (used in the CIBERSORT and FARDEEP methods) and that estimated by the unsupervised DECONICA and LINSEED methods to identify five peripheral blood cells (T lymphocytes, B lymphocytes, NK cells, neutrophils and monocytes), using the GSE64385 data. To do this, we chose to apply a clustering analysis in the statistical software R, using the `hclust()` function, and selecting the 'Ward.D2' minimum variance method as the clustering technique. To show the results, a *heatmap* has been made, showing the expression of the marker genes of each cell type, as well as including a dendogram with the grouping of the data obtained using the clustering algorithm.



The first *heatmap* shows that DECONICA considers the same number of genes to identify cell types, which is not the case for the other two signature matrices. Moreover, as this method is based on the search for components (in our case cell types) that present asymmetric distributions, without considering the sign of the asymmetry, some cell types may contain marker genes with an absence of expression in them, which does not make biological meaning. The intersection between the genes contained in the three matrices will also be studied:



The LM22 matrix is considered as a reference matrix, as it has been studied by the creators of CIBERSORT for use in deconvolution methods. It is observed that the matrix estimated by DECONICA contains 500 marker genes, of which 67 are found in the LM22 matrix, while the matrix calculated by LINSEED only contains 117 marker genes, but all of them are contained in the LM22 matrix, so it can be said that LINSEED is more robust in the selection of cell signatures.

ANEXOS: Gráficos y tablas suplementarios

Mediante este documento, correspondiente a los datos suplementarios, se profundizará en aspectos más específicos del Trabajo de Fin de Grado: **Análisis e implementación de algoritmos de deconvolución de mezclas celulares complejas basados en expresión de genes (firmas génicas) y aplicación a muestras de tumores**.

En primer lugar, se mostrarán tres tablas que aportan información sobre el estudio. Las dos primeras, resumen las características principales de los datos utilizados y los tipos celulares analizados en cada conjunto de datos. En última tabla, se recogen las propiedades fundamentales de los métodos de deconvolución implementados en este trabajo. Después, se expondrán una serie de gráficos, comenzando por los obtenidos durante el procedimiento del método LINSEED (nivel de significación de los genes, gráfico SVD, proyección de los datos en el subespacio simplex, gráfico de proporciones para cada tipo celular y la representación de la comparación entre las frecuencias estimadas y reales) para los conjuntos de datos GSE64385 GSE20300 (S1-S10). A continuación, se presentarán los correspondientes a los distintos enfoques comparativos entre las proporciones estimadas y las proporciones reales, mostrándose primero los resultados obtenidos para la mezcla GSE64385, y en segundo lugar los obtenidos para GSE20300, en cada uno de los tipos de gráficos. Primero, se expondrán los gráficos de correlaciones (S11 y S12) y los *heatmaps* (S13 y S14), seguidos de los *cell signature plot* (uno para cada tipo celular), en los que aparecen representadas ambas puntuaciones (estimadas y reales) en forma de puntos, y los *bar mixture plots*, que engloban en un mismo gráfico las frecuencias relativas estimadas y originales, para cada tipo celular en cada muestra analizada. Por último, para el análisis de los marcadores celulares, se ha realizado un *heatmap* para cada matriz de firmas considerada en el estudio: la matriz LM22, la matriz firmas estimada por DECONICA y la calculada por LINSEED, además de una figura ilustrativa, que nos muestra la intersección de los genes entre las tres matrices de firmas mencionadas.

LISTADO DE TABLAS

Tabla 1. Resumen de los conjuntos de datos utilizadas en el trabajo.

Nº acceso	Plataforma de expresión génica	Nº muestras	Nº genes	Fuente biológica	Tipos celulares	Referencia
GSE64385	Microarray HGU133 Plus 2.0 - *Affymetrix*	12	54675	PBMCs y celulas HCT116	5	(Becht et al., 2016)
GSE20300	Microarray HGU133 Plus 2.0 - *Affymetrix*	24	54675	Sangre periférica	4	(Shen-Orr et al., 2010)
GSE107011	RNA-Seq HiSeq 2000 - *Illumina*	13	17487	PBMCs	17	(Monaco et al., 2019)
GSE106898	Microarray Human HT-12 V4.0 - *Illumina*	13	17487	PBMCs	11	(Monaco et al., 2019)

Tabla 2. Clasificación de los tipos celulares estudiados.

Cell types	Datasets			
	GSE107011	**GSE106898**	**GSE64385**	**GSE20300**
Lymphocytes	B Naive	B Naive	B cells	Lymphocytes
	B Memory	B Memory		
	T CD4 Naive	T Naive	T cells	
	T CD8 Naive			
	T CD4 Memory	T Memory		
	T CD8 Memory			
	Plasmablasts	Plasmablasts		
T Innate	T γδ Vd2			
	T γδ non-Vd2			
	MAIT			
NK	NK	NK	NK	
pDCs	pDCs	pDCs		
mDCs	mDCs	mDCs		
Monocytes C	Monocytes C	Monocytes	Monocytes	Monocytes
Monocytes NC+I	Monocytes NC+I			
Granulocytes	Neutrophils LD	Neutrophils LD	Neutrophils	Neutrophils
	Basophils LD	Basophils LD		
Eosinophils				Eosinophils

Tabla 3. Resumen de las características principales de los métodos de deconvolución.

Método	Algoritmo de deconvolución	Supervisado	Selección recursiva de las variables	Software	Referencia
DECONICA	ICA	No	Sí	R (GitHub)	(Czerwińska, 2018)
LINSEED	Simplex	No	Sí	R (GitHub)	(Zaitsev et al., 2019)
ABIS	RLM	Sí	No	R (CRAN)	(Monaco et al., 2019)
FARDEEP	aLTS	Sí	No	R (CRAN)	(Hao et al., 2019)
CIBERSORT	*v*-SVR	Sí	No	R (cibersortX)	(Newman et al., 2015)

LISTADO DE GRÁFICOS
GRÁFICOS OBTENIDOS DURANTE LA EJECUCIÓN DEL MÉTODO LINSEED (GSE64385)

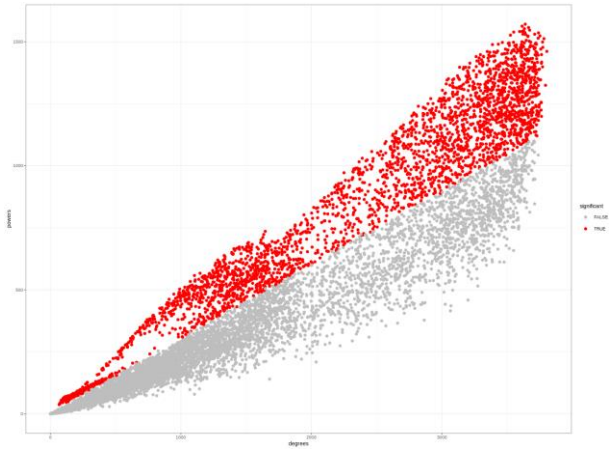


Gráfico S1. Nivel de significación de los genes en GSE64385 tras evaluar la colinealidad y la correlación de Spearman por pares.

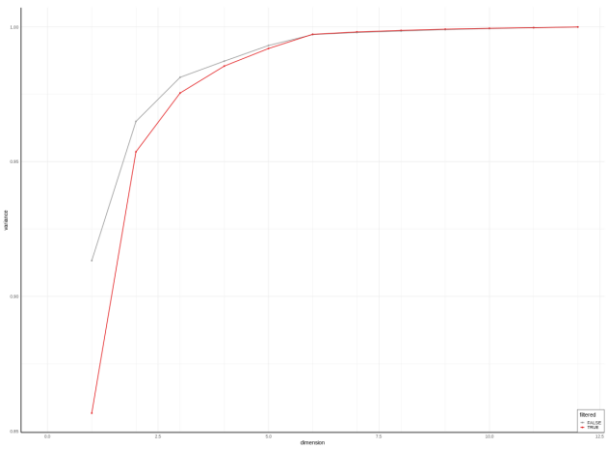


Gráfico S2. Selección del número de tipos celulares mediante la descomposición de los datos (GSE64385) con SVD.

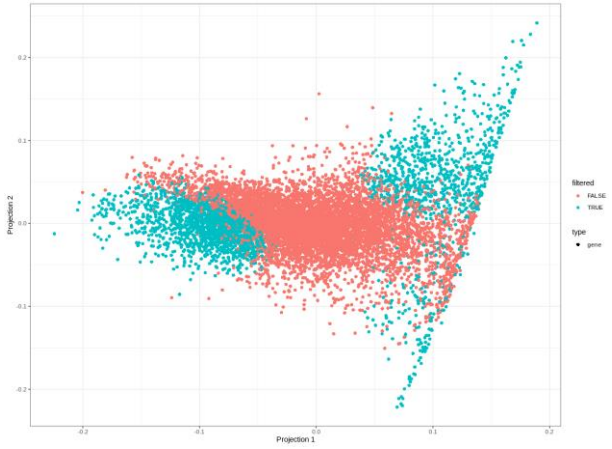


Gráfico S3. Proyección de los datos (GSE64385) en el subespacio simplex tras el filtrado de los genes no significativos.

Gráfico S4. Proporciones de los tipos celulares en cada muestra (GSE64385).

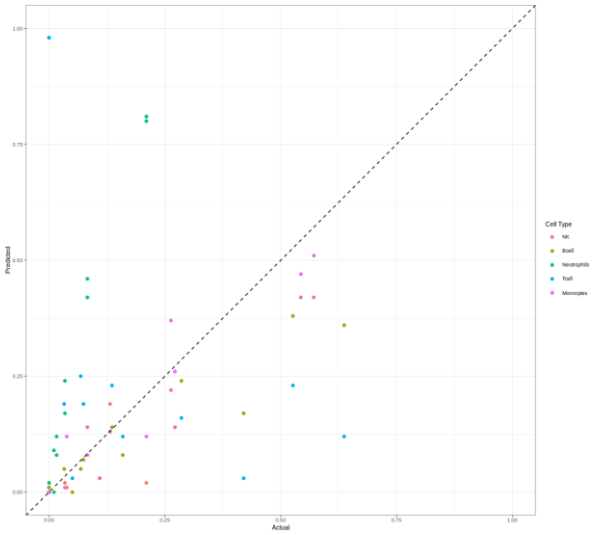


Gráfico S5. Comparación de las frecuencias relativas observadas y estimadas (GSE64395).

GRÁFICOS OBTENIDOS DURANTE LA EJECUCIÓN DEL MÉTODO LINSEED (GSE20300)

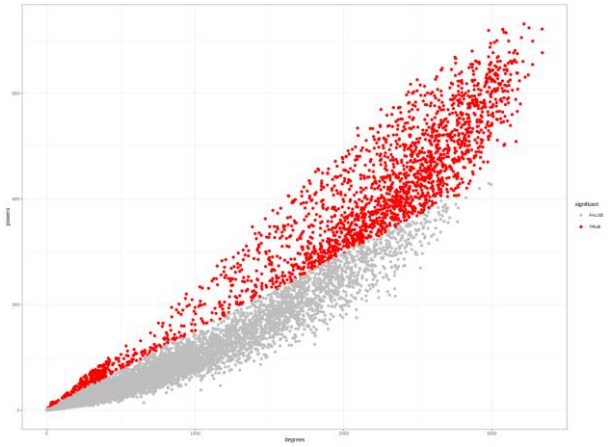


Gráfico S6. Nivel de significación de los genes en GSE20300 tras evaluar la colinealidad y la correlación de Spearman por pares.

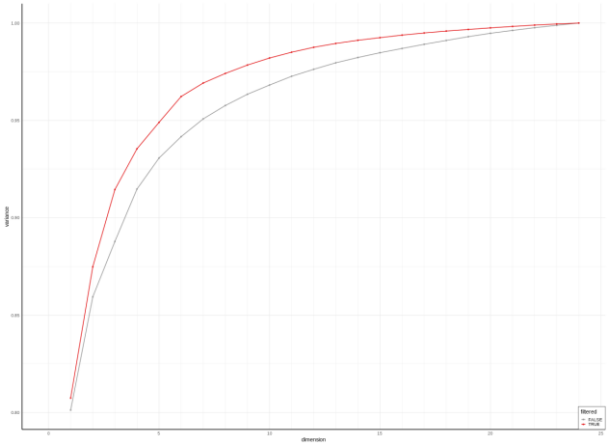


Gráfico S7. Selección del número de tipos celulares mediante la descomposición de los datos (GSE20300) con SVD.

Gráfico S8. Proyección de los datos (GSE20300) en el subespacio simplex tras el filtrado de los genes no significativos.

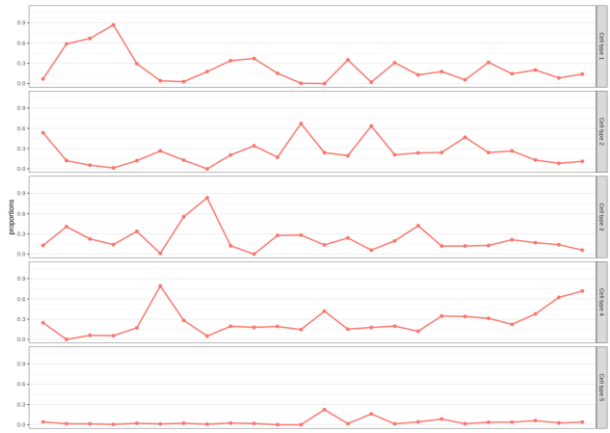


Gráfico S9. Proporciones de los tipos celulares en cada muestra (GSE20300).

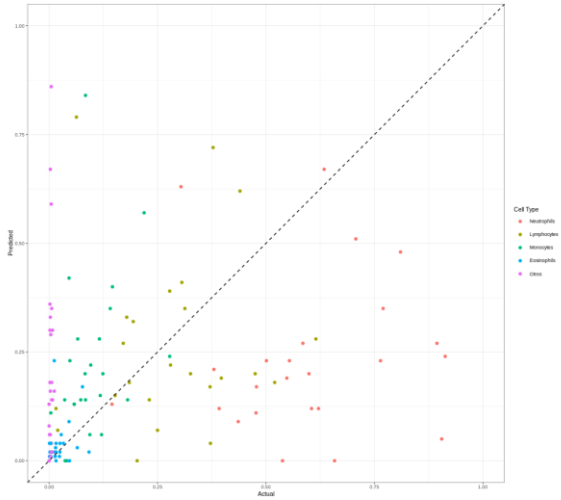


Gráfico S10. Comparación de las frecuencias relativas observadas y estimadas (GSE20300).

GRÁFICOS DE CORRELACIONES (*CORRPLOT*)

GSE64385

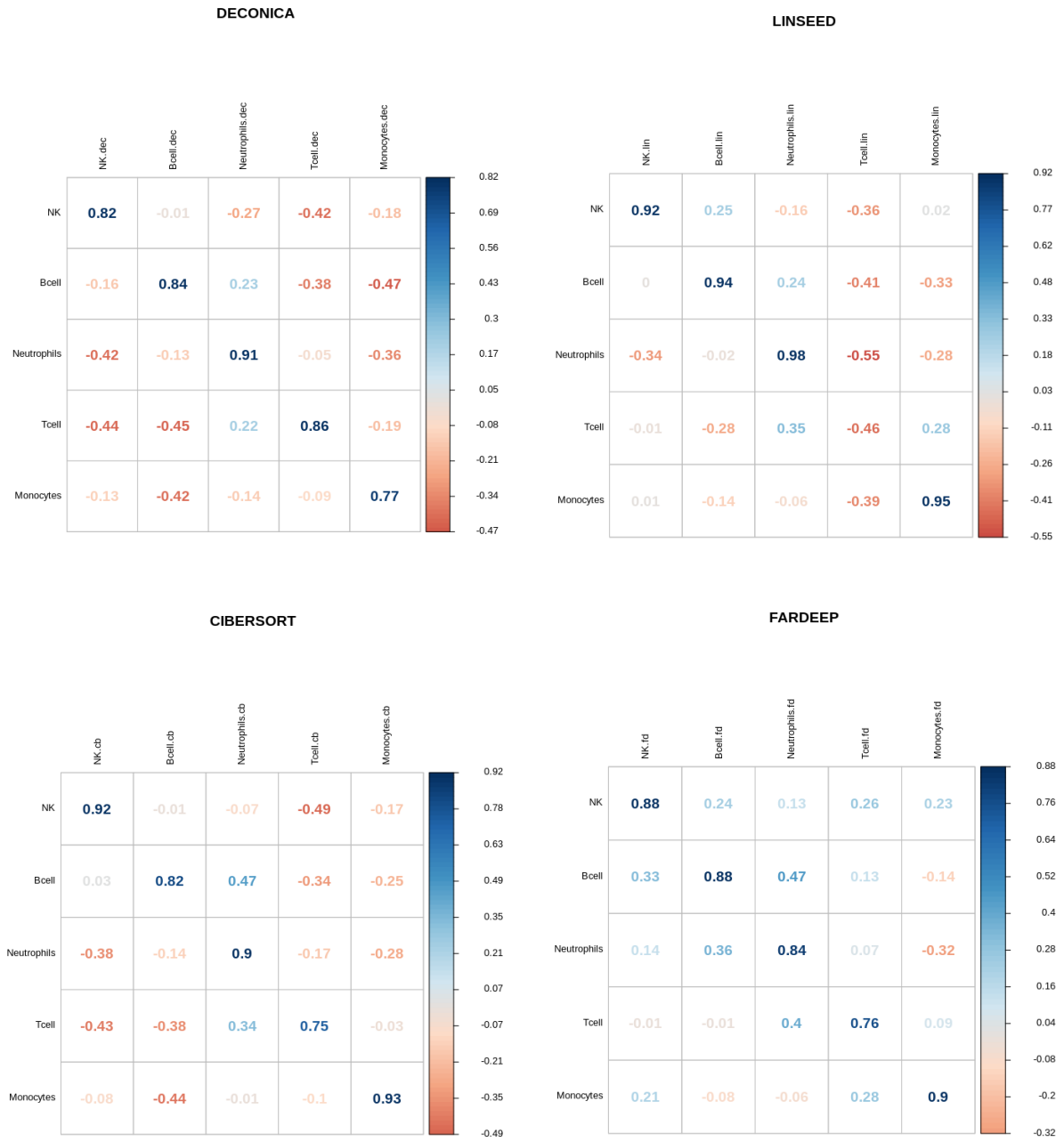


Gráfico S11. Gráfico de correlaciones (*corrplot*) para las frecuencias estimadas por DECONICA, LINSEED, CIBERSORT y FARDEEP en la mezcla GSE20300.

