



UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

# **Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq**

**Jonathan Carvajal Veloza**

Universidad Nacional de Colombia

Facultad de Ingeniería

Bogotá, Colombia

2025

# **Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq**

**Jonathan Carvajal Veloza**

Trabajo final de maestría presentado como requisito parcial para optar al título de:

**Magíster en Bioinformática – Modalidad profundización**

Directora:

Luz Dary Gutiérrez Castañeda. Ph.D.,

Codirector:

César Payán Gómez Ph.D.,

Línea de Investigación:

Bioinformática funcional y estructural

Grupo de Investigación:

Bioinformática y biología de sistemas

Universidad Nacional de Colombia

Facultad de Ingeniería

Bogotá, Colombia

2025

*Dedicatoria*

*A mi mamá, a mi papá, a mi familia, a mis amigos, a mis hermanos gatunos y perrunos, y a todas las personas que, de una u otra forma, han creído en mí, me han apoyado, guiado, inspirado y animado a seguir mi camino.*

## Agradecimientos

Agradezco profundamente a los directores de este trabajo, Luz Dary Gutiérrez y César Payán por su confianza, orientación y acompañamiento desde el primer momento en el que planteamos este proyecto.

A los profesores de la Maestría en Bioinformática de la Universidad Nacional de Colombia, cuyas enseñanzas en las asignaturas cursadas me brindaron las bases necesarias para profundizar en los temas desarrollados en este trabajo.

A los docentes y auxiliares del Instituto de Ciencias Básicas de la Fundación Universitaria de Ciencias de la Salud – FUCS por su guía y apoyo constante desde el primer día.

A mis amigos y compañeros del semillero de Ciencias Básicas en Salud por su compañía y motivación continua.

A mi familia y amigos.

A mí, por llegar hasta aquí.

## Resumen

### **Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq**

El conocimiento del microambiente tumoral ha mostrado un papel importante en pronóstico y respuesta a tratamientos del cáncer. El cáncer de ovario seroso de alto grado (HGSOC) se caracteriza por su heterogeneidad, quimioresistencia y mal pronóstico. La caracterización de su composición celular mediante métodos experimentales es compleja y costosa. Mediante análisis de deconvolución es posible estimar el contenido celular a partir de datos genómicos de tejido completo. El objetivo de este trabajo fue comparar el desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de HGSOC a partir de datos de RNA-seq. Se construyeron *pseudobulks* a partir de scRNA-seq de HGSOC y se evaluaron con los métodos de deconvolución CIBERSORTx, TOAST, Linseed y CDSeqR. Posteriormente, se analizaron datos *bulk* de TCGA-OV (n = 150) considerando únicamente tumores serosos primarios, estadios FIGO IIIC/IV, pacientes blancos, mayores de 20 años y con supervivencia > 365 días. En la comparación de *pseudobulk*, CIBERSORTx obtuvo la mayor precisión (r=0.91, MAE=0.039, RMSE=0.061), seguido de TOAST (r=0.63, MAE=0.065, RMSE=0.091). En el análisis de supervivencia, TOAST identificó que a mayor proporción de células plasmáticas mayor supervivencia global, mientras que a una mayor proporción de células T y NK menor supervivencia. Con CIBERSORTx se encontró que a mayor proporción de células plasmáticas mayor supervivencia. En conclusión, CIBERSORTx fue el método más robusto y preciso, mientras que TOAST ofreció mayor sensibilidad para la asociación con supervivencia. Los métodos reference-free resultaron poco confiables en tumores heterogéneos.

**Palabras clave:** deconvolución, HGSOC, CIBERSORTx, TOAST, TCGA, microambiente tumoral, supervivencia.

## Abstract

### **Comparison of deconvolution methods for identifying cellular composition and its association with survival in high-grade serous ovarian cancer using RNA-seq data**

Knowledge of the tumor microenvironment has shown an important role in cancer prognosis and treatment response. High-grade serous ovarian cancer (HGSOC) is characterized by heterogeneity, chemoresistance, and poor prognosis. The characterization of its cellular composition through experimental methods is complex and costly. From a bioinformatics perspective, it is possible to use existing bulk genomic data available in public repositories. The objective of this study was to compare the performance of deconvolution methods for identifying cellular composition and their association with survival in HGSOC samples using RNA-seq data. Pseudobulks were constructed from HGSOC scRNA-seq and evaluated with the deconvolution methods CIBERSORTx, TOAST, Linseed, and CDSeqR. Subsequently, bulk RNA-seq data from TCGA-OV (n = 150) were analyzed, considering only primary serous tumors, FIGO stage III/IV, white patients, age > 20 years, and overall survival > 365 days. In pseudobulk comparisons, CIBERSORTx achieved the highest accuracy (r = 0.91, MAE = 0.039, RMSE = 0.061), followed by TOAST (r = 0.63, MAE = 0.065, RMSE = 0.091). In the survival analysis, TOAST identified that a higher proportion of plasma cells was associated with better overall survival, whereas higher proportions of T and NK cells were associated with worse outcomes. With CIBERSORTx, a higher proportion of plasma cells was also associated with improved survival. In conclusion, CIBERSORTx was the most robust and accurate method, whereas TOAST showed greater sensitivity in detecting survival associations. Reference-free methods proved unreliable in heterogeneous tumors.

**Keywords:** deconvolution, HGSOC, CIBERSORTx, TOAST, TCGA, tumor microenvironment, survival.

Este Trabajo Final de maestría fue calificado en **SEPTIEMBRE** de 2025 por el siguiente evaluador:

**FABIO AUGUSTO GONZÁLEZ OSORIO**  
Profesor del Departamento de Ingeniería de Sistemas e Industrial  
Facultad de Ingeniería  
Universidad Nacional de Colombia, Sede Bogotá

# Contenido

**Lista de figuras** IX

**Lista de tablas** XII

## **1 Introducción**

## **2 Marco teórico**

2.1	Cáncer de ovario .....	1
2.2	Microambiente tumoral .....	2
2.3	Bioinformática y cáncer .....	4
2.4	Análisis computacional de deconvolución .....	5

## **3 Objetivos**

3.1	Objetivo general .....	1
3.2	Objetivos específicos .....	1

## **4 Capítulo 1: Determinación del desempeño de métodos de deconvolución a partir de datos de scRNA-seq de muestras de HGSOC** 3

4.1	Introducción .....	3
4.2	Metodología .....	4
4.2.1	Procesamiento de datos .....	4
4.2.2	Generación de <i>pseudobulks</i> .....	5
4.2.3	Deconvolución .....	5
4.2.4	Generación de GEPs y anotación de tipos celulares .....	7
4.2.5	Evaluación del desempeño de los algoritmos de deconvolución .....	8

4.2.6	Flujo computacional .....	8
4.3	Resultados.....	9
4.3.1	Generación de <i>pseudobulks</i> .....	9
4.3.2	Deconvolución.....	12
4.3.3	Evaluación del desempeño de los algoritmos de deconvolución .....	15
4.4	Discusión .....	26
<b>5</b>	<b>Capítulo 2: Estimación de proporciones celulares mediante deconvolución y su asociación con la supervivencia en cáncer de ovario seroso de alto grado a partir de datos de TCGA</b>	<b>30</b>
5.1	Introducción.....	30
5.2	Metodología.....	31
5.2.1	Procesamiento de datos.....	31
5.2.2	Deconvolución.....	31
5.2.3	Análisis de supervivencia .....	31
5.3	Resultados .....	32
5.3.1	Deconvolución.....	32
5.3.2	Análisis de supervivencia .....	35
5.4	Discusión.....	40
<b>6</b>	<b>Conclusiones y recomendaciones</b>	<b>44</b>
6.1	Conclusiones .....	44
6.2	Recomendaciones .....	45