GSE20300

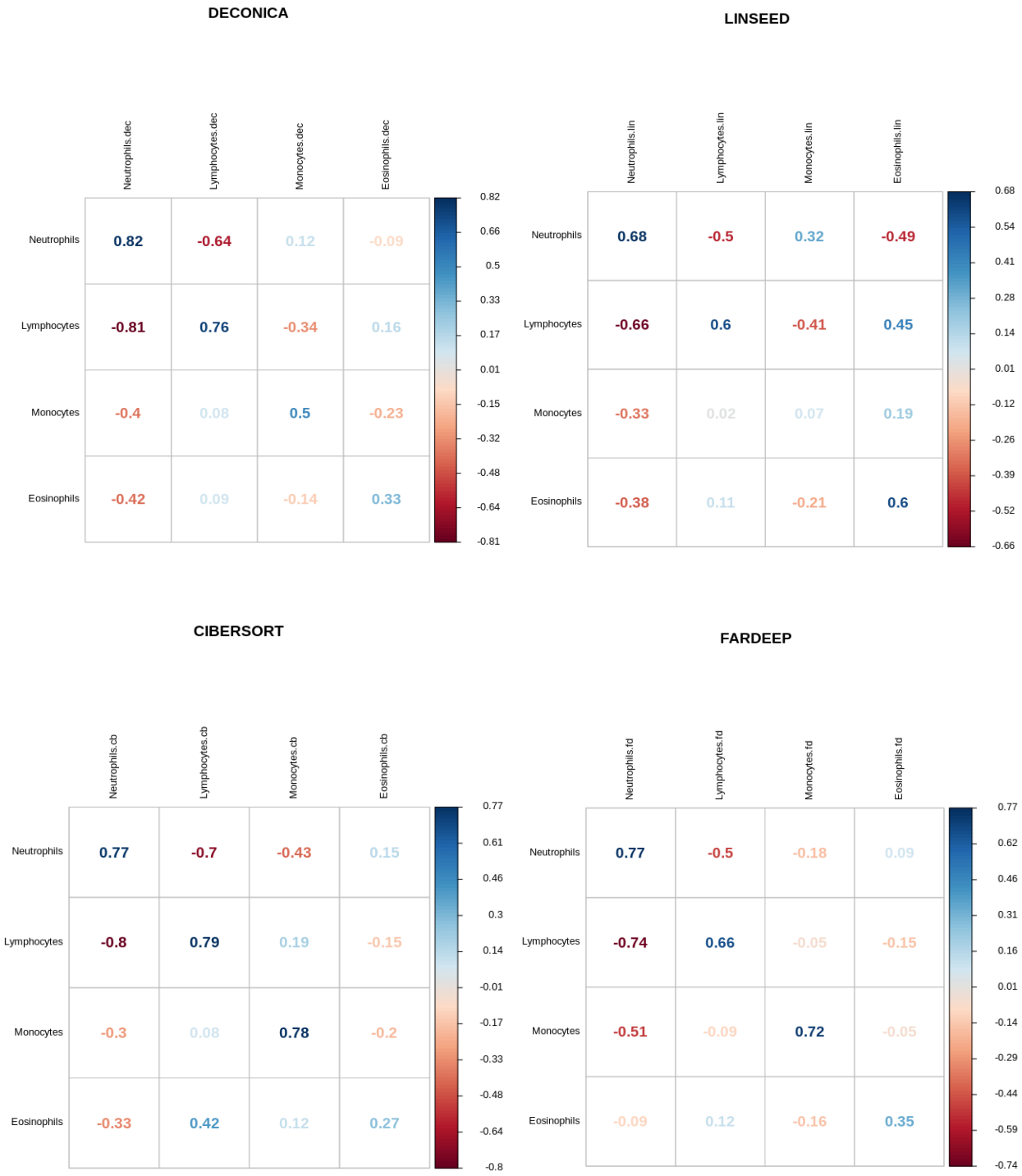


Gráfico S12. Gráfico de correlaciones (*corrplot*) para las frecuencias estimadas por DECONICA, LINSEED, CIBERSORT y FARDEEP en la mezcla GSE20300.

GSE64385

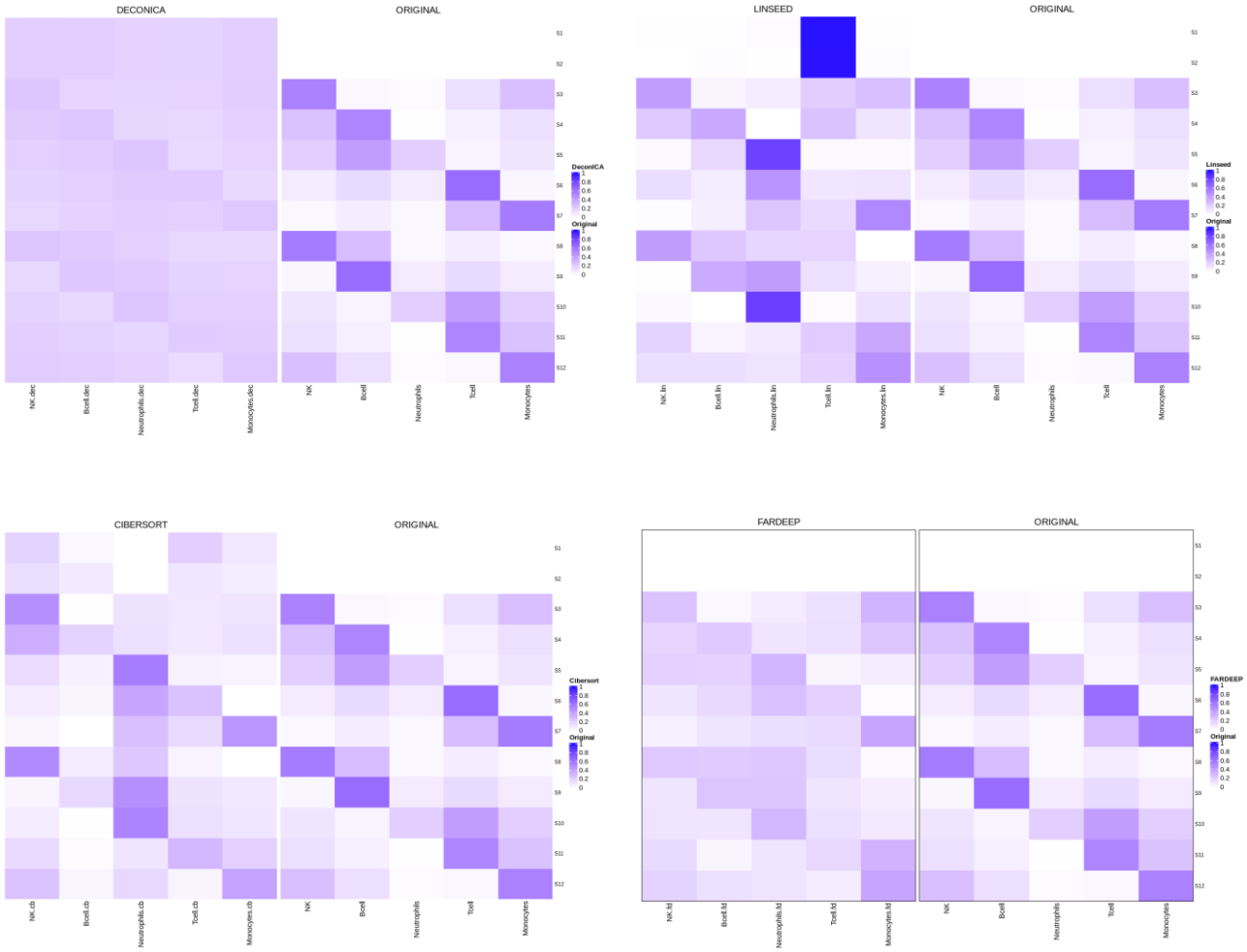


Gráfico S13. *Heatmap* de las proporciones estimadas por DECONICA, LINSEED, CIBERSORT y FARDEEP en la primera mezcla de células sanguíneas GSE64385.

GSE20300

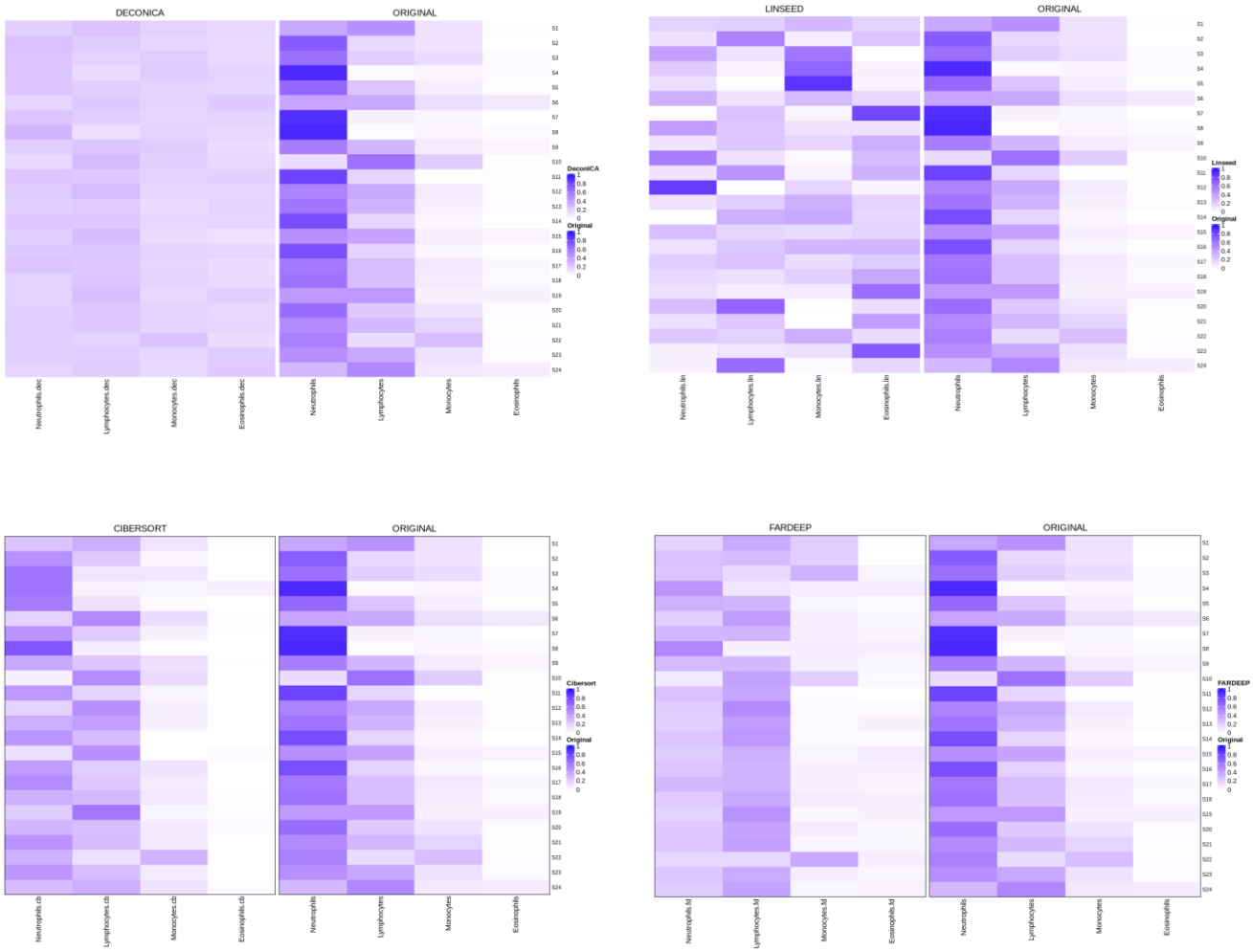
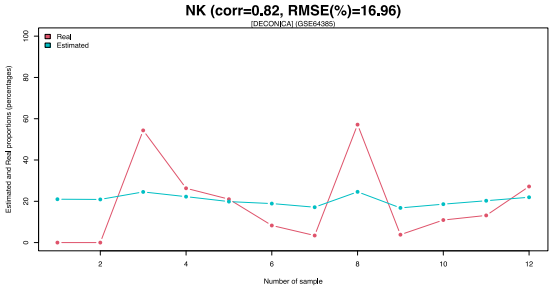
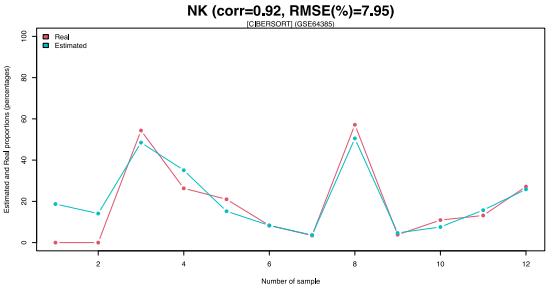
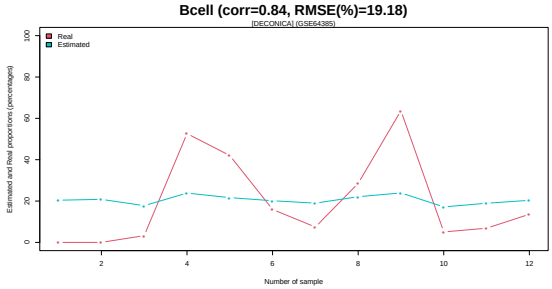
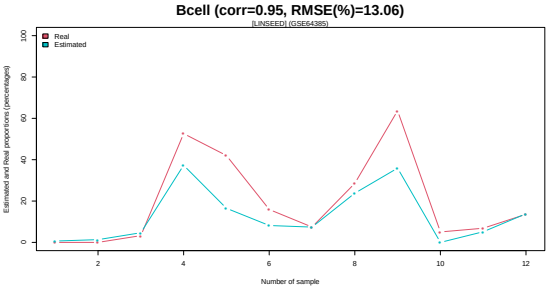
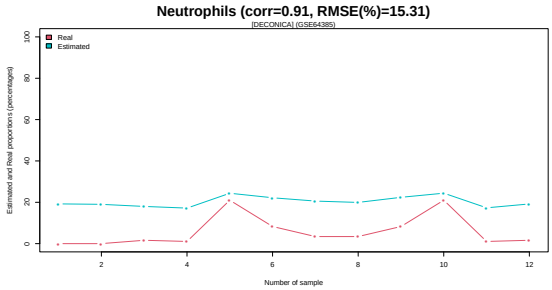
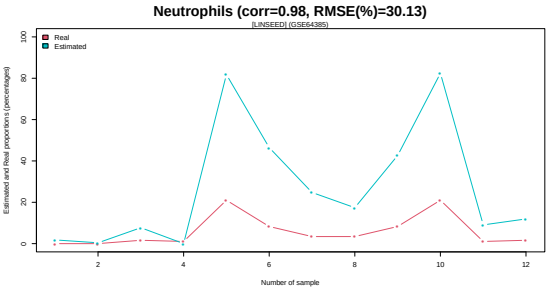
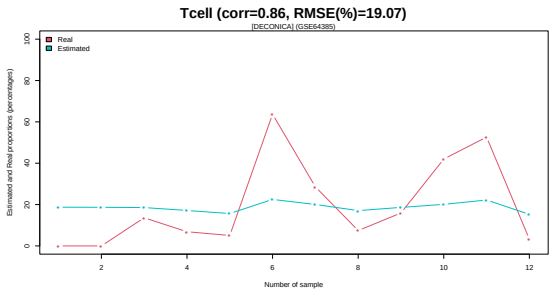
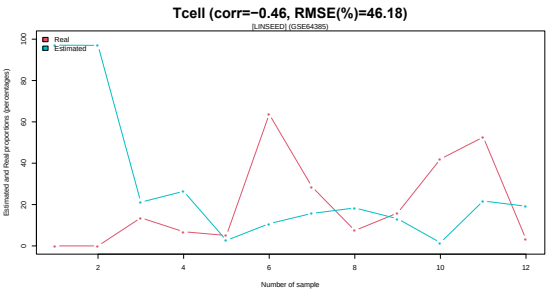
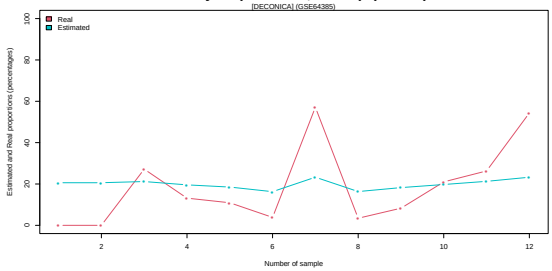
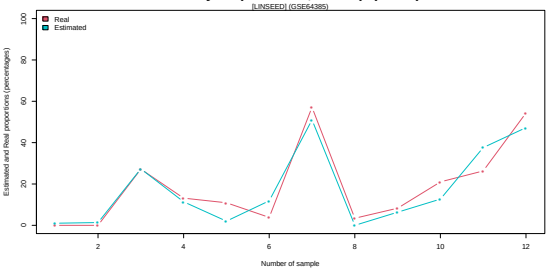


Gráfico S14. *Heatmap* de las proporciones estimadas por DECONICA, LINSEED, CIBERSORT y FARDEEP en la segunda mezcla de células sanguíneas GSE20300.