## Lista de figuras

	Pág.
<b>Figura 4-1. Composición celular de 4 escenarios (<i>even, realistic, sparse y weighted</i>) generados a partir de <i>pseudobulks simulados</i> con SimBu para cada <i>dataset</i> analizado.</b> En cada escenario se representan 50 muestras simuladas (columnas), con la altura proporcional de cada barra. Cada color representa un tipo celular. ....	9
<b>Figura 4-2. Proporciones celulares estimadas mediante deconvolución con TOAST en <i>pseudobulks simulados</i> bajo los escenarios <i>even, realistic, sparse y weighted</i>.</b> En cada escenario se representan las proporciones estimadas para 50 muestras (columnas), con la altura proporcional de cada barra indicando la fracción estimada para cada tipo celular. Se muestra la proporción celular para cada dataset.....	13
<b>Figura 4-3. Comparación del desempeño de métodos de deconvolución en métricas globales (Pearson, MAE y RMSE).</b> Diagrama de cajas de la distribución de valores de desempeño para CDSegR, Linseed, TOAST y CIBERSORTx en términos de MAE, correlación de Pearson y RMSE. Cada caja representa la variabilidad del método a través de múltiples escenarios de <i>pseudobulks simulados</i> , los puntos corresponden a valores individuales. ....	15
<b>Figura 4-4. Comparación del desempeño de métodos de deconvolución en métricas por tipo celular (RMSE, MAE y Pearson).</b> <i>Mapa de calor</i> que muestran el desempeño promedio de CDSegR, Linseed, TOAST y CIBERSORTx en distintos tipos celulares, evaluado mediante correlación de Pearson, RMSE y MAE. Las filas corresponden a métodos de deconvolución y las columnas a tipos celulares. Los gradientes de color indican concordancia (Pearson) o magnitud del error (RMSE, MAE).....	17
<b>Figura 4-5. Comparación del desempeño de métodos de deconvolución en función de métricas (Pearson, MAE y RMSE) por tipo celular.</b> Distribución de valores de RMSE, MAE y correlación de Pearson para cada tipo celular, comparando cuatro métodos de deconvolución: CDseqR, Linseed, TOAST y CIBERSORTx. Cada caja resume la	

variabilidad entre simulaciones, mientras que los puntos representan valores individuales.

..... 19

**Figura 4-6. Comparación de proporciones celulares reales vs. estimadas con TOAST en cuatro escenarios de simulación (*even, realistic, sparse y weighted*).** Relación entre proporciones celulares reales y estimadas mediante TOAST en pseudobulks simulados bajo los escenarios *even, realistic, sparse y weighted*. Cada punto corresponde a un tipo celular en una muestra simulada. Se incluyen los valores de correlación de Pearson (*r*), RMSE y MAE para cada escenario. ....21

**Figura 4-7. Comparación de proporciones celulares reales vs. estimadas con Linseed en cuatro escenarios de simulación (*even, realistic, sparse y weighted*).** Relación entre proporciones celulares reales y estimadas mediante Linseed en pseudobulks simulados bajo los escenarios *even, realistic, sparse y weighted*. Cada punto corresponde a un tipo celular en una muestra simulada. Se incluyen los valores de correlación de Pearson (*r*), RMSE y MAE para cada escenario. ....22

**Figura 4-8. Comparación de proporciones celulares reales vs. estimadas con CDSeqR en cuatro escenarios de simulación (*even, realistic, sparse y weighted*).** Relación entre proporciones celulares reales y estimadas mediante CDSeqR en pseudobulks simulados bajo los escenarios *even, realistic, sparse y weighted*. Cada punto corresponde a un tipo celular en una muestra simulada. Se incluyen los valores de correlación de Pearson (*r*), RMSE y MAE para cada escenario. ....24

**Figura 4-9. Comparación de proporciones celulares reales vs. estimadas con CIBERSORTx en cuatro escenarios de simulación (*even, realistic, sparse y weighted*).** Relación entre proporciones celulares reales y estimadas mediante CIBERSORTx en pseudobulks simulados bajo los escenarios *even, realistic, sparse y weighted*. Cada punto corresponde a un tipo celular en una muestra simulada. Se incluyen los valores de correlación de Pearson (*r*), RMSE y MAE para cada escenario.....25

**Figura 5-1. Proporciones celulares estimadas mediante deconvolución con TOAST, Linseed, CDSeqR y CIBERSORTx en muestras de pacientes obtenidas de TCGA.** Proporciones celulares estimadas con TOAST, Linseed, CDSeqR y CIBERSORTx a partir de 150 muestras de pacientes con cáncer de ovario (TCGA-OV). Cada barra corresponde a una muestra individual, y la altura proporcional de los segmentos representa la fracción estimada de cada tipo celular en la mezcla..... 33

---

**Figura 5-2. Comparación de proporciones celulares estimadas por TOAST, Linseed, CDseqR y CIBERSORTx entre pacientes vivos (verde) y fallecidos (rojo) de la cohorte obtenida de TCGA-OV.** Cada panel corresponde a un método de deconvolución, y las cajas representan la distribución de fracciones estimadas para cada tipo celular..... 35

**Figura 5-3. Curvas de Kaplan–Meier para tipos celulares con asociaciones significativas de supervivencia estimadas por TOAST en TCGA-OV.** Curvas de Kaplan–Meier de la cohorte TCGA-OV estratificadas por cuartiles (Q1 = baja, Q4 = alta proporción) para los tipos celulares cuya proporción estimada mediante TOAST mostró una asociación significativa con la supervivencia global. Se muestran los valores de  $p$  de la prueba *log-rank* y el número de pacientes en riesgo a lo largo del tiempo..... 37

**Figura 5-4. Curva de Kaplan–Meier para el tipo celular con asociación significativa de supervivencia estimado por CIBERSORTx en TCGA-OV.** Curva de Kaplan–Meier de la cohorte TCGA-OV estratificada por cuartiles (Q1 = baja, Q4 = alta proporción) para el tipo celular cuya proporción estimada mediante CIBERSORTx mostró una asociación significativa con la supervivencia global. Se muestran los valores de  $p$  de la prueba *log-rank* y el número de pacientes en riesgo a lo largo del tiempo. .... 38

## Lista de tablas

<b>Tabla 4-1. Composición de <i>pseudobulks</i> simulados SimBu por dataset y escenario.</b> .....	11
<b>Tabla 5-1. Resultados del análisis de supervivencia según proporciones celulares inferidas por distintos métodos de deconvolución en TCGA-OV.</b> Resultados del análisis de supervivencia global mediante curvas de Kaplan–Meier estratificadas por cuartiles extremos (Q1 vs. Q4) de proporciones celulares estimadas con CDSeqR, CIBERSORTx, Linseed y TOAST en la cohorte TCGA-OV. Se muestran los valores de <i>p</i> ( <i>log rank</i> ) para las comparaciones entre grupos. ....	39

# 1 Introducción

El cáncer de ovario (CO) es el cáncer ginecológico con mayor mortalidad en el mundo. Debido a la inespecificidad de sus síntomas, el 80% de CO se diagnostican en estadios tardíos (III y IV) donde los tratamientos son poco efectivos y la probabilidad de supervivencia a cinco años es del 29.2%<sup>1</sup>. El desenlace clínico desfavorable es atribuido principalmente a la compleja y heterogénea naturaleza del microambiente tumoral (TME). El microambiente tumoral es definido como el conjunto de tipos celulares que rodean el tumor e interactúan con él están directamente relacionadas con la agresividad del tumor<sup>2,3</sup>. Los tratamientos oncológicos usados actualmente para CO suelen centrarse en eliminar las células tumorales, sin considerar el microambiente tumoral (TME), cuya heterogeneidad contribuye a la resistencia terapéutica y a la recurrencia tumoral y dificulta el diseño de intervenciones eficaces en CO.