CELL SIGNATURE PLOT

GSE64385

DECONICA	LINSEED
	
	
	
	
	

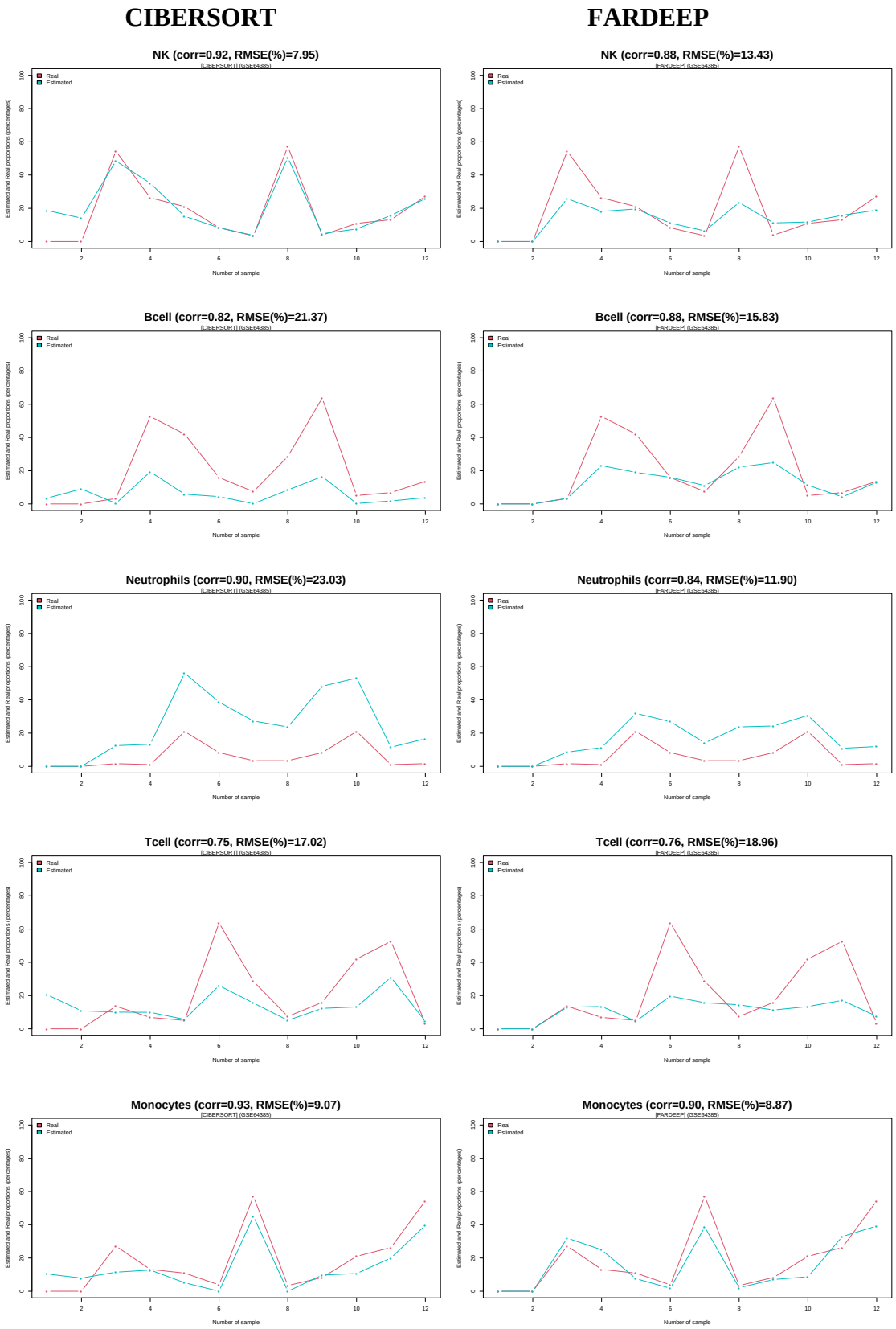
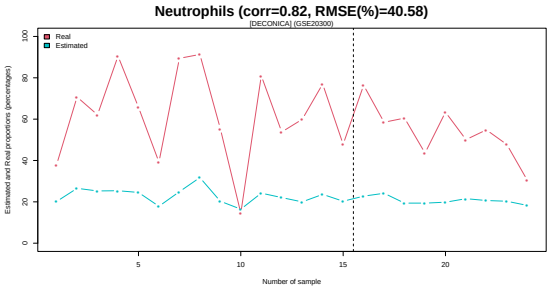
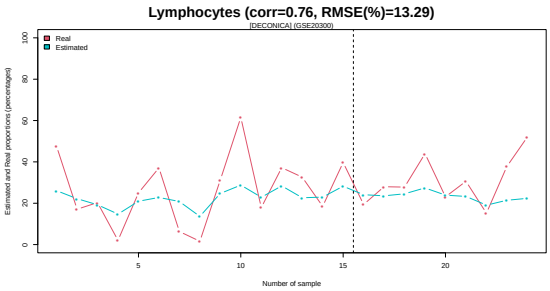


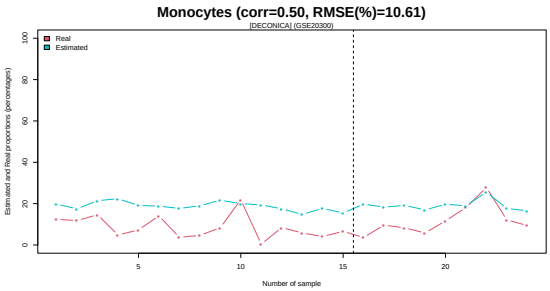
Gráfico S15. Representación de los tipos celulares (cell signature plot) para GSE64385.

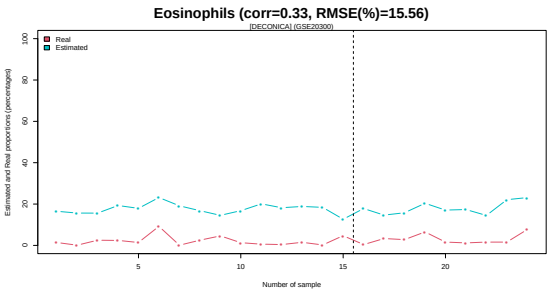
GSE20300

DECONICA

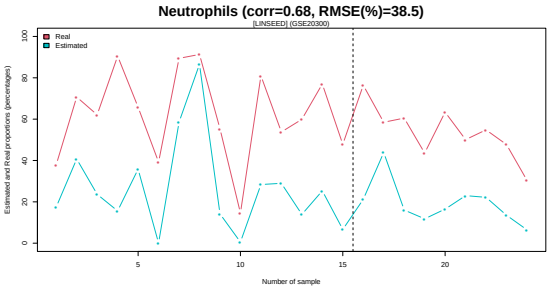


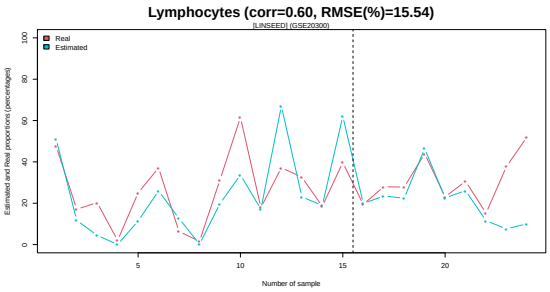


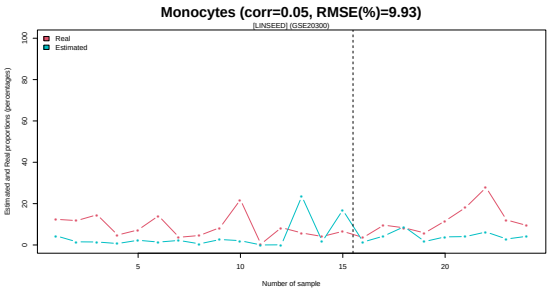


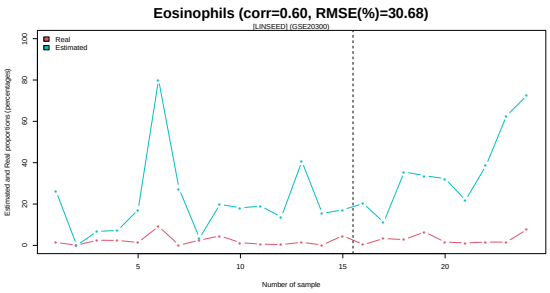


LINSEED









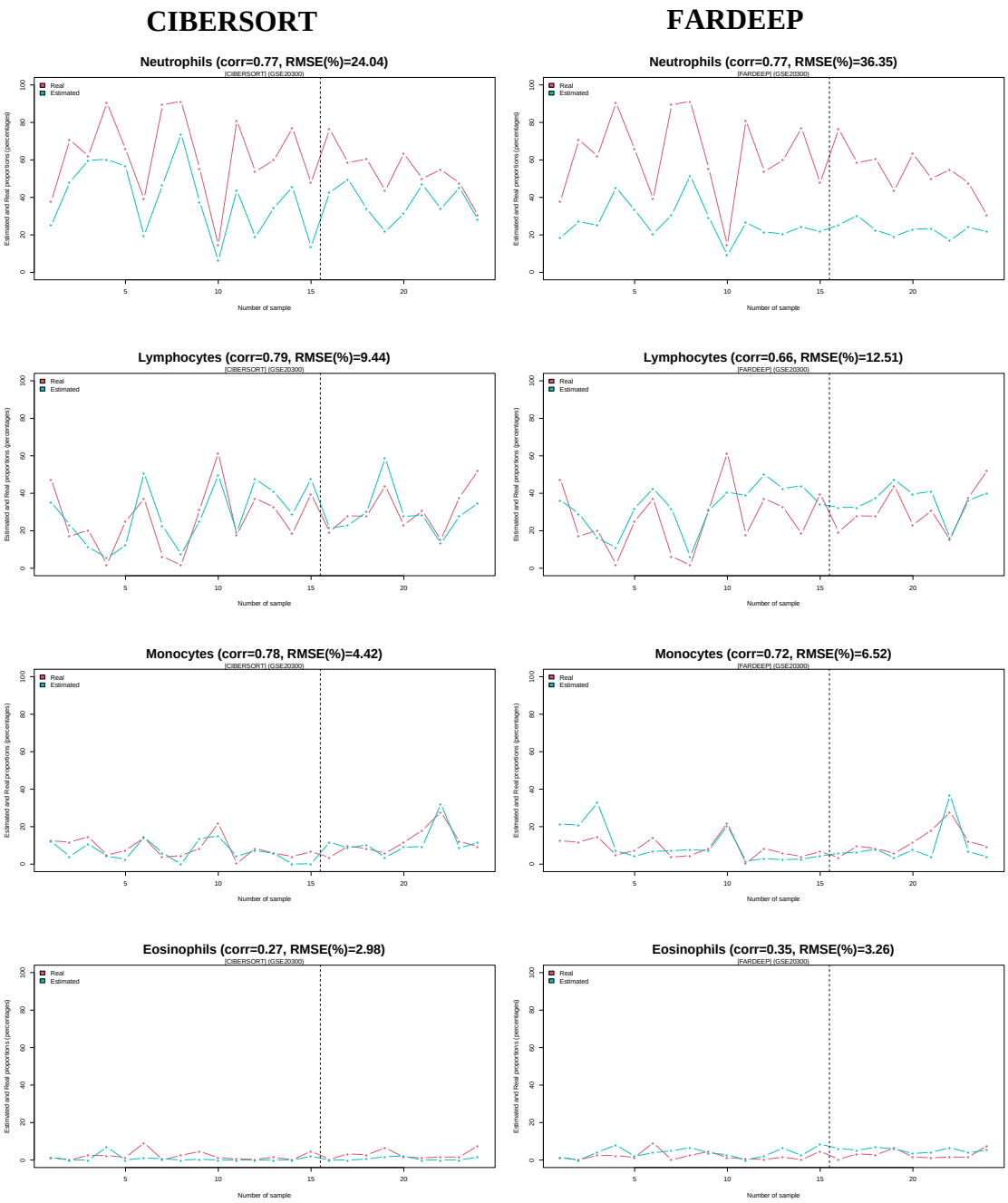


Gráfico S16. Representación de los tipos celulares (*cell signature plot*) para GSE20300. La línea dibujada separa los pacientes estables tras el trasplante renal (de la muestra 1 a la muestra 15) y los que han sufrido un rechazo agudo (de la muestra 16 a la muestra 24).

BAR MIXTURE PLOT

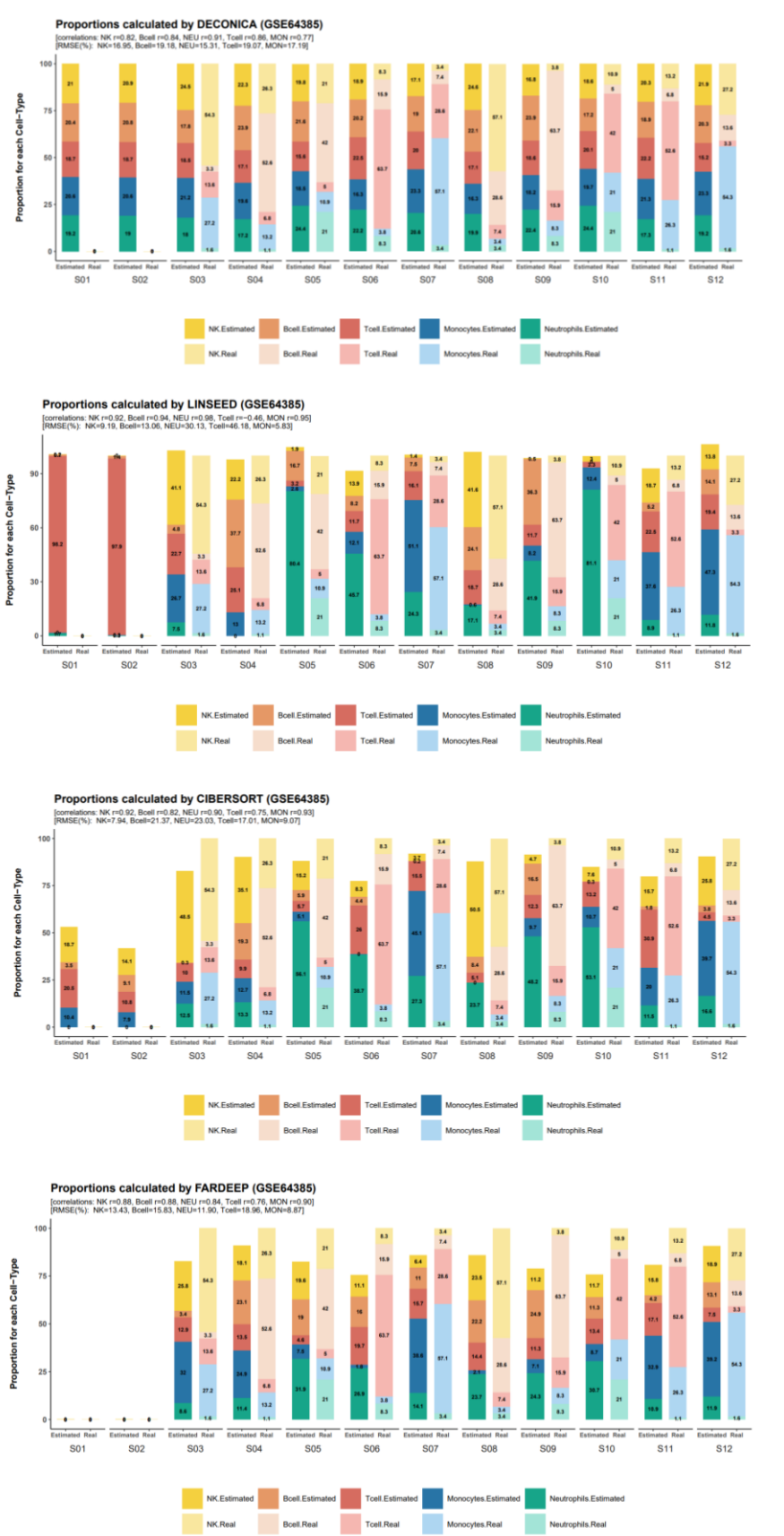


Gráfico S17. Gráfico de barras que representa las frecuencias relativas de cada tipo celular en cada muestra (*bar mixture plot*) para GSE64385.