El TME está conformado por componentes celulares y no celulares. Entre los componentes celulares se encuentran las células tumorales, células madre tumorales, células endoteliales, fibroblastos asociados a cáncer (CAFs), pericitos y diferentes tipos de células inmunes. En cuanto a los componentes no celulares se encuentran las moléculas señalizadoras que secretan algunos de estos tipos celulares como las quimiocinas, citoquinas, metaloproteinasas, y la matriz extracelular<sup>4,5</sup>. Como producto de esta interacción las células neoplásicas son capaces de potenciar los procesos conocidos como “sellos” o “*hallmarks*” del cáncer (invasión y metástasis, angiogénesis, resistencia a la muerte celular y proliferación)<sup>2,6</sup>. Además, se ha identificado que el TME juega un papel importante en el desarrollo, respuesta a quimioterapéuticos y en la supervivencia global de las pacientes<sup>2</sup>. Por ejemplo, tumores con altas proporciones de células con actividad pro tumoral como los CAFs, los macrófagos con fenotipo M2 y las células supresoras derivadas de mieloides (MDSCs), se han asociado con peor pronóstico, ya que promueven la progresión tumoral, la metástasis y la resistencia a la quimioterapia<sup>7</sup>. Por otra parte, tumores con mayor infiltración de linfocitos T CD8+ citotóxicos se han asociado con una respuesta antitumoral más efectiva y por ende con una mayor supervivencia<sup>7</sup>.

Varias aproximaciones han sido realizadas con el objetivo de identificar los componentes del TME, uno de ellos es a través de la identificación de perfiles transcriptómicos mediante la secuenciación del ARN (RNA)<sup>8</sup>. Aunque el RNA-seq de tejido completo (*bulk* RNA-seq) ha permitido clasificar el cáncer de ovario de alto grado (HGSOC) en subtipos moleculares, esta aproximación diluye las señales transcriptómicas de las poblaciones celulares minoritarias. se pierde información sobre la heterogeneidad funcional de las células del TME y su contribución específica a la progresión tumoral y la resistencia a la terapia <sup>9,10</sup>. Para superar esta limitación se han desarrollado tecnologías experimentales capaces de analizar células a nivel individual, como la secuenciación de RNA de célula única (scRNA-seq), la microdissección por captura láser (LCM) o la citometría de flujo con clasificación celular (FACS). Si bien estas técnicas permiten caracterizar la heterogeneidad con alta resolución y sin necesidad de información previa, su elevado costo y complejidad técnica las hacen poco viables para estudios de gran escala o para aplicaciones clínicas rutinarias <sup>9,11,12</sup>.

En este contexto, los métodos computacionales de deconvolución son una alternativa bioinformática costo-efectiva<sup>10</sup>. Estas herramientas infieren la abundancia de distintos tipos celulares a partir de datos de expresión génica de muestras heterogéneas, aprovechando la disponibilidad de grandes volúmenes de transcriptomas de tejido completo en los repositorios públicos<sup>9</sup>. Además de estimar proporciones celulares, algunos métodos permiten reconstruir perfiles de expresión específicos para cada tipo celular. Algoritmos como CIBERSORTx y MuSiC han demostrado un desempeño robusto en la estimación de proporciones celulares en distintos escenarios biológicos <sup>13,14</sup>. Por lo tanto, el objetivo de este trabajo es comparar el rendimiento de diversos métodos de deconvolución en la identificación de la composición celular y explorar su asociación con la supervivencia en cáncer de ovario seroso de alto grado, utilizando datos de RNA-seq.

## 2 Marco teórico

### 2.1 Cáncer de ovario

El cáncer de ovario (CO) es el cáncer ginecológico con mayor mortalidad en el mundo. De acuerdo con Globocan el CO en 2022, la carga global de cáncer de ovario se estimó en aproximadamente 324 398 casos nuevos y 206 839 muertes<sup>15</sup>. En Colombia, para el año 2022 se registraron 2253 casos nuevos de cáncer de ovario causando 1445 muertes en ese mismo año por esta patología<sup>16</sup>. El subtipo histológico más frecuente es el carcinoma seroso de alto grado (HGSOC) presentándose en un 70% de los tumores epiteliales<sup>17</sup>. El 80% de los diagnósticos del CO se realizan en estadios tardíos (III y IV) donde las pacientes tienen un 29.2% de probabilidad de supervivencia a cinco años<sup>1</sup>.

El CO es una enfermedad heterogénea y es clasificada en tres tipos de acuerdo a la Organización Mundial de la Salud: epitelial, de células germinales y de células estromales, siendo el epitelial el más común con un 90% de los casos<sup>18-20</sup>. A su vez, de acuerdo a la histología, el CO de tipo epitelial se clasifica en: seroso de alto grado (HGSOC), seroso de bajo grado (LGSOC), endometroide (EC), de células claras (CCC) y mucinoso (MC)<sup>18-21</sup>. El HGSOC es el subtipo histológico de CO más frecuente (70% de los tumores epiteliales)<sup>18</sup>.

De acuerdo con la clasificación quirúrgica dada por la Federación Internacional de Ginecología y Obstetricia (FIGO), el CO se clasifica en cuatro estadios (I, II, III y IV). En el estadio I el tumor está confinado a los ovarios o las trompas de Falopio y al fluido peritoneal, en el estadio II la enfermedad está limitada a la cavidad pélvica, en el estadio III el tumor afecta uno o ambos ovarios/trompas con diseminación peritoneal fuera de la pelvis y/o metástasis en ganglios linfáticos retroperitoneales y en el estadio IV se presenta metástasis a distancia fuera de la cavidad peritoneal, estos últimos son tumores de mal pronóstico, donde la paciente no se lleva a cirugía citorrreductora<sup>15</sup>.

En cuanto a las características genéticas, en el HGSOC la carcinogénesis es promovida principalmente por mutaciones en los genes TP53 y BRCA1/2, lo que lleva a una deficiencia en los mecanismos de reparación del ADN por recombinación homóloga<sup>22</sup>.

En cuanto al diagnóstico, aproximadamente dos tercios de los casos de CO son diagnosticados en estadios avanzados (FIGO III-IV), lo que se asocia a tasas de supervivencia a 5 años del 41 % y 29.2%, respectivamente<sup>1</sup>. Mientras que las pacientes diagnósticas den estadios tempranos (FIGO I-II), presentan tasas de supervivencia de 93% y 74%, respectivamente<sup>23</sup>.

## 2.2 Microambiente tumoral

El HGSOC se desarrolla dentro de un TME complejo, compuesto por células epiteliales malignas, diversos subconjuntos de células inmunes, fibroblastos estromales, células endoteliales, adipocitos y matriz extracelular (ECM). Cada uno de estos componentes contribuye a la iniciación tumoral, la progresión, el pronóstico y la respuesta terapéutica<sup>4</sup>. Dentro de los múltiples tipos celulares que pueden conformar el TME del CO algunos de los tipos que se encuentran son:

### **Fibroblastos Asociados al Cáncer (CAFs)**

Son esencialmente fibroblastos normales residentes en el tejido, los cuales experimentan activación y transformación a CAFs por la influencia del ambiente inflamatorio y la hipoxia que se encuentra en el tumor. Los CAFs fomentan la proliferación, invasión y metástasis de las células tumorales por secreción de citoquinas proinflamatorias y factores que remodelan la matriz extracelular. Asimismo, induce la transición epitelio-mesénquima (EMT), inmunosupresión y la quimiorresistencia del tumor<sup>4,24,25</sup>. Una alta proporción de CAFs ha sido asociada con estadios más avanzados del CO y con una menor supervivencia de la enfermedad<sup>4</sup>.

### **Macrófagos Asociados a Tumor (TAMs)**

Los macrófagos son un tipo celular que es reclutado a la cavidad peritoneal durante la progresión del CO donde mantienen una amplia interacción con las células tumorales. Típicamente, los macrófagos son clasificados de acuerdo con su fenotipo como M1 (pro-

inflamatorios anti-tumorales) y M2 (anti-inflamatorios pro-tumorales). Los TAMs predominan en el fenotipo M2 por lo que promueven la invasión, angiogénesis, aumento de la inmunosupresión y quimiorresistencia<sup>24,25</sup>. De acuerdo con lo anterior, una mayor proporción de macrófagos M2 ha sido asociada con mal pronóstico, mientras que una mayor proporción del fenotipo M1 se asocia con una mejor supervivencia global de la enfermedad<sup>25</sup>.

### **Linfocitos T Infiltrantes de Tumor**

Los linfocitos T infiltrantes (TIL, por su sigla en inglés), comprenden principalmente linfocitos T CD8<sup>+</sup> citotóxicos, linfocitos T CD4<sup>+</sup> cooperadores y células T reguladoras (Tregs), aunque también pueden incluir linfocitos B y células NK<sup>4</sup>. Las células T CD8<sup>+</sup> y los subtipos efectores de CD4<sup>+</sup> (especialmente Th1) reconocen antígenos tumorales y ejercen funciones antitumorales, inhibiendo el crecimiento del cáncer<sup>4,24</sup>. Una alta densidad de TIL CD8<sup>+</sup> se asocia con los denominados “tumores calientes”, caracterizados por una infiltración inmune activa y una mejor supervivencia global en HGSOE<sup>24</sup>.

Por otra parte, las células Tregs son inmunosupresoras, se encargan de mantener el equilibrio de la respuesta del sistema inmune<sup>24,26</sup>. Estas células contribuyen al crecimiento tumoral mediante la supresión de las células T antitumorales y supresión de linfocitos T antitumorales<sup>25</sup>. Altas proporciones de células Tregs se han vinculado con mal pronóstico y mayor riesgo de mortalidad en CO<sup>4,7,25</sup>.

### **Linfocitos B**

Las células B son un componente relevante de los linfocitos infiltrantes de tumor, actúan como células presentadoras de antígenos y activan las células T antitumorales al presentar los antígenos específicos del cáncer (neoantígenos)<sup>27</sup>. Forman parte esencial de las estructuras linfoides terciarias (TLS), donde coexisten con células T y células dendríticas mieloides. El papel pronóstico de las células B en CO es debatido. Algunos autores han identificado su presencia especialmente cuando coexisten con TIL CD8<sup>+</sup> y formando las TLS y se han asociado con una mayor supervivencia global en el HGSOE<sup>27,28</sup>. En contraste, cuando se encuentran altas proporciones de subtipos como las células B reguladoras se promueve la expansión de células Tregs y por ende un ambiente pro-tumoral por lo que también pueden estar asociadas con un mal pronóstico<sup>25,28</sup>.

### **Células Natural Killer (*NK cells*)**

Son linfocitos que residen en sangre periférica y son altamente efectores antitumorales. Las células NK comprometen entre 5% a 10% de los linfocitos circulantes y tienen la capacidad de eliminar células diana sin requerir sensibilización previa. Para inducen apoptosis de las células diana Este tipo celular se ha descrito con funciones tanto pro-tumorales como anti-tumorales<sup>24</sup>. La co-infiltración de células NK y células T citotóxicas en tumores se ha correlacionado con una mejor supervivencia. Asimismo, la co-infiltración de células NK con células B se ha correlacionado con un mal pronóstico en CO metastásico<sup>24,25</sup>. La actividad de las células NK puede verse suprimida por factores inmunosupresores del TME, como las citoquinas derivadas de TAMs y otras células inmunoregulatoras como las supresoras de origen mieloide (MDSCs)<sup>24</sup>.

## **2.3 Bioinformática y cáncer**

Desde la bioinformática se han realizado distintos acercamientos para el estudio del cáncer. Bases de datos como TCGA del instituto nacional de salud de Estados Unidos (NIH), cBioPortal, GTEx Portal y Gene Expression Omnibus (GEO), entre otras, registran datos genómicos y clínicos de distintos proyectos relacionados con cáncer<sup>29-33</sup>. La información depositada en las bases de datos es útil para llevar a cabo posteriores análisis complementarios e identificar diferentes factores genéticos y celulares que influyen en el desarrollo de los distintos tipos de cáncer.

El estudio de los datos genómicos y clínicos registrados en estas bases de datos permite análisis de distintos tipos, entre los cuales se encuentran: análisis de expresión diferencial de genes, análisis de vías de señalización, construcción de redes de interacción proteína-proteína e identificación de patrones moleculares como biomarcadores<sup>34</sup>. Además de las bases de datos relacionadas con los datos genómicos, también existen bases de datos que permiten el análisis funcional de los genes que se encuentran diferencialmente expresados, tales como: KEGG, DAVID, REACTOME y Gene Ontology (GO), entre otras<sup>35-40</sup>. Además, los avances tecnológicos recientes permiten el análisis de datos genómicos a gran escala para estudiar temas como: predicción de interacción ADN/RNA - proteína, predicción de estados de metilación, predicción de corte y empalme de RNA, análisis de redes de regulación<sup>41</sup>.

## 2.4 Análisis computacional de deconvolución

La deconvolución computacional es un enfoque bioinformático que permite estimar la composición de tipos celulares en tejidos heterogéneos, como muestras tumorales, a partir de datos de expresión génica de tejido completo<sup>42,43</sup>. Esta metodología es útil para comprender diferentes procesos como la progresión de enfermedades como el cáncer, donde la composición del microambiente tumoral influye en el pronóstico y la respuesta al tratamiento<sup>44,45</sup>. Desde un punto de vista matemático, la deconvolución se modela como una combinación lineal (Ecuación 1):

$$B = S \times P \quad (1)$$

donde  $B$  es el nivel de expresión heterogéneo observado,  $S$  los perfiles de expresión de cada tipo celular y  $P$  sus proporciones relativas. Estas últimas están restringidas a valores no negativos y su suma debe ser igual a uno<sup>14,43</sup>. Los métodos de deconvolución se clasifican basados en referencia, libres de referencia y parcialmente libres de referencia.

### Basados en referencia (RB)

Estos métodos requieren perfiles de expresión de tipos celulares puros o datos de secuenciación de RNA de única célula (scRNA-seq) como entrada. Suelen ser los más precisos, sin embargo, dependen de la disponibilidad de datos de referencia y pueden verse afectados por las diferentes técnicas de secuenciación usadas y a la fase preanalítica para la recolección de las muestras. Los métodos RB utilizan enfoques como regresión de mínimos cuadrados (OLS, NNLS, W-NNLS), regresión robusta, regresión de vectores de soporte (SVR), modelado bayesiano y máxima verosimilitud (MLE)<sup>43</sup>.

### Libres de referencia (RF)

Estos métodos no necesitan datos de referencia externos, sino que estiman simultáneamente perfiles y proporciones a partir de los datos de entrada. Utilizan análisis factorial como la factorización de matrices no negativas (NMF) o análisis de componentes independientes (ICA). La selección de genes marcadores y de variables informativos es clave, priorizando aquellas con baja varianza intra-tipo celular y alta varianza inter-tipo celular para distinguir y cuantificar los diferentes tipos celulares presentes en la muestra heterogénea<sup>43</sup>.