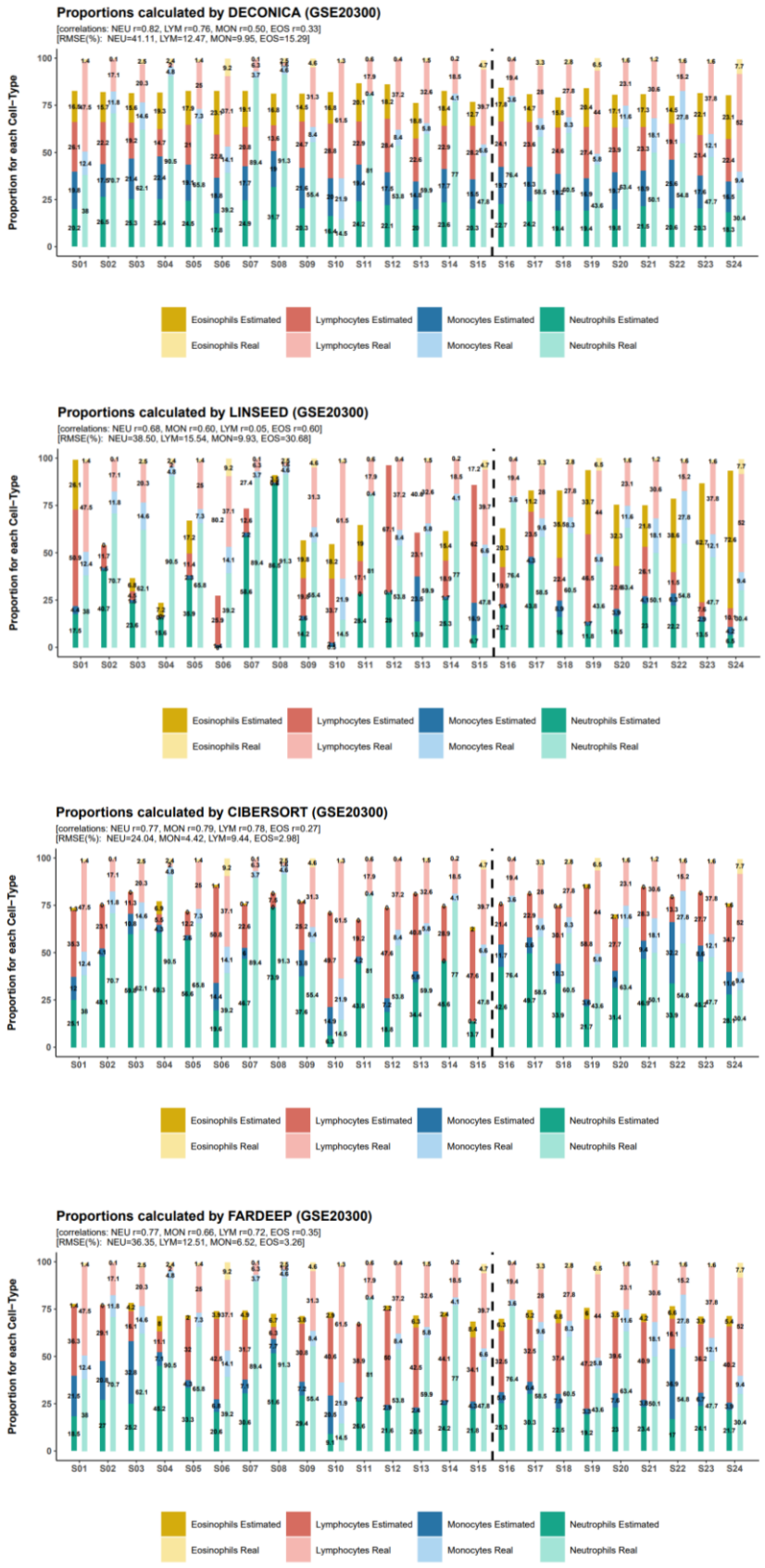


Gráfico S18. Gráfico de barras que representa las frecuencias relativas de cada tipo celular en cada muestra (*bar mixture plot*) para GSE20300. La línea dibujada separa los pacientes estables tras el trasplante renal (de la muestra 1 a la muestra 15) y los que han sufrido un rechazo agudo (de la muestra 16 a la muestra 24).

GRÁFICOS DE CORRELACIONES (*CORRPLOT*)

Datos con señal de expresión detectada mediante microarrays

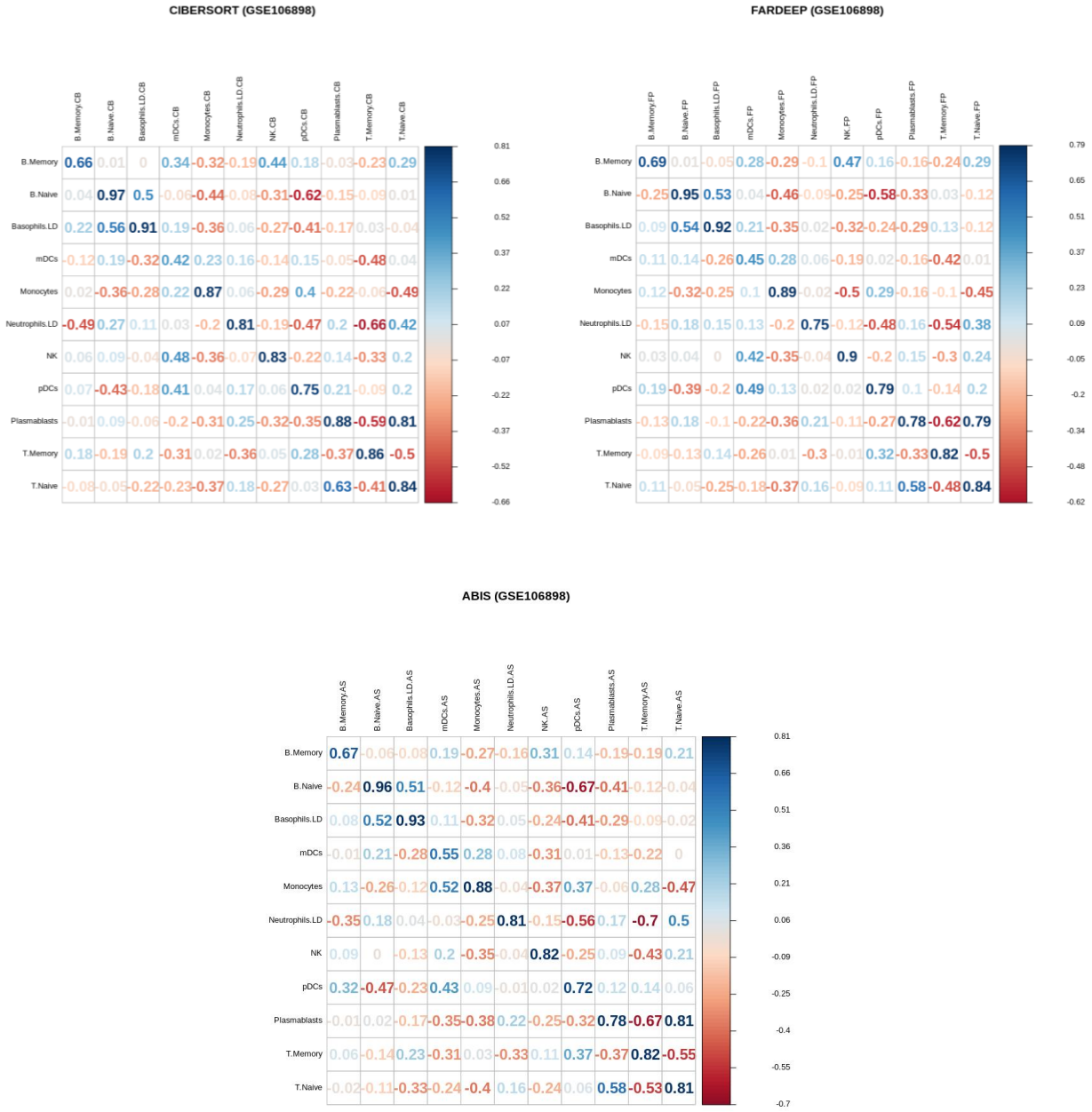
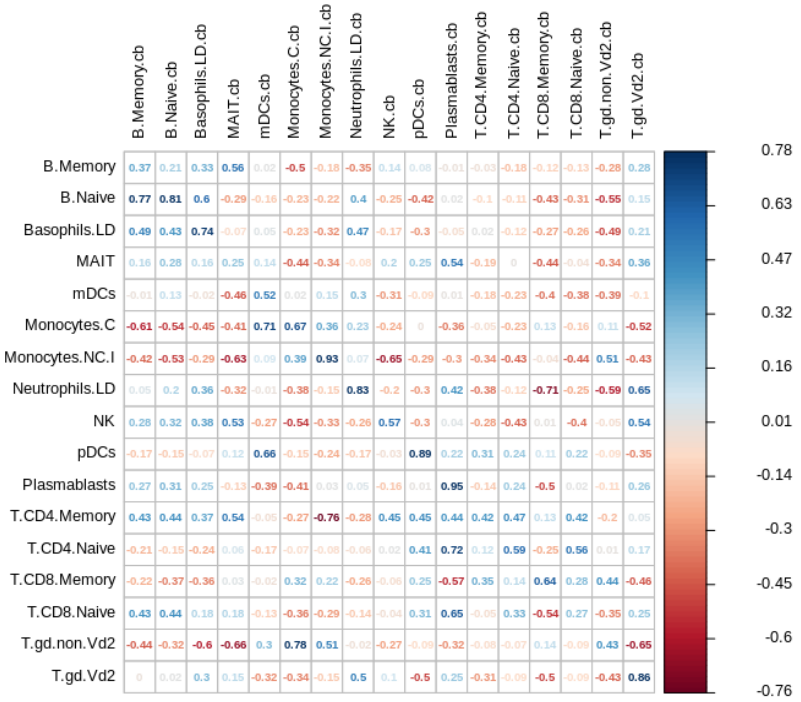


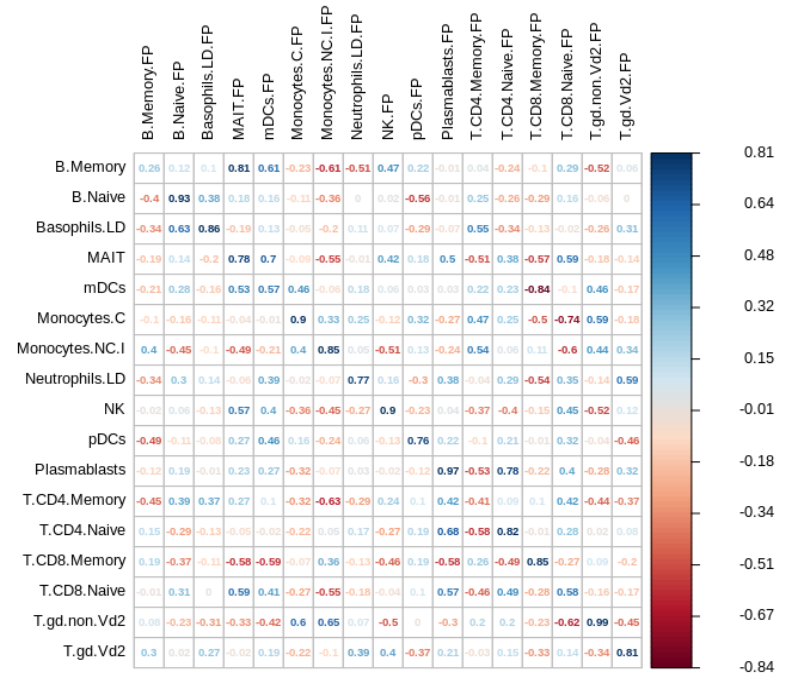
Gráfico S19. Gráfico de correlaciones (*corrplot*) para las frecuencias estimadas en la mezcla GSE106898.

Datos con señal de expresión detectada mediante RNA-Seq

CIBERSORT



FARDEEP



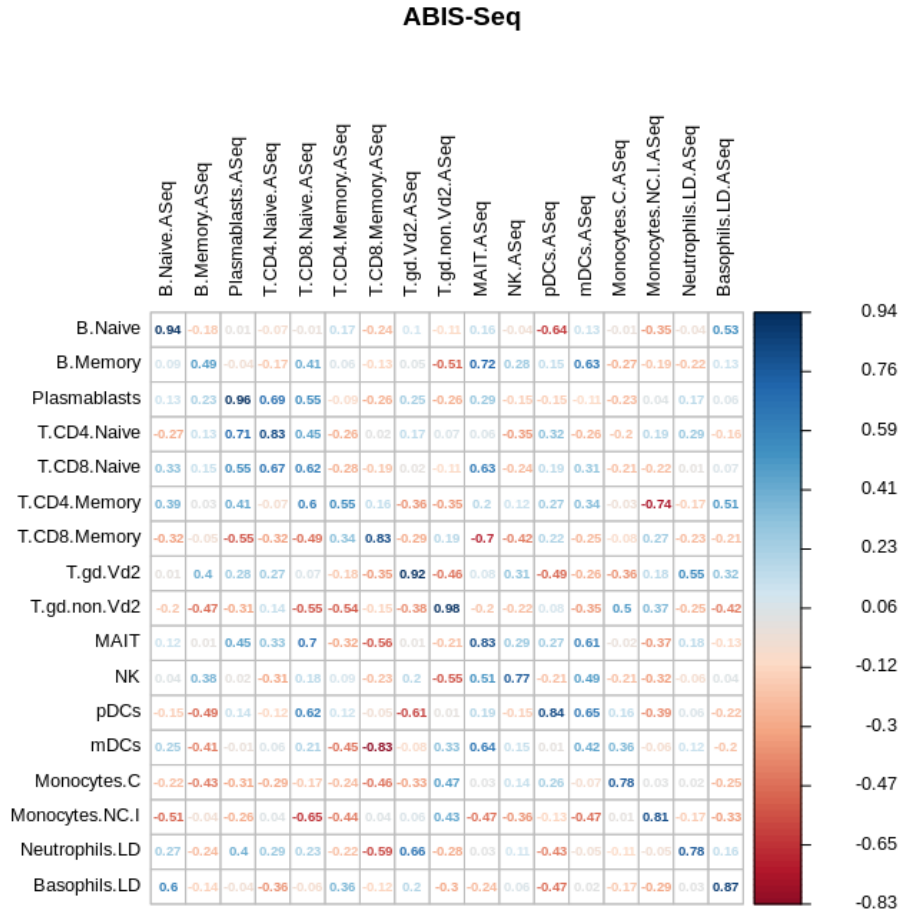


Gráfico S20. Gráfico de correlaciones (*corrplot*) para las frecuencias estimadas en la mezcla GSE107011.