### **Parcialmente libres de referencia (PRF)**

Los métodos PRF son similares a los RF, pero incorporan información biológica adicional, como listas de genes marcadores, lo cual mejora la precisión y facilita la asignación de etiquetas a los diversos tipos celulares<sup>43</sup>.

Entre los diferentes métodos reportados en la literatura destacamos:

### **CIBERSORTx**

Es una nueva versión del algoritmo CIBERSORT (Cell type Identification By Estimating Relative Subsets Of known RNA Transcripts) que emplea regresión de vectores de soporte (v-SVR) para descomponer muestras de RNA-seq en sus diferentes tipos celulares<sup>11</sup>. CIBERSORTx es un método de enfoque RB, es decir, requiere una matriz de firma con perfiles de expresión específicos de cada tipo celular. A diferencia de otros métodos, este no se limita a una matriz de referencia fija, sino que permite construir matrices personalizadas a partir de datos de scRNA-seq o de poblaciones celulares purificadas, lo que le da gran flexibilidad para estudiar distintos tejidos. CIBERSORTx permite obtener perfiles de expresión específicos de cada tipo celular desde datos de tejido completo<sup>12</sup>. Esto puede hacerse a nivel grupal, generando un perfil promedio por tipo celular, o en modo de alta resolución, que permite obtener perfiles a nivel de muestra individual<sup>11,12</sup>.

El algoritmo modela la expresión génica de una muestra heterogénea como una ecuación lineal (Ecuación 1). Para estimar dichas proporciones, implementa v-SVR con kernel lineal, una técnica de aprendizaje automático que optimiza los coeficientes bajo dos principios: 1. minimizar los errores de predicción dentro de un margen definido ( $\epsilon$ -tube) y 2. penalizar desviaciones grandes mediante un parámetro de regularización (C). Esta formulación hace que el modelo sea robusto al ruido, a la colinealidad entre perfiles de referencia y a la presencia de tipos celulares no modelados explícitamente<sup>11</sup>. Además, CIBERSORTx incorpora esquemas de normalización entre plataformas lo que permite la comparabilidad entre datos de distinto origen<sup>11</sup>.

CIBERSORTx se ha consolidado como un método de referencia en la deconvolución transcriptómica<sup>9,13</sup>. Numerosos estudios de benchmarking lo han destacado por su buen desempeño tanto en la estimación de proporciones celulares como en la reconstrucción de perfiles de expresión específicos de cada tipo de célula<sup>13,14,42,45-47</sup>.

### **TOAST (-P) - Tools for the Analysis of heterogeneous Tissues**

Es un método PRF que mejora la estimación de la composición celular mediante la búsqueda iterativa de selección de características específicas de tipos celulares<sup>48</sup>. Su algoritmo funciona de la siguiente manera:

1. Selecciona genes con alta variabilidad y realiza una primera estimación libre de referencia.
2. Usa esas proporciones en un modelo lineal para identificar genes diferencialmente expresados entre tipos celulares.
3. Actualiza la lista de genes marcadores y repite el proceso, refinando progresivamente la estimación.

Este método posee una variante denominada TOAST/+P que permite incorporar conocimiento previo de proporciones en un marco bayesiano, aumentando la precisión<sup>49</sup>.

### **MuSiC - *MULTI-Subject Single Cell deconvolution***

Es un método RB que estima la composición celular en muestras de tejido completo utilizando como referencia datos de scRNA-seq de múltiples individuos<sup>50</sup>. Se fundamenta en una regresión de mínimos cuadrados no negativos ponderados (W-NNLS), donde los genes con menor variabilidad interindividual reciben mayor peso. Para solucionar el problema de colinealidad entre perfiles de tipos celulares cercanos, en MuSiC se implementó un procedimiento recursivo guiado por árboles que agrupa primero poblaciones similares y luego refina las proporciones en pasos sucesivos. Esta estrategia permite diferenciar con mayor precisión los tipos celulares estrechamente relacionados<sup>50</sup>. Debido a esto, ha demostrado ser un algoritmo robusto y consistente, manteniendo buena precisión incluso cuando los datos de tejido completo y los de referencia provienen de estudios o plataformas distintas<sup>43,46,50</sup>.

### **Linseed - Linearity of Transcriptional Signatures**

Es un método de deconvolución RF basado en la propiedad de linealidad mutua de los genes específicos de un tipo celular. En este método se propone que genes de un mismo tipo celular mantienen relaciones lineales entre sí a través de diferentes mezclas. Para identificar y seleccionar estos genes mutuamente lineales, Linseed construye una red donde los genes son nodos y los enlaces entre pares de genes se ponderan por su

coeficiente de linealidad mutua y correlación de Spearman. La significancia de los genes se estima mediante simulaciones de modelos nulos, generando un valor  $p$  para cada gen, que se define como la probabilidad de observar la linealidad mutua combinada de todos los enlaces asociados a dicho gen. Los genes con un valor  $p$  inferior a un umbral de significancia estadística de 0.01 se identifican como el conjunto de genes mutuamente lineales, permitiendo un filtrado robusto para distinguir los genes específicos de las "esquinas" del simplex de los irrelevantes o ruidosos<sup>51</sup>. Este método utiliza una normalización por fila y estima el número de tipos celulares ( $K$ ) mediante descomposición de valores singulares (SVD). Posteriormente, utiliza algoritmos geométricos como SISAL para identificar los vértices del simplex, que corresponden a las firmas celulares<sup>51</sup>.

### **CDseqR - Deconvolution using Sequencing data**

Es una herramienta de deconvolución RF que estima simultáneamente los perfiles de expresión génica específicos de tipos celulares (GEPs) y las proporciones de tipos celulares a partir de datos de RNA-seq. Este método emplea un modelado bayesiano jerárquico con variables aleatorias multinomiales de Dirichlet y un muestreador de Gibbs para la estimación de parámetros<sup>52</sup>.

### **Limitaciones y fuentes de sesgo de los análisis de deconvolución.**

El análisis computacional de deconvolución es una herramienta poderosa para caracterizar el TME pero no está exento de limitaciones y sesgos que deben considerarse con cuidado<sup>42</sup>. Uno de los retos principales es su fuerte dependencia de la matriz de firma de referencia. Es decir, si la referencia carece de ciertas poblaciones presentes en la muestra o existe inconsistencia técnica entre las plataformas utilizadas para generar la referencia (por ejemplo, diferencias entre scRNA-seq y RNA-seq), se pueden introducir sesgos en la asignación a los diferentes tipos celulares<sup>42,43,53</sup>.

A esto se suma la alta similitud transcripcional entre subtipos celulares, que genera multicolinealidad, por lo que la señal de un tipo celular se atribuye erróneamente a otro<sup>12</sup>. Este problema es especialmente relevante en enfoques de regresión, estrategias como la regularización o modelos bayesianos jerárquicos pueden atenuarlo. También existen sesgos sistemáticos derivados de diferencias en el tamaño celular o el contenido de ARNm por célula, que pueden llevar a sobre o subestimar ciertas poblaciones, y que en algunos casos se corrigen mediante normalizaciones ajustadas por tamaño celular estimado<sup>54,55</sup>.

Otros desafíos incluyen la dificultad para detectar con precisión poblaciones celulares raras.

La validación también representa un obstáculo: establecer un “*ground truth*” o valor verdadero de referencia en tejidos complejos es difícil, las simulaciones *in silico* son de utilidad, sin embargo, *datasets* empíricos con validación mediante inmunohistoquímica o FACS siguen siendo importantes para hacer una validación más realista<sup>44</sup>.

Sin embargo, a pesar de estas limitaciones, la deconvolución ofrece un valor importante, ya que permite explotar los datos de expresión génica masiva ya disponibles en bases de datos públicas o privadas, para generar nuevas hipótesis, relacionar o asociar la composición celular con variables clínicas y explorar el TME a una escala que sería costosa si se hiciera mediante estudios genómicos de célula única.

### **Benchmarking**

La comparación del desempeño o benchmarking de los métodos de deconvolución es fundamental para seleccionar la herramienta más adecuada en la investigación particular que se planea realizar. La precisión con la que un método estima las proporciones celulares se mide comparando las proporciones estimadas con unas proporciones “*ground truth*” obtenidas experimentalmente de distintas maneras tales como: mezclas celulares *in vitro*, simulaciones *in silico* a partir de datos de scRNA-seq, o mediante validación con citometría de flujo<sup>13,42</sup>.

Las simulaciones *in silico* a partir de datos de expresión de datos de scRNA-seq, agregando perfiles de células individuales en proporciones de tipos celulares predefinidas y conocidas se usan para generar “*pseudobulks*” o perfil transcriptómico simulado<sup>10</sup>. Esta metodología permite simular los datos obtenidos por secuenciación de biopsia o tejido tumoral heterogéneo denominado “*bulk*”<sup>55</sup>. Esta aproximación ofrece una “*ground truth*”, lo que permite medir con precisión el desempeño de los algoritmos de deconvolución, no solo en un escenario si no también simulando escenarios biológicos complejos<sup>14,45</sup>.

SimBu (del Inglés, Simulation of *Bulk* RNA-seq data) es un paquete de R diseñado específicamente para la generación de *pseudobulks* teniendo en cuenta sesgos biológicos<sup>55</sup>. Este paquete usa de datos scRNA-seq anotados y agrega los transcriptomas

de células individuales para construir perfiles *bulk* simulados. Su elemento distintivo es la capacidad de modelar y corregir sesgos debidos a diferencias en el contenido de ARNm entre tipos celulares, incorporando factores de escala que ajustan la contribución de cada célula a la muestra final<sup>55</sup>.

El proceso matemático puede describirse como la suma ponderada de perfiles de expresión de células individuales (Ecuación 2):

$$v_{bulk} = \sum_{i=1}^n (\text{perfil}_i \times \text{factor\_escala}_i) \quad (2)$$

donde  $i$  recorre las células seleccionadas para el *pseudobulk* y el factor de escala  $i$  ajusta la contribución según el sesgo de ARNm.

SimBu incluye varios escenarios predefinidos que determinan cómo se calculan las fracciones de tipos celulares y, por tanto, cuántas células de cada tipo se muestrean:

**random:** las fracciones se generan aleatoriamente a partir de una distribución uniforme y se escalan para sumar 1.

**weighted:** se asigna una fracción fija a un tipo celular de interés y las demás se generan aleatoriamente.

**even, custom, mirror\_db, pure:** permiten recrear composiciones específicas, simular muestras puras o imitar distribuciones observadas en *datasets* reales.

Una vez generados los *pseudobulks* con proporciones celulares conocidas, estos se convierten en la base de referencia para aplicar y comparar diferentes métodos de deconvolución, permitiendo calcular métricas de rendimiento, entre las más frecuentes que permiten evaluar la precisión de la predicción se encuentran<sup>14,42</sup>:

### **Coefficiente de Correlación de Pearson ( $r$ )**

El coeficiente de correlación de Pearson es una medida de la fuerza y dirección de una relación lineal entre dos variables continuas. En el contexto de la deconvolución, la correlación de Pearson se emplea para evaluar la relación lineal entre las proporciones celulares estimadas ( $x_j$ ) y las proporciones verdaderas ( $y_j$ ) para un tipo de célula específico  $k$  a lo largo de múltiples muestras  $j$  (Ecuación 3)<sup>56</sup>. También puede aplicarse

para medir la concordancia entre los perfiles de expresión génica estimados y los verdaderos.

$$r = \frac{\sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{j=1}^n (x_j - \bar{x})^2} \sqrt{\sum_{j=1}^n (y_j - \bar{y})^2}} \quad (3)$$

Donde:

$x_j$  : proporción estimada para la muestra  $j$

$y_j$  : proporción verdadera para la muestra  $j$

$\bar{x}$  : media de las proporciones estimadas  $x_j$  , calculada sobre las  $n$  muestras

$\bar{y}$  : media de las proporciones verdaderas  $y_j$  , calculada sobre las  $n$  muestras

$n$  : número total de muestras

El coeficiente  $r$  toma valores entre -1 y 1. Un valor de 1 indica una correlación lineal positiva perfecta, a medida que  $x$  aumenta,  $y$  también lo hace de forma proporcional. Un valor de -1 refleja una correlación lineal negativa perfecta, donde el incremento de  $x$  se asocia a una disminución proporcional de  $y$ . Un valor de 0 señala ausencia de correlación lineal. En el contexto de la deconvolución, un  $r$  cercano a +1 es deseable, ya que implica que las proporciones estimadas reproducen con alta fidelidad las proporciones verdaderas a lo largo de las muestras<sup>14</sup>.

### **Coefficiente de Correlación de Spearman ( $\rho$ )**

El coeficiente de correlación de Spearman, también conocido como correlación de rangos, mide la fuerza y la dirección de una relación monótona entre dos variables (una siempre aumenta o disminuye con la otra, aunque no necesariamente de forma lineal). A diferencia del coeficiente de Pearson, no exige que la relación sea estrictamente lineal ni que los datos sigan una distribución normal. En lugar de trabajar con los valores originales, Spearman convierte los datos a rangos y calcula sobre ellos la correlación de Pearson (Ecuación 4)<sup>56</sup>.

$$\rho_s = \frac{\sum_{i=1}^n (R_x(i) - \bar{R}_x)(R_y(i) - \bar{R}_y)}{\sqrt{\sum_{i=1}^n (R_x(i) - \bar{R}_x)^2} \sqrt{\sum_{i=1}^n (R_y(i) - \bar{R}_y)^2}} \quad (4)$$

Donde:

$R_x(j)$  : rango de la proporción estimada (o valor de expresión) en la muestra  $j$ ,

$R_y(j)$  : rango de la proporción verdadera (o valor de expresión) en la muestra  $j$ ,

$\overline{R_x}, \overline{R_y}$  : medias de los rangos  $R_x(j)$  y

$R_y(j)$ , respectivamente; para  $n$  observaciones, ambas son  $\frac{n+1}{2}$ ,

$n$  : número total de muestras.

El coeficiente  $\rho$  al igual que Pearson toma valores entre -1 y 1. Donde 1 indica una relación monótona perfectamente ascendente: a medida que una variable aumenta, la otra también lo hace de forma consistente. Un valor de -1 refleja una relación monótona perfectamente descendente: cuando una crece, la otra disminuye de manera constante. Por su parte, un valor de 0 sugiere ausencia de relación monótona. En el contexto de la deconvolución, la correlación de Spearman es especialmente útil para evaluar si un método conserva el orden relativo de las proporciones celulares o de los niveles de expresión génica entre las muestras, incluso cuando la relación no es estrictamente lineal<sup>14</sup>.

### Error Cuadrático Medio (RMSE)

Es una de las métricas de error más comunes para cuantificar la magnitud del error entre las estimaciones generadas por un modelo y los valores reales observados. En el contexto de la deconvolución, mide qué tan lejos, en promedio, están las proporciones celulares estimadas ( $y_j$ ) de las proporciones verdaderas ( $\hat{y}_j$ ) a lo largo de todas las muestras. Se calcula como la raíz cuadrada del promedio de los cuadrados de estas diferencias, tal como se expresa en la siguiente ecuación (Ecuación 5)<sup>43,44</sup>:

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (\hat{y}_j - y_j)^2} \quad (5)$$

Donde:

$\hat{y}_j$  : proporción celular estimada para la muestra  $j$ ,

$y_j$  : proporción celular verdadera para la muestra  $j$ ,

$n$  : número total de muestras evaluadas.