Datos con señal de expresión detectada mediante microarrays

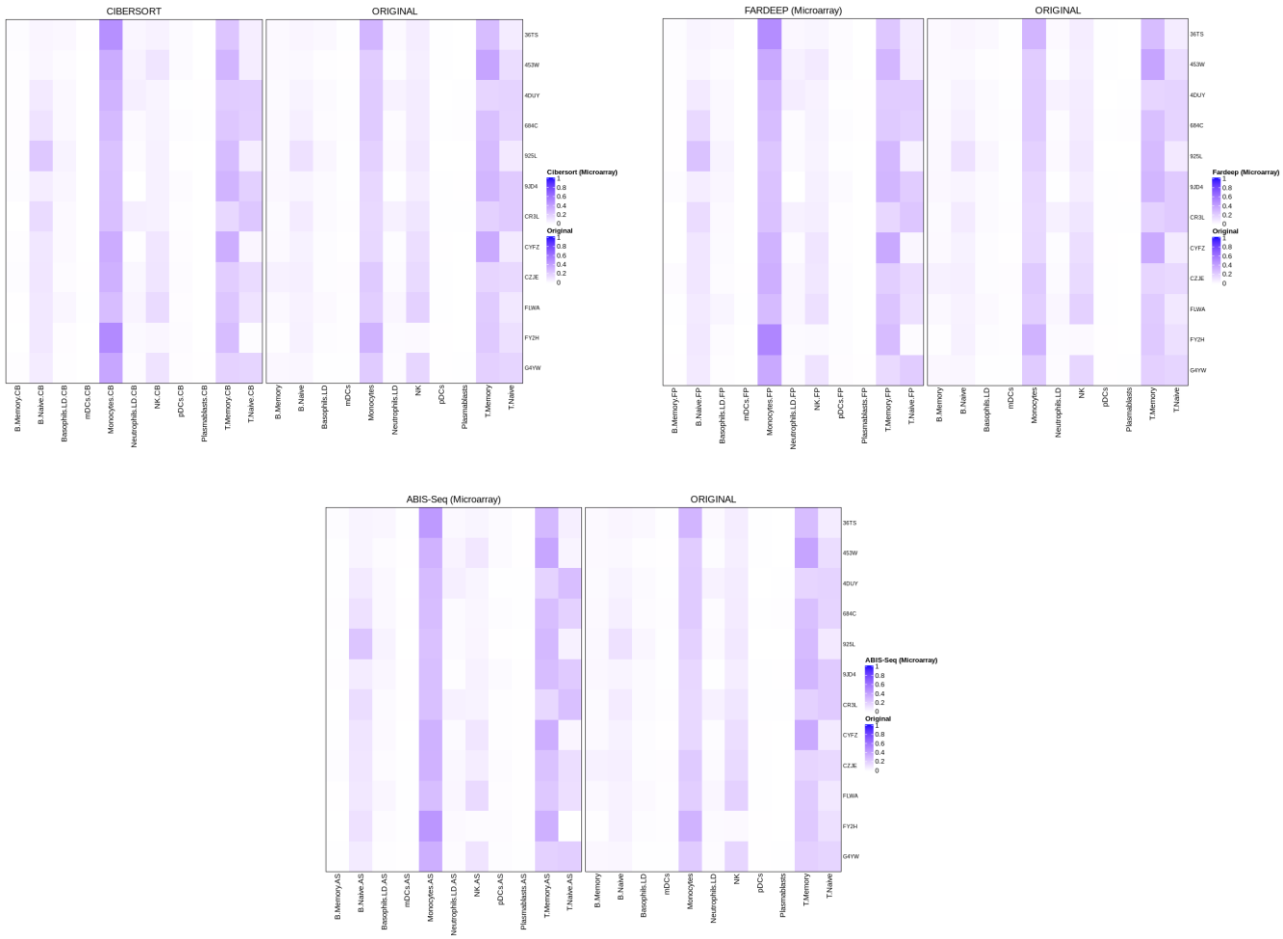


Gráfico S21. *Heatmap* que representa los resultados obtenidos tras la descomposición de la mezcla GSE107011 con señal expresión génica analizada mediante microarrays.

Datos con señal de expresión detectada mediante RNA-Seq

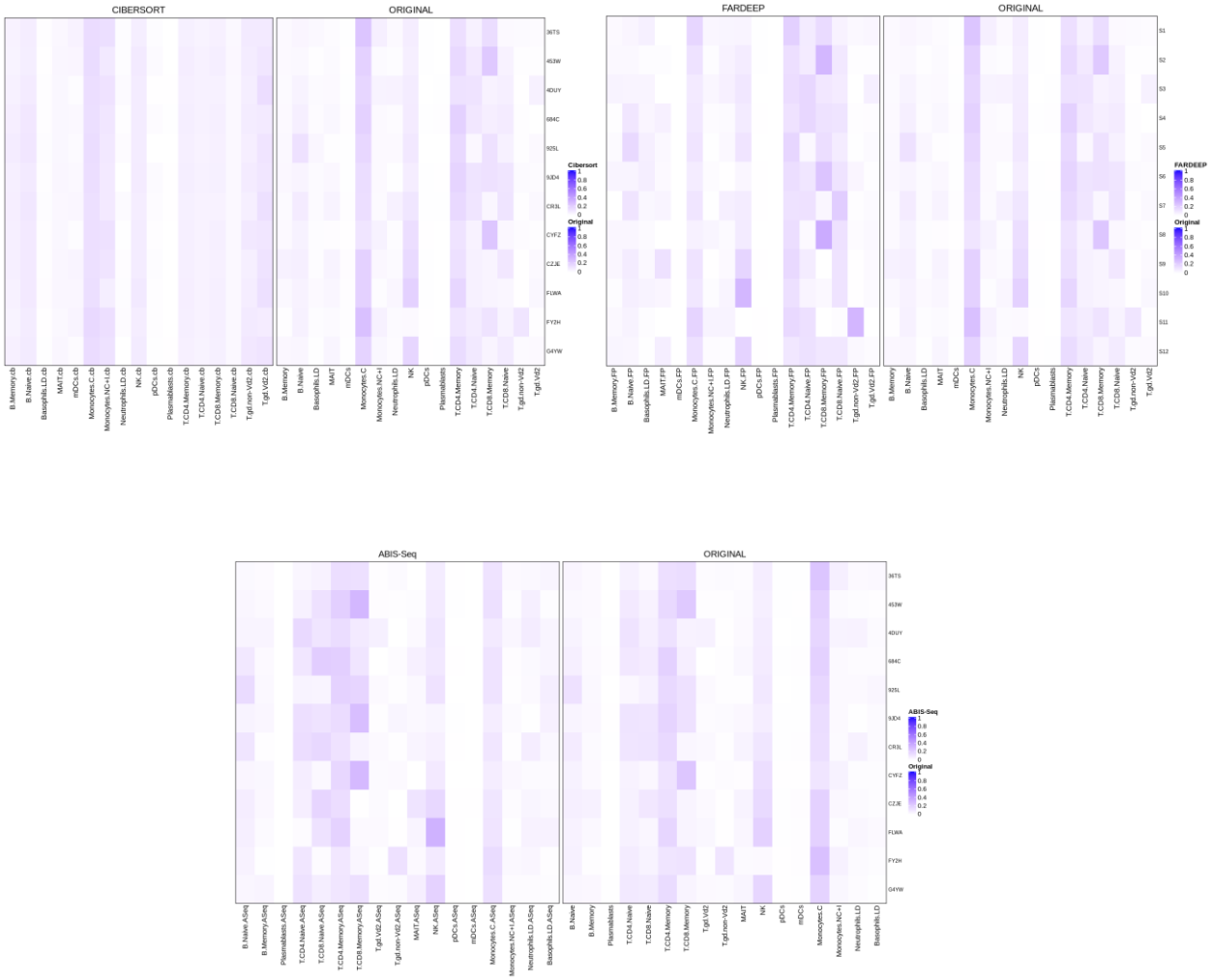
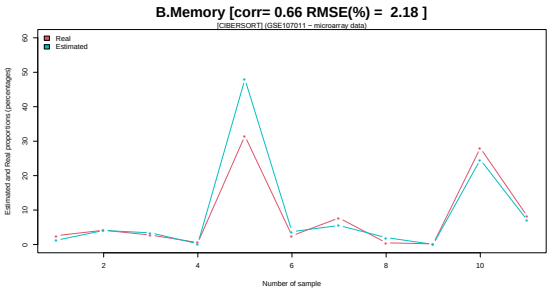


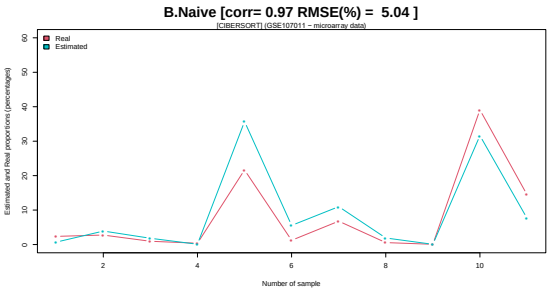
Gráfico S22. *Heatmap* que representa los resultados obtenidos tras la descomposición de la mezcla GSE107011 con señal expresión génica analizada mediante RNA-Seq.

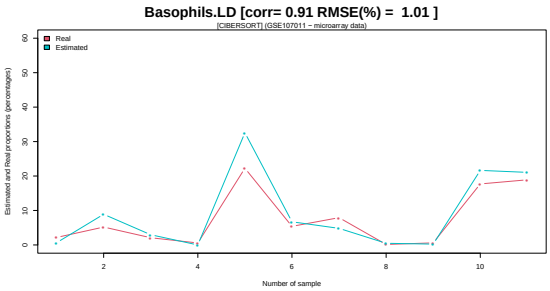
CELL SIGNATURE PLOT

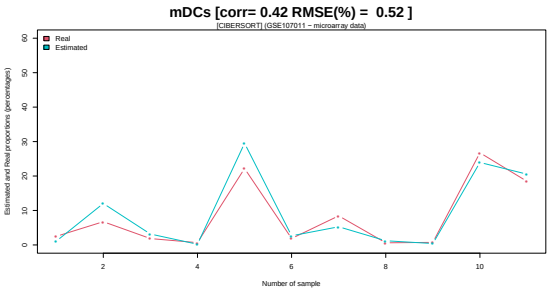
Datos con señal de expresión detectada mediante microarrays (GSE106898)

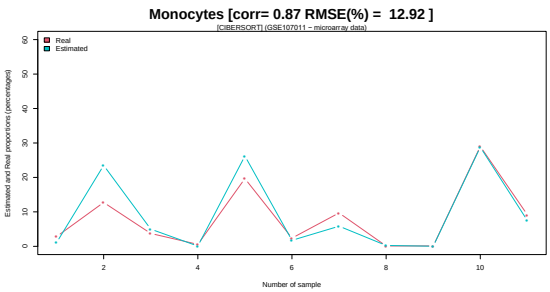
CIBERSORT

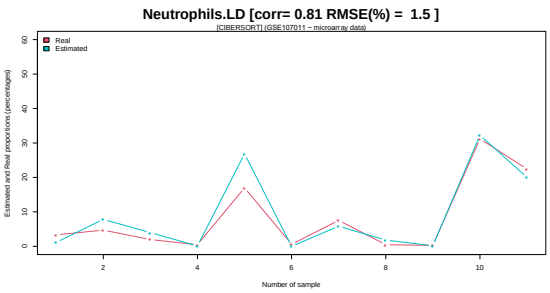


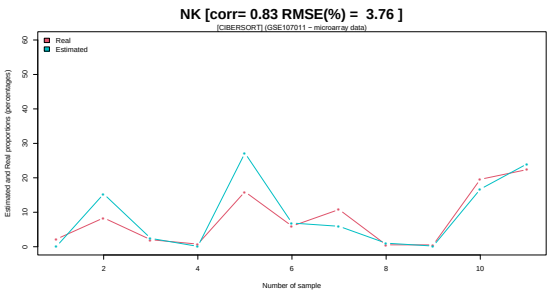


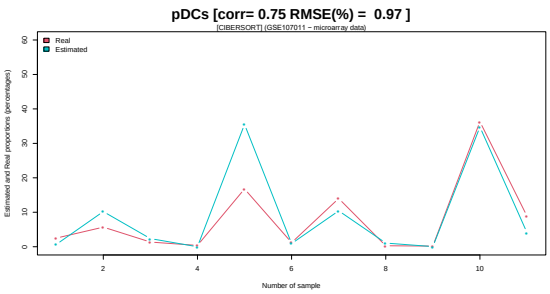


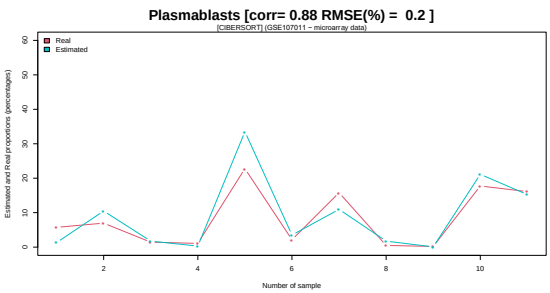


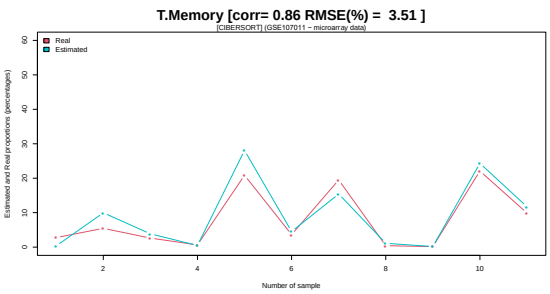












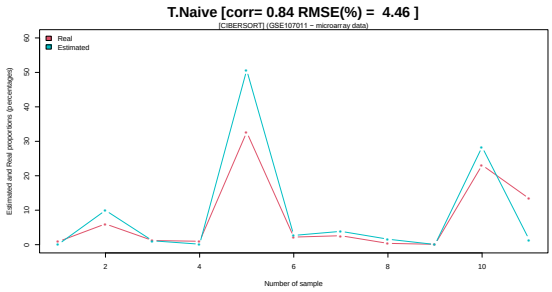
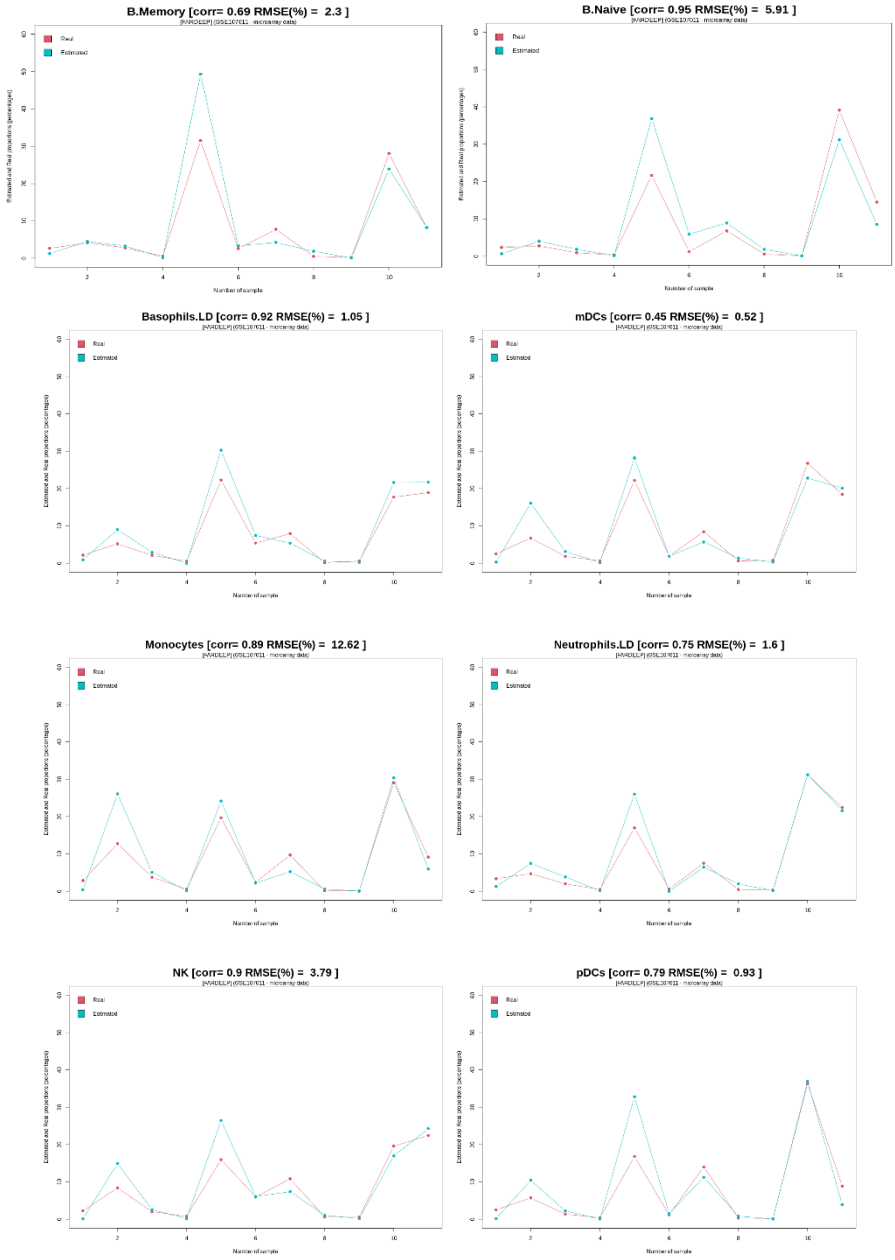


Gráfico S23. Representación de los tipos celulares (*cell signature plot*) estimados por CIBERSORT en la mezcla GSE106898 (datos de expresión obtenidos mediante microarrays).