Un RMSE más bajo indica un mejor rendimiento. Se utiliza para cuantificar el error entre las proporciones estimadas y las proporciones verdaderas de las mezclas *pseudobulk*. En el benchmarking de deconvolución, el RMSE es crucial porque, a diferencia de la correlación, penaliza fuertemente los errores grandes. Si un método produce estimaciones con errores grandes, el RMSE será alto, incluso si la correlación de Pearson es buena.

### **Error Absoluto Medio (MAE o mAD)**

El MAE (Mean Absolute Error) o mAD (Mean Absolute Deviance) es una métrica de error que calcula la magnitud promedio de los errores entre las predicciones y las observaciones reales. Se define como el promedio de las diferencias absolutas entre los valores estimados y verdaderos (Ecuación 6)<sup>14,42,43</sup>.

$$MAE = \frac{1}{n} \sum_{j=1}^n |\hat{y}_j - y_j| \quad (6)$$

Donde:

$\hat{y}_j$  : proporción celular estimada para la muestra  $j$ ,

$y_j$  : proporción celular verdadera para la muestra  $j$ ,

$n$  : número total de muestras evaluadas.

Un MAE más bajo indica un mejor rendimiento, al igual que el RMSE. A menudo se utiliza junto con el RMSE y la correlación para proporcionar una visión completa del rendimiento. Es especialmente útil para evaluar la exactitud en la estimación de las proporciones celulares.

### **Deconvolución y análisis de supervivencia**

El análisis computacional de deconvolución se ha consolidado como una herramienta prometedora para estudios de supervivencia en cáncer, al ofrecer una visión más detallada del microambiente tumoral (TME), reconocido como un factor clave en la progresión de la enfermedad y en la respuesta a terapias oncológicas<sup>9,45</sup>. Una de sus principales ventajas es la posibilidad de aplicarlo a datos genómicos de tejido completo ya existentes, como los disponibles en repositorios públicos (TCGA, GEO), lo que amplía su alcance y permite generar hipótesis clínicas sin necesidad de generar nuevos *datasets*<sup>9</sup>.

No obstante, esta metodología también presenta algunas limitaciones ya mencionadas anteriormente en el documento. Por ejemplo, la heterogeneidad funcional dentro de un

mismo tipo celular, por ejemplo, entre células T efectoras y reguladoras, puede dificultar la interpretación de las firmas de expresión usadas como referencia<sup>13</sup>. Además, factores técnicos, como los sesgos introducidos por la disociación tisular en tecnologías de célula única o la selección del número de componentes en métodos no supervisados, pueden afectar la calidad de las estimaciones<sup>12,57</sup>.

A pesar de estas limitaciones, algunos estudios han demostrado el potencial de la deconvolución para generar hallazgos clínicamente relevantes. Por ejemplo, se ha observado que altos niveles de macrófagos predicen peor supervivencia en cáncer de mama y vejiga, mientras que una mayor infiltración de células T CD8+ se asocia con un mejor pronóstico en melanoma o cáncer de cabeza y cuello<sup>43,58</sup>. En cáncer hepático, la presencia de ciertos tipos celulares, como hepatocitos y células B maduras, se vinculó con una supervivencia más prolongada, en contraste con otros linajes como células madre bipotenciales o T reguladoras<sup>13</sup>. También se ha usado CIBERSORTx para predecir la respuesta a inmunoterapia en melanoma a partir de la proporción de células T CD8+ agotadas<sup>12</sup>. En cáncer gástrico, genes asociados al metabolismo de la lisina mostraron relación con la infiltración inmune y la supervivencia, siendo integrados en modelos de riesgo pronóstico<sup>59</sup>. Por otro lado, a nivel pan-cáncer, un estudio que combinó nueve herramientas de deconvolución logró clasificar tumores en 41 grupos distintos, estableciendo vínculos entre el TME, la supervivencia y el perfil genómico en 33 tipos de cáncer<sup>60</sup>.

## **3 Objetivos**

### **3.1 Objetivo general**

Comparar el desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq.

### **3.2 Objetivos específicos**

Determinar el desempeño de métodos de deconvolución a partir de datos de scRNA-seq de muestras de cáncer de ovario seroso de alto grado.

Identificar mediante los métodos de deconvolución de mejor desempeño las proporciones celulares a partir de datos de RNA-seq de muestras de cáncer de ovario seroso de alto grado.

Relacionar las proporciones celulares halladas en los métodos de deconvolución con la supervivencia de pacientes con cáncer de ovario seroso de alto grado.



# 4 Capítulo 1: Determinación del desempeño de métodos de deconvolución a partir de datos de scRNA-seq de muestras de HGSOC

## 4.1 Introducción

En este capítulo se estudia el rendimiento de distintos métodos de deconvolución aplicados a datos de scRNA-seq obtenidos de muestras de HGSOC. Si bien en los últimos años ha crecido la disponibilidad de algoritmos para el análisis computacional de deconvolución, elegir la herramienta más adecuada para un estudio específico sigue siendo un factor para tener en cuenta en el diseño del estudio a realizar<sup>11,53</sup>.

Investigaciones previas han examinado el desempeño de estos métodos bajo diversos escenarios de simulación y condiciones de preprocesamiento. Por ejemplo, Hippen *et al.* en 2023 estudiaron cómo influyen los factores técnicos en la deconvolución de tejido tumoral ovárico, centrándose en métodos RB<sup>45</sup>. Estos enfoques supervisados dependen de matrices de referencia previamente definidas que contienen perfiles de expresión génica de marcadores celulares, lo cual limita su aplicabilidad en contextos donde la identidad o la cantidad de poblaciones celulares es incierta, desconocida o particularmente compleja, como en el TME del HGSOC<sup>9,57</sup>. Por el contrario, los métodos de deconvolución basados en perfiles de referencia flexibles (PRF) y los enfoques no supervisados constituyen alternativas menos dependientes del conocimiento previo sobre los tipos celulares presentes<sup>52,57,61</sup>.

Este capítulo se centra en determinar el desempeño de métodos PRF y RF en datos provenientes de muestras de HGSOC, con el propósito de determinar su posible utilidad en la caracterización de la composición celular en contextos de estudios clínicos y de investigación.

#### 4 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

---

Se seleccionaron los métodos de deconvolución Linseed<sup>51</sup>, CDseqR<sup>52</sup>, CIBERSORTx<sup>11</sup> y TOAST<sup>48,49</sup>, dado que permiten estimar proporciones celulares a partir de datos de *bulk* RNA-seq. Estos métodos representan enfoques complementarios, combinando estrategias RF (Linseed, CDseqR), RB (CIBERSORTx) y PRF (TOAST), con el objetivo de hacer una evaluación comparativa robusta de la composición celular en el microambiente tumoral.

Para el análisis del desempeño de estos métodos de deconvolución se usaron los datos de scRNA-seq reportados por Hippen et al (2023), dado que estos autores reportaron en la base de datos GEO (Gene Expression Omnibus) datos generados a partir de muestras de tumores frescos de pacientes con HGSOC. En su estudio los investigadores evaluaron exhaustivamente los efectos de las diferencias de protocolo, la viabilidad de generar perfiles de referencia a partir de scRNA-seq, el impacto de la disociación tisular y de los métodos de enriquecimiento de RNA, por lo que resulta útil para evaluar los métodos de deconvolución.