FARDEEP



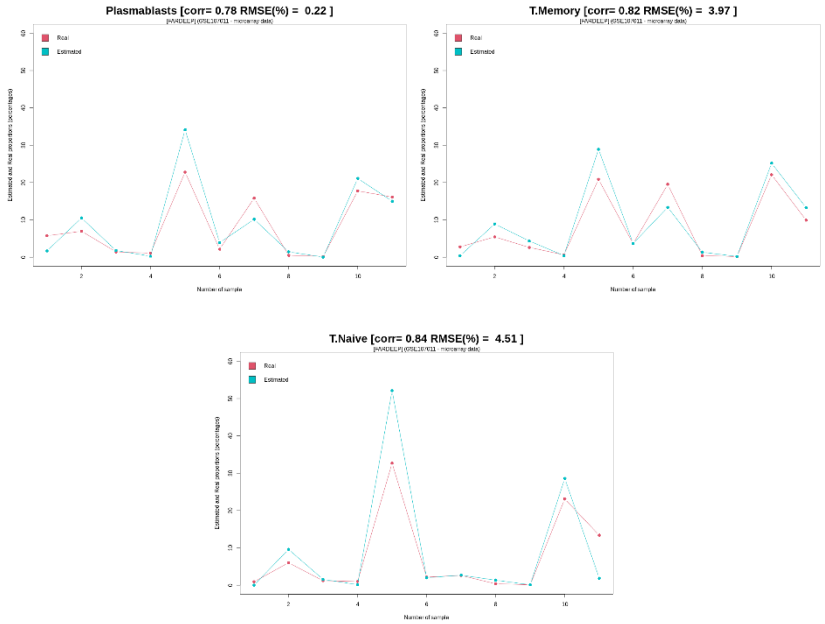
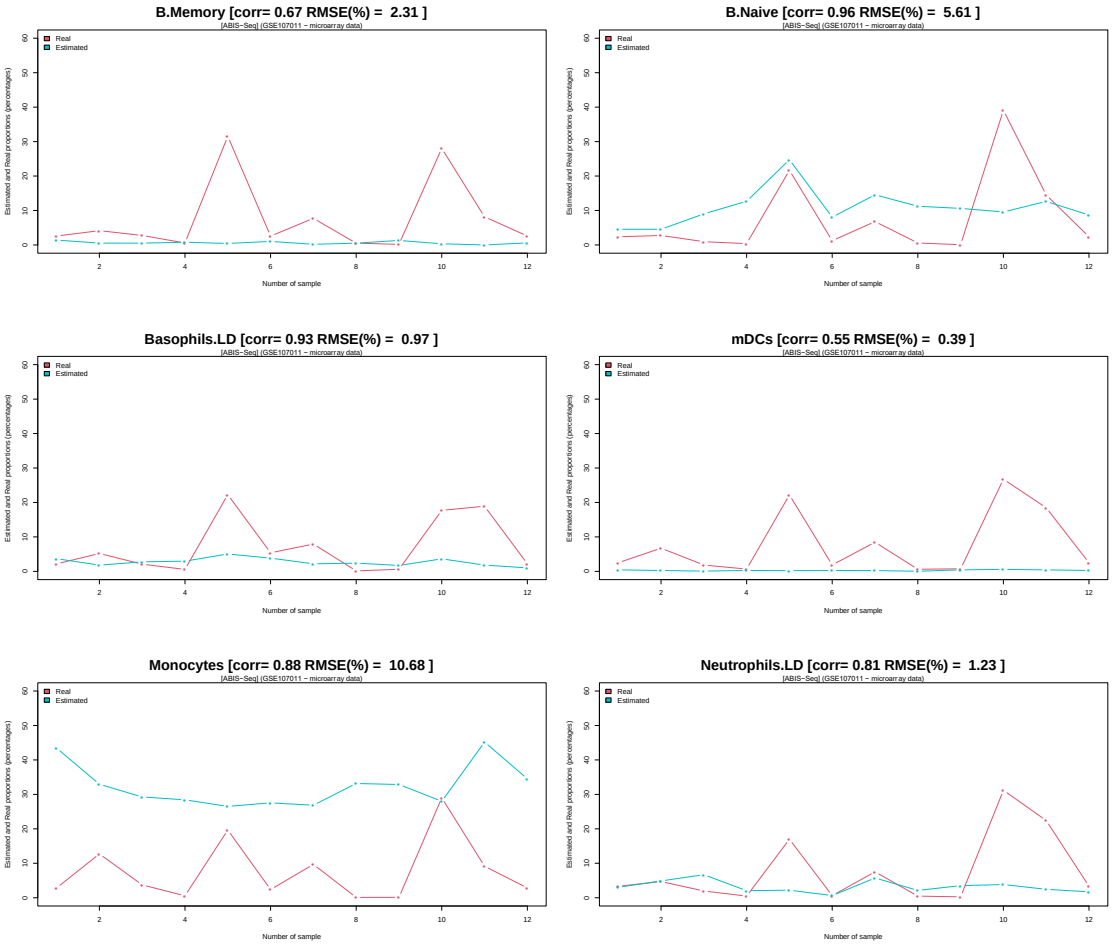


Gráfico S24. Representación de los tipos celulares (*cell signature plot*) estimados por FARDEEP en la mezcla GSE106898 (datos de expresión de microarrays).

ABIS



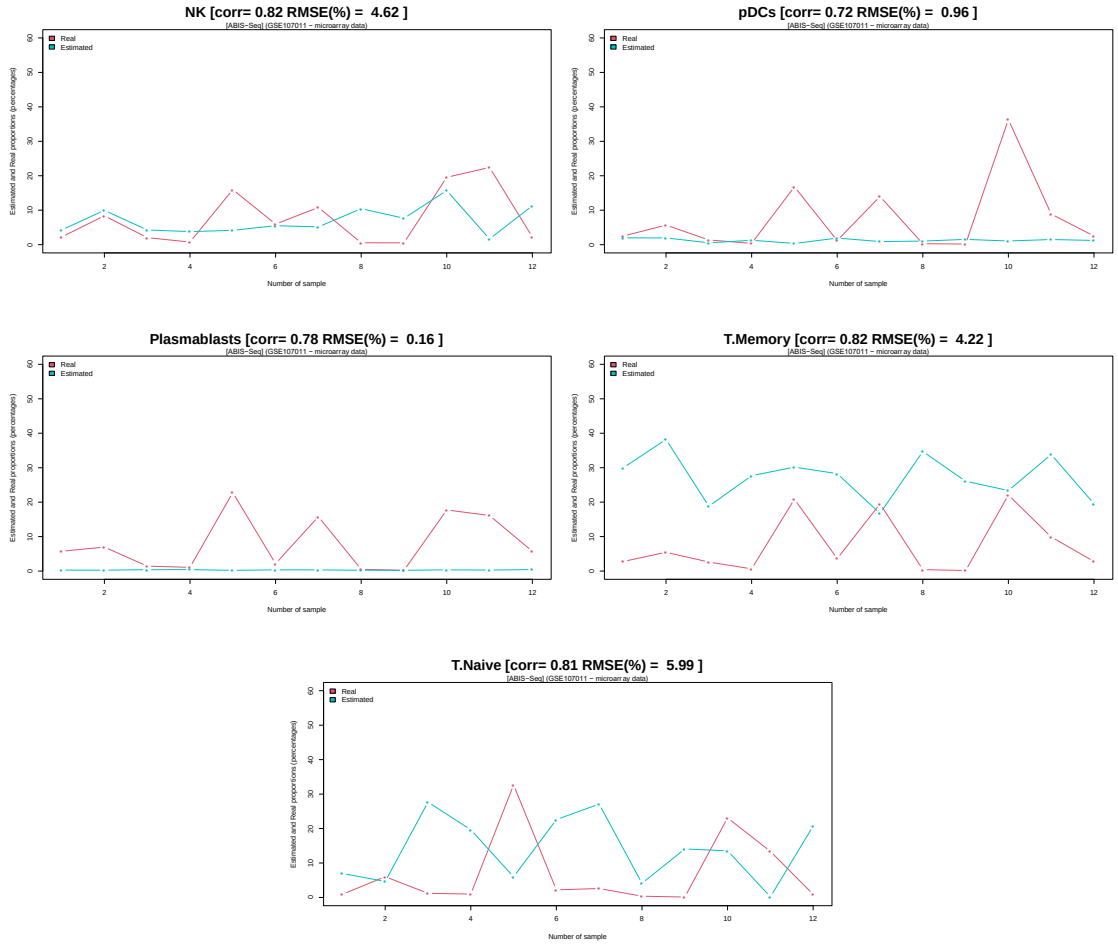
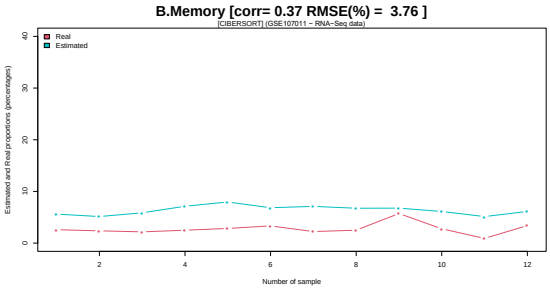
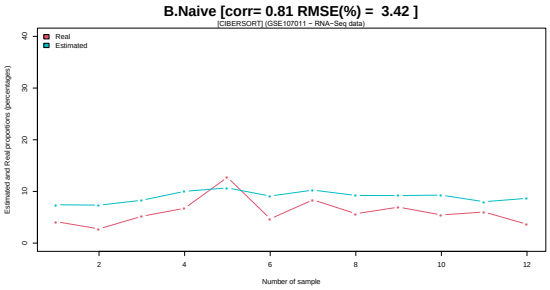


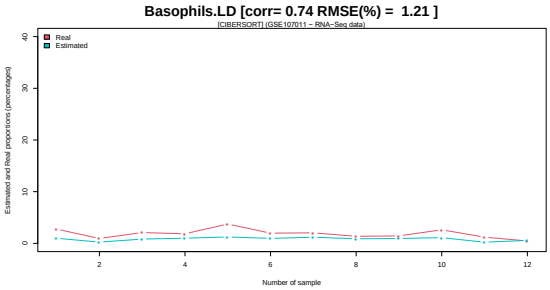
Gráfico S25. Representación de los tipos celulares (*cell signature plot*) estimados por ABIS en la mezcla GSE106898 (datos de expresión obtenidos mediante microarrays).

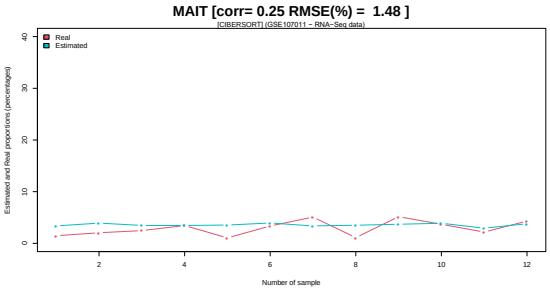
Datos con señal de expresión detectada mediante RNA-Seq (GSE107011)

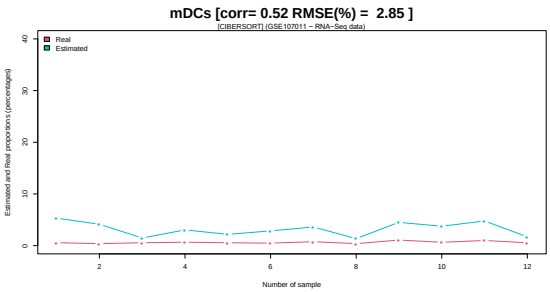
CIBERSORT

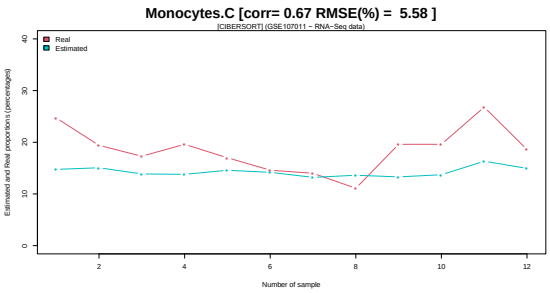


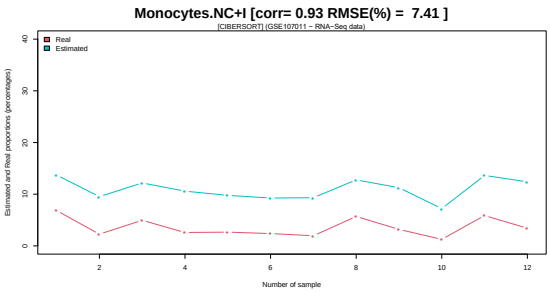


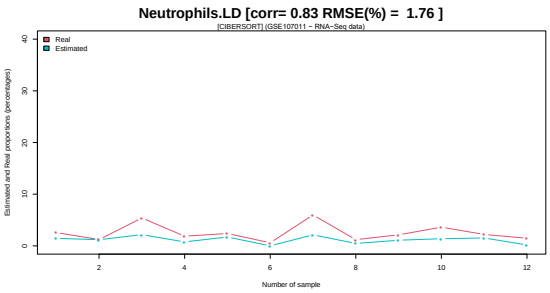


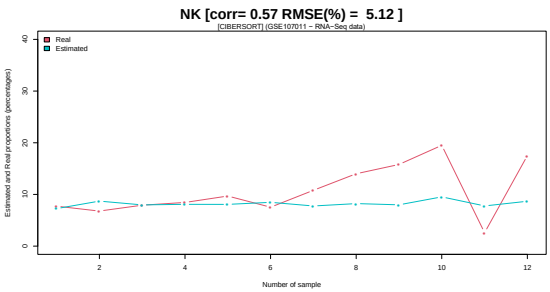


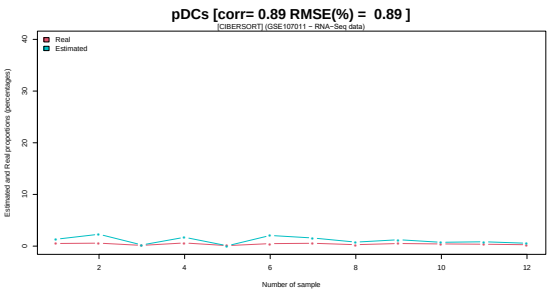












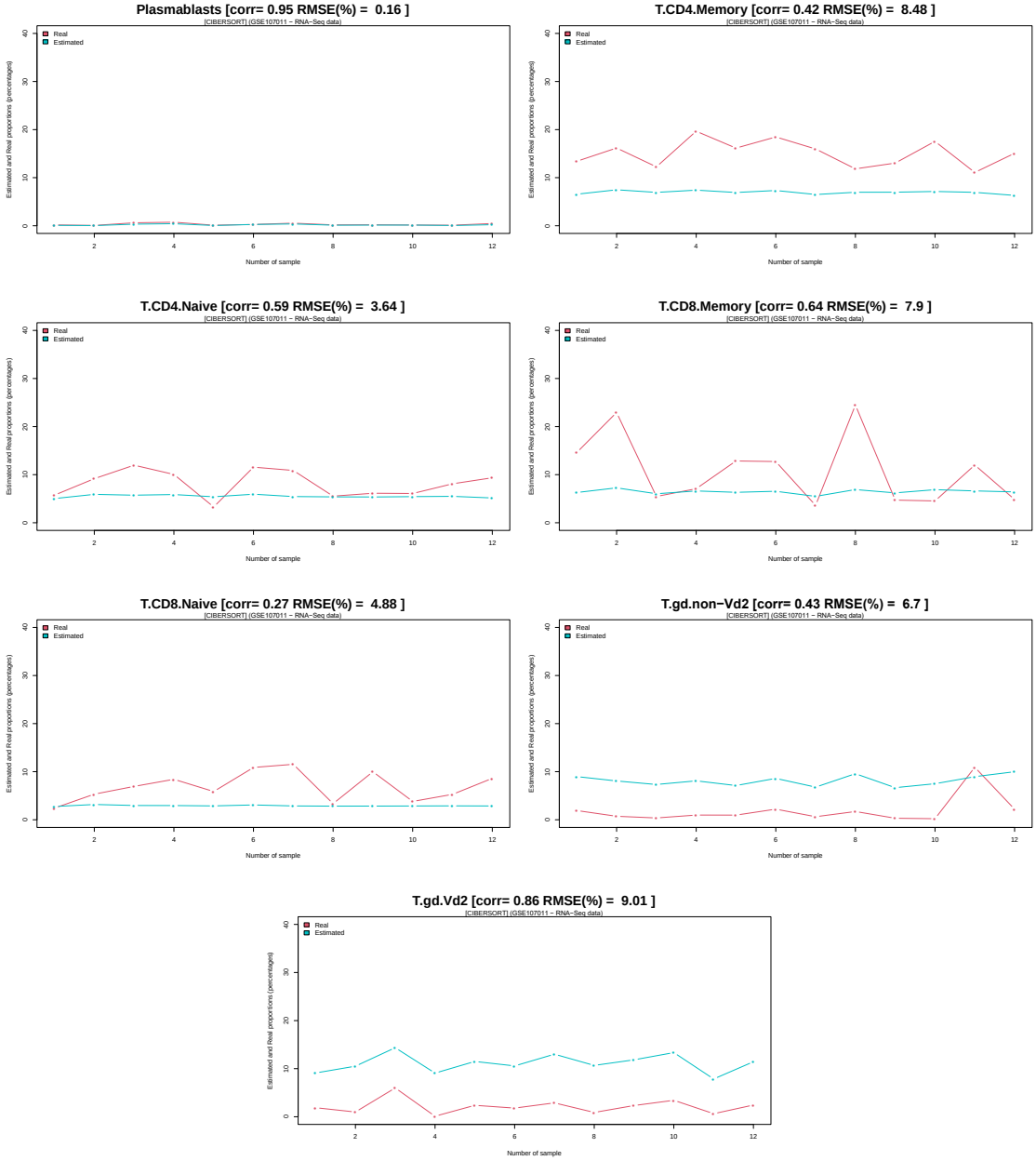
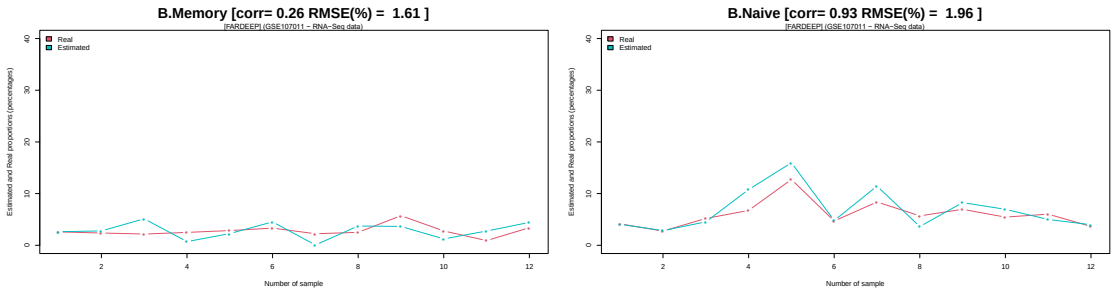
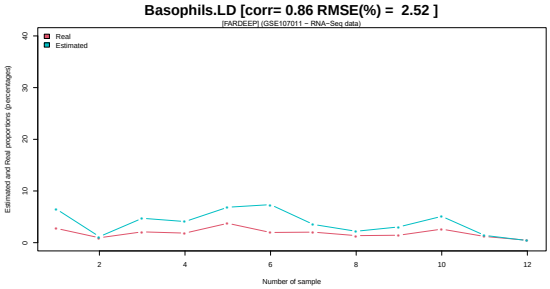
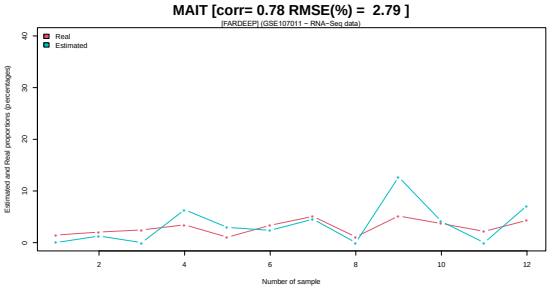


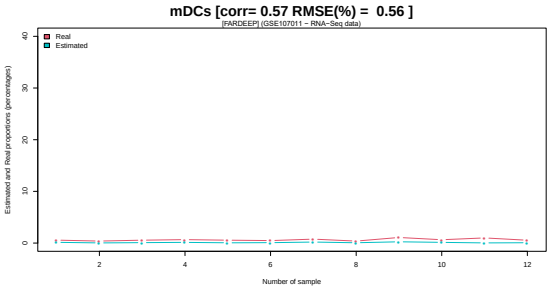
Gráfico S26. Representación de los tipos celulares (*cell signature plot*) estimados por CIBERSORT en la mezcla GSE107011 (datos de expresión obtenidos mediante RNA-Seq).

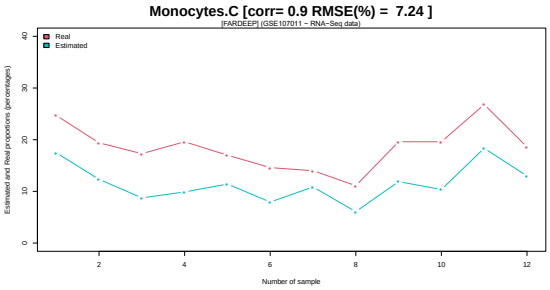
FARDEEP

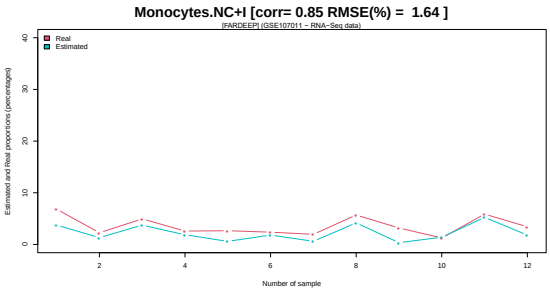


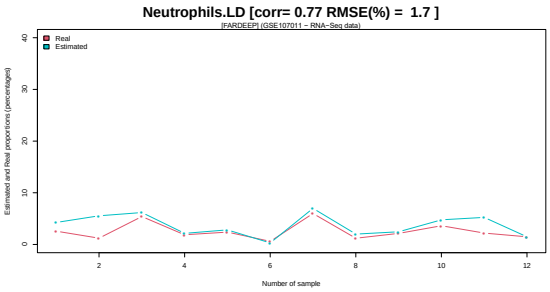

Basophils.LD [corr= 0.86 RMSE(%) = 2.52]
[FARDEEP] (GSE107011 – RNA-Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample

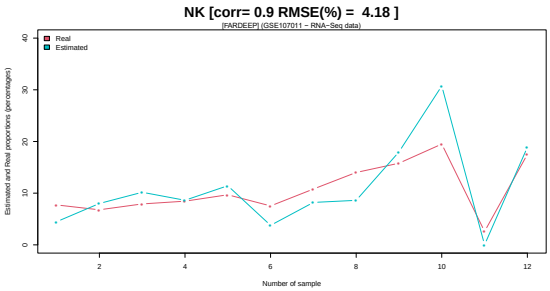

MAIT [corr= 0.78 RMSE(%) = 2.79]
[FARDEEP] (GSE107011 – RNA-Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample

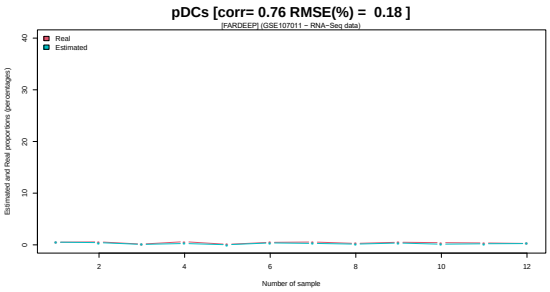

mDCs [corr= 0.57 RMSE(%) = 0.56]
[FARDEEP] (GSE107011 – RNA-Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample

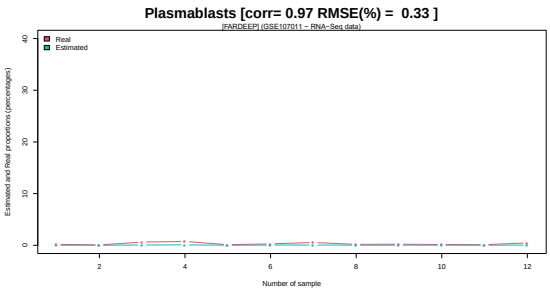

Monocytes.C [corr= 0.9 RMSE(%) = 7.24]
[FARDEEP] (GSE107011 – RNA-Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample

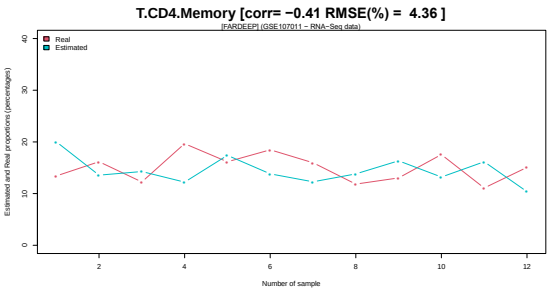

Monocytes.NC+I [corr= 0.85 RMSE(%) = 1.64]
[FARDEEP] (GSE107011 – RNA-Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample


Neutrophils.LD [corr= 0.77 RMSE(%) = 1.7]
[FARDEEP] (GSE107011 – RNA-Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample


NK [corr= 0.9 RMSE(%) = 4.18]
[FARDEEP] (GSE107011 – RNA-Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample


pDCs [corr= 0.76 RMSE(%) = 0.18]
[FARDEEP] (GSE107011 – RNA-Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample


Plasmablasts [corr= 0.97 RMSE(%) = 0.33]
[FARDEEP] (GSE107011 – RNA-Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample


T.CD4.Memory [corr= −0.41 RMSE(%) = 4.36]
[FARDEEP] (GSE107011 – RNA-Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample

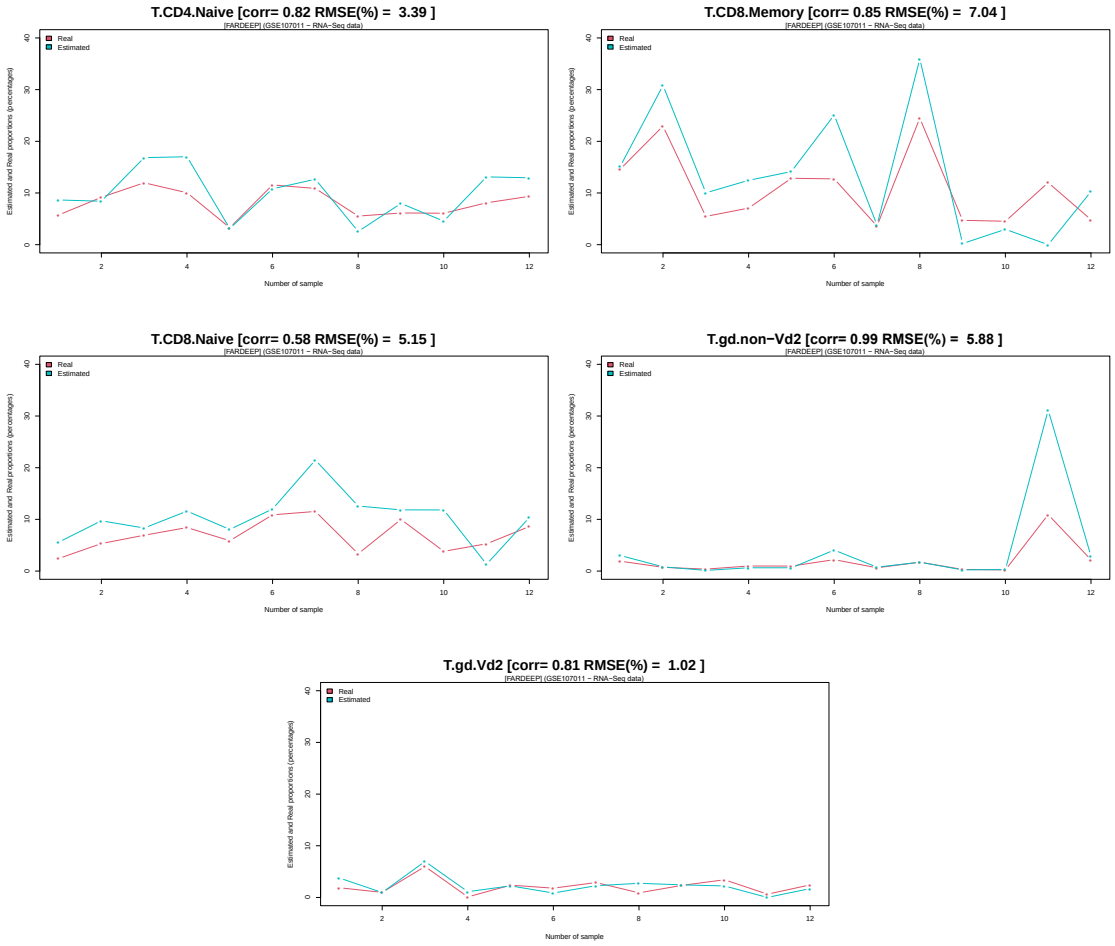
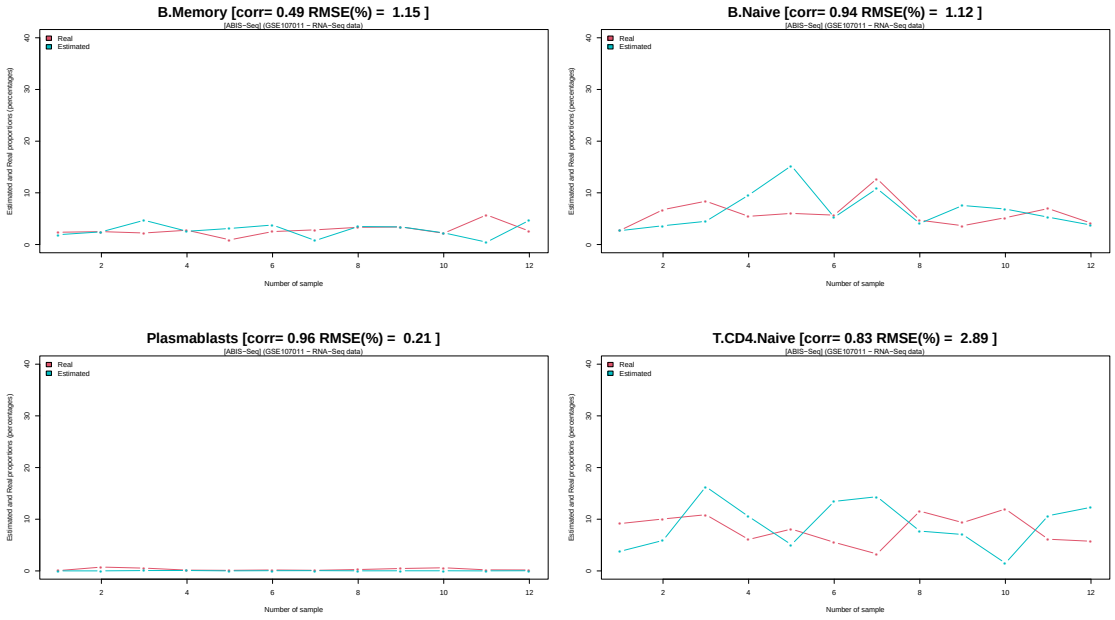
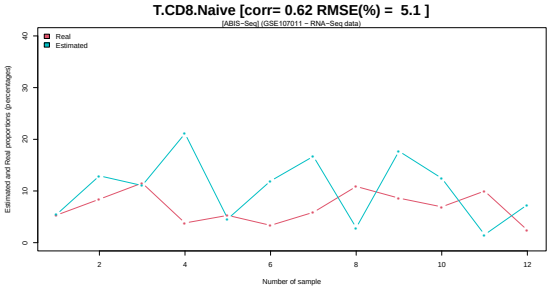
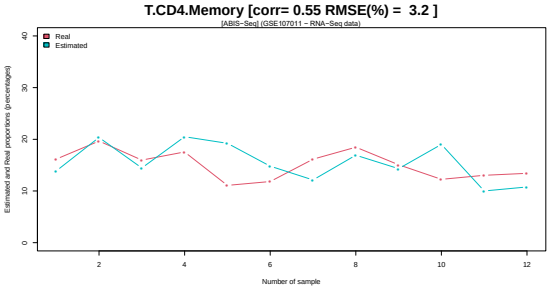


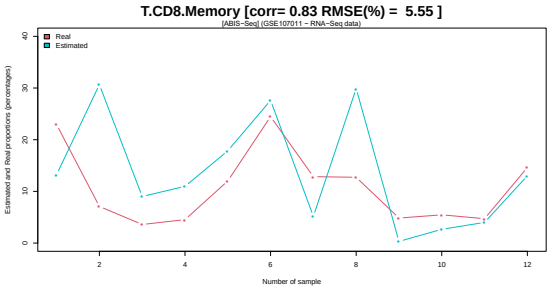
Gráfico S27. Representación de los tipos celulares (*cell signature plot*) estimados por FARDEEP en la mezcla GSE107011 (datos de expresión obtenidos mediante RNA-Seq).