## 4.2 Metodología

### 4.2.1 Procesamiento de datos

Se descargaron seis muestras (GSM6720925\_2251, GSM6720926\_2267, GSM6720927\_2283, GSM6720928\_2293, GSM6720929\_2380 y GSM6720931\_2467) del set de datos GSE217517 de scRNA-seq de HGSOC publicados por Hippen et al. en 2023<sup>45</sup>. Se obtuvieron los perfiles de expresión para cada muestra de célula única junto con los metadatos de anotación reportados por los investigadores en su repositorio de GitHub ([https://GitHub.com/greenelab/deconvolution\\_pilot](https://GitHub.com/greenelab/deconvolution_pilot)). El preprocesamiento se llevó a cabo en R usando el paquete Seurat. Durante el control de calidad, se retuvieron únicamente las células con más de 200 genes detectados y con un número de genes expresados dentro del rango definido por los percentiles 5 y 95 de la distribución global, descartando aquellas con un número excesivo de genes que podrían corresponder a dobletes. Asimismo, se eliminaron células con un porcentaje elevado de transcritos mitocondriales (>10%). Aquellas células sin anotación de tipo celular en los metadatos fueron descartadas del análisis.

## 4.2.2 Generación de *pseudobulks*

Para la generación de los *pseudobulks* se utilizó *SimBu*, que incorpora el sesgo derivado del contenido de ARNm entre tipos celulares<sup>55</sup>. A partir de cada objeto Seurat de scRNA-seq, que contenía metadatos de ID y tipo celular, se simularon 50 muestras de RNA-seq de tejido completo por escenario. Cada muestra contenía 6,000 células. Se consideraron cuatro escenarios distintos: *realistic*, *even*, *weighted* y *sparse*.

El escenario *realistic* replica la distribución celular observada en los datos de célula única. El escenario *even* representa una situación en la que todos los tipos celulares están presentes en proporciones aproximadamente iguales. En el escenario *weighted*, se fijó un 70 % de células epiteliales en la mezcla, con el objetivo de simular su predominancia característica en tumores epiteliales. Por último, el escenario *sparse* incluyó únicamente los tipos celulares más comunes en muestras tumorales (células epiteliales, fibroblastos, células endoteliales, macrófagos y linfocitos T), permitiendo así evaluar el rendimiento de los métodos de deconvolución ante la ausencia de ciertos tipos celulares. Para cada simulación se exportaron matrices de conteos con sus respectivas fracciones de celulares reales. Adicionalmente se generó una matriz adicional normalizada en CPM (counts per million).

## 4.2.3 Deconvolución

La deconvolución de perfiles *pseudobulk* se realizó utilizando los paquetes *Linseed*<sup>51</sup>, *CDseqR*<sup>52</sup>, *CIBERSORTx*<sup>11</sup> y *TOAST*<sup>48,49</sup>. La selección de métodos en este estudio se realizó teniendo en cuenta que Hippen et al. habían evaluado principalmente enfoques RB. Por ello, se decidió incluir también métodos RF y PRF que han mostrado buen desempeño en la literatura reciente<sup>43</sup>, dando preferencia a aquellos con una implementación accesible desde el punto de vista computacional. Aunque se consideraron otras alternativas RF, como *CellDistinguisher*<sup>62</sup> o *CAM3.0*<sup>63</sup>, no fue posible integrarlas en el análisis por limitaciones prácticas. *CIBERSORTx* se mantuvo en la comparación, aun cuando ya había sido evaluado por Hippen, debido a su amplio uso actual y su relevancia como punto de referencia en la mayoría de los estudios de deconvolución. Para cada escenario de simulación, el número de tipos celulares ( $k$ ) se estimó a partir de las fracciones conocidas, para los algoritmos que era requerido.

## 6 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

---

Para Linseed, el análisis se llevó a cabo considerando los 10.000 genes más variables, aplicando un umbral de significancia de  $p < 0.01$  con 100 permutaciones. Estos parámetros se seleccionaron de acuerdo con las recomendaciones dadas por los autores del paquete en su GitHub (<https://GitHub.com/ctlab/LinSeed>). A partir de la proyección en el simplex y la identificación de vértices mediante smartSearchCorners, se estimaron tanto las proporciones celulares como las firmas de expresión asociadas a cada tipo celular.

CDseqR se ejecutó con 700 iteraciones MCMC, 10 bloques de genes y un subconjunto aleatorio de 250 genes por iteración, con un factor de dilución de 3. Estos parámetros se seleccionaron de acuerdo con las recomendaciones dadas por los autores del paquete en su GitHub ([https://GitHub.com/kkang7/CDSeq\\_R\\_Package](https://GitHub.com/kkang7/CDSeq_R_Package)). El algoritmo produjo como salida las proporciones celulares estimadas y las firmas génicas asociadas a cada población, las cuales fueron exportadas para su análisis posterior.

CIBERSORTx se ejecutó en el modo Fractions a través de Docker. En primer lugar se usó este modo para la generación de las matrices de firmas requeridas por este método para la deconvolución a partir de los objetos Seurat de scRNA-seq. Posteriormente se llevó a cabo el análisis utilizando 100 permutaciones y deshabilitando la normalización por cuantiles (QN = FALSE), a fin de preservar las distribuciones originales de expresión. Como salida, el algoritmo produjo las proporciones celulares estimadas para cada muestra.

TOAST se usó en su versión PRF, por lo cual, a partir de los objetos Seurat de célula única se seleccionaron hasta 30 genes marcadores significativos mediante la función *ChooseMarker* para cada tipo celular, restringiendo el análisis a los genes comunes con el *pseudobulk* correspondiente. Finalmente, se aplicó la función MDeconv incluida en el paquete de TOAST sobre las matrices *pseudobulk* filtradas, generando como salida las proporciones celulares estimadas y los perfiles de expresión específicos de cada población.

#### 4.2.4 Generación de GEPs y anotación de tipos celulares

Los perfiles de expresión génica (GEPs) se generaron a partir de objetos Seurat previamente anotados. Los conteos fueron normalizados a unidades de CPM y posteriormente agregados por tipo celular, utilizando la media cuando el número de células era igual o superior a 500, y la mediana en los casos con menos de 500 células.

Para asegurar tanto la especificidad como la robustez de los perfiles, se aplicaron filtros sucesivos: un umbral mínimo de abundancia, exigiendo una expresión  $\geq 1$  CPM en al menos dos tipos celulares, un criterio de especificidad basado en *fold change*, seleccionando genes cuyo cociente entre la expresión máxima y la segunda mayor fuera  $\geq 1.5$ ; y un filtro de variabilidad, reteniendo únicamente aquellos genes con un coeficiente de variación por encima del percentil 50. Los GEPs resultantes fueron exportados en dos formatos: lineal (CPM) y logarítmico ( $\log_2[\text{CPM} + 1]$ ).

Para la anotación de los componentes estimados por los métodos no supervisados (Linseed, CDseqR), se implementó un enfoque de consenso multi-métrica, inspirado en el estudio de Jin y Liu en 2021 donde compararon algoritmos de deconvolución en diferentes ambientes dinámicos<sup>14</sup>. En primer lugar, se calcularon correlaciones de Spearman entre cada firma estimada (S) y los perfiles de expresión génica de referencia (GEP), asignando de forma preliminar cada componente al tipo celular con el coeficiente de correlación más alto. Complementariamente, se calcularon valores de  $\log_2$  fold change ( $\log_2\text{FC}$ ) y Z-scores a nivel génico. El Z-score se definió como la diferencia entre la expresión de un gen en el tipo celular de interés y la media de los demás tipos celulares