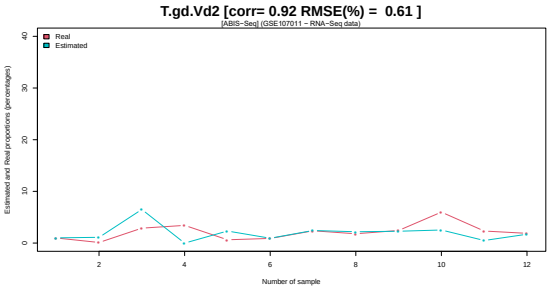
ABIS

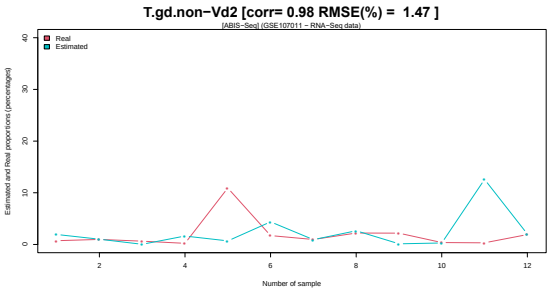


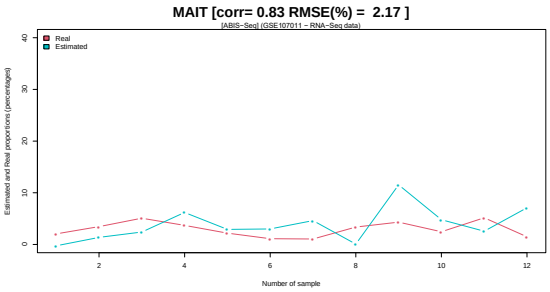

T.CD8.Naive [corr= 0.62 RMSE(%) = 5.1]
[ABIS−Seq] (GSE107011 − RNA−Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample

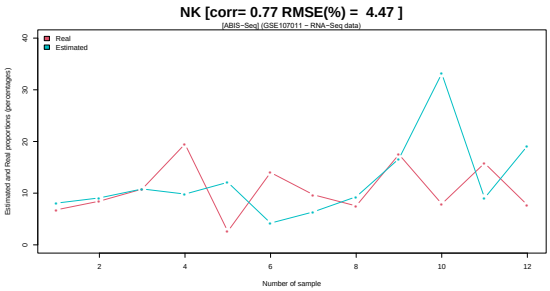

T.CD4.Memory [corr= 0.55 RMSE(%) = 3.2]
[ABIS−Seq] (GSE107011 − RNA−Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample

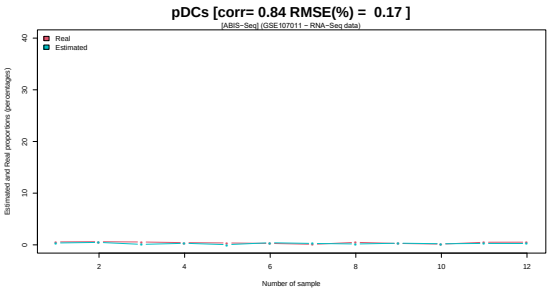

T.CD8.Memory [corr= 0.83 RMSE(%) = 5.55]
[ABIS−Seq] (GSE107011 − RNA−Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample

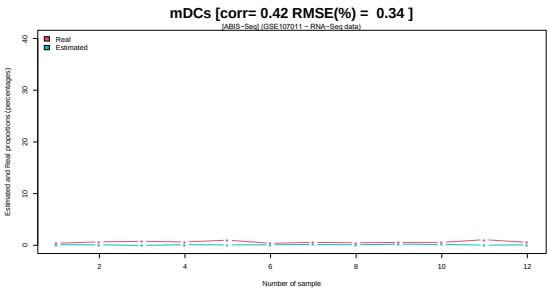

T.gd.Vd2 [corr= 0.92 RMSE(%) = 0.61]
[ABIS−Seq] (GSE107011 − RNA−Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample

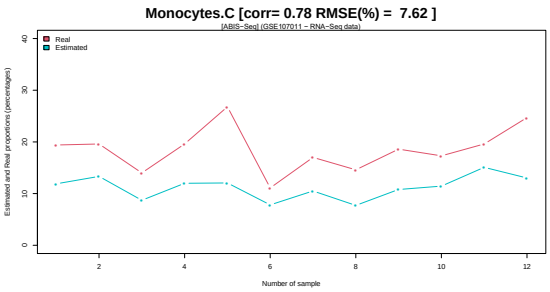

T.gd.non−Vd2 [corr= 0.98 RMSE(%) = 1.47]
[ABIS−Seq] (GSE107011 − RNA−Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample


MAIT [corr= 0.83 RMSE(%) = 2.17]
[ABIS−Seq] (GSE107011 − RNA−Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample


NK [corr= 0.77 RMSE(%) = 4.47]
[ABIS−Seq] (GSE107011 − RNA−Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample


pDCs [corr= 0.84 RMSE(%) = 0.17]
[ABIS−Seq] (GSE107011 − RNA−Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample


mDCs [corr= 0.42 RMSE(%) = 0.34]
[ABIS−Seq] (GSE107011 − RNA−Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample


Monocytes.C [corr= 0.78 RMSE(%) = 7.62]
[ABIS−Seq] (GSE107011 − RNA−Seq data)
Real
Estimated
Estimated and Real proportions (percentages)
Number of sample

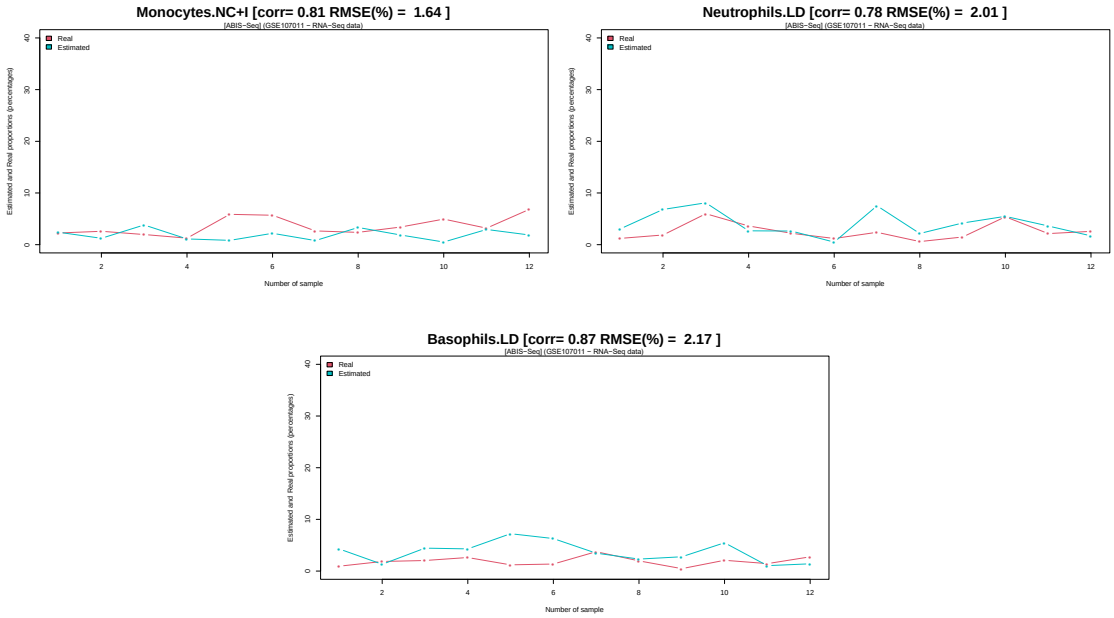
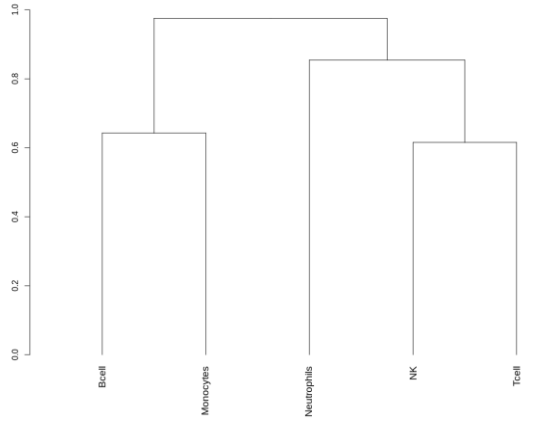


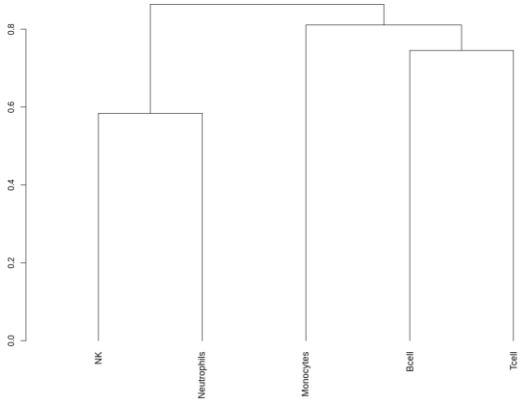
Gráfico S28. Representación de los tipos celulares (*cell signature plot*) estimados por ABIS en la mezcla GSE107011 (datos de expresión obtenidos mediante RNA-Seq).

DENDOGRAMAS

MATRIZ DE FIRMAS ESTIMADA POR DECONICA



MATRIZ DE FIRMAS ESTIMADA POR LINSEED



MATRIZ DE FIRMAS LM22

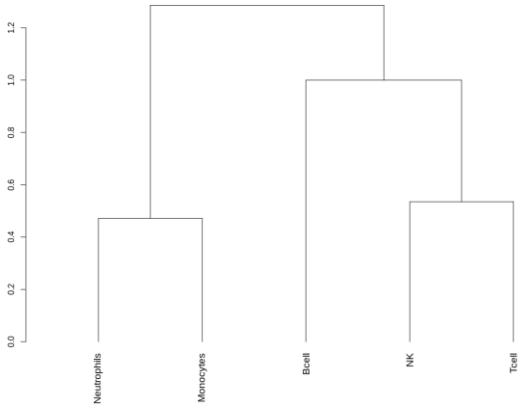


Gráfico S29. Dendograma que muestra la clasificación de los tipos celulares en cada matriz de firmas.

HEATMAPS

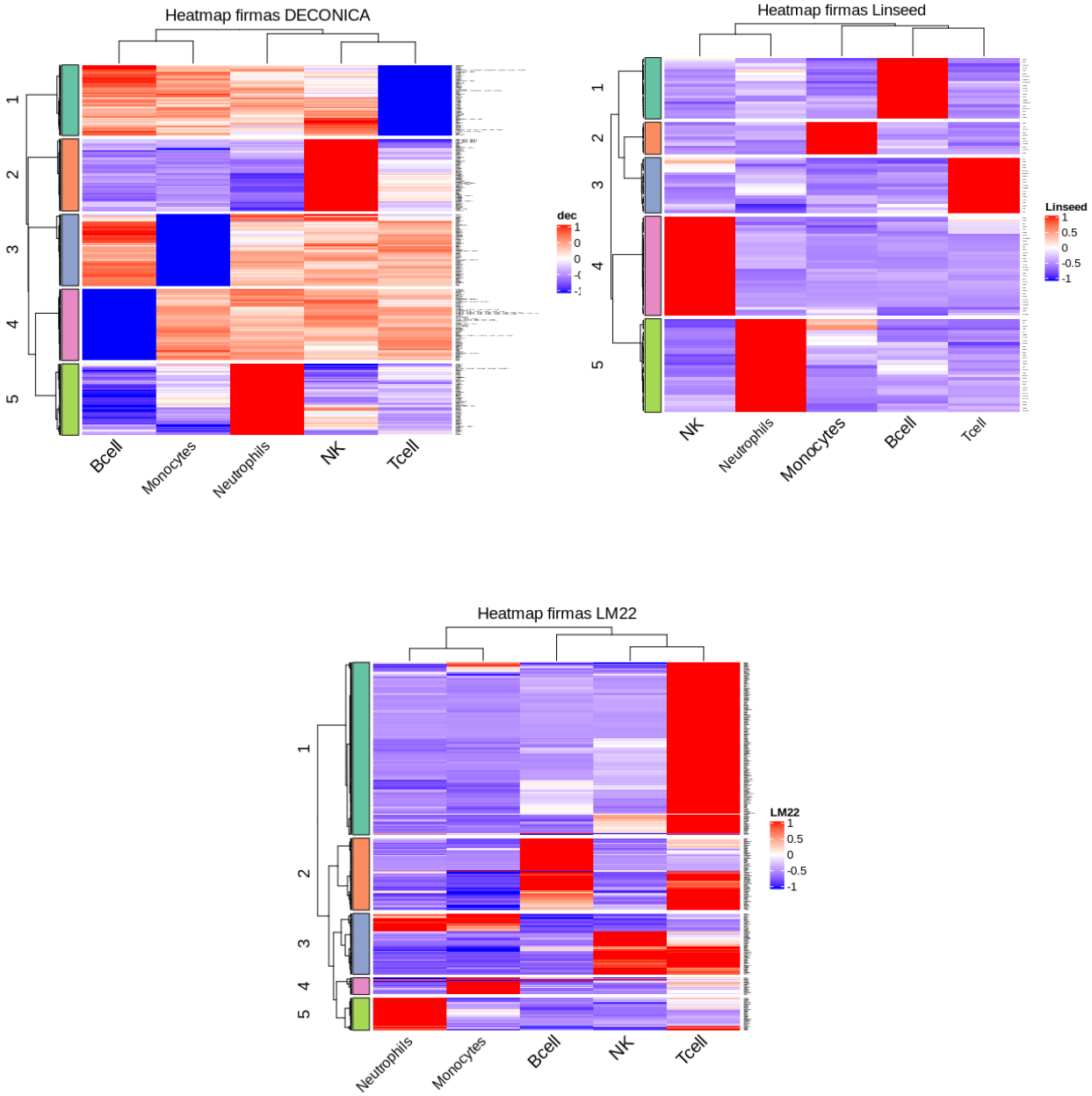


Gráfico S30. *Heatmap* que representa la expresión de los genes marcadores en las matrices de firmas.

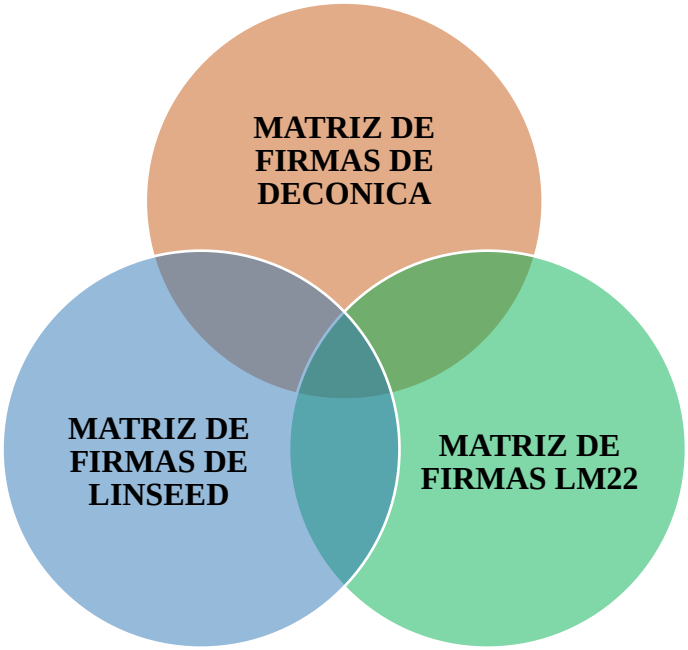


Gráfico S31. Representación del número de genes coincidentes entre las tres matrices estudiadas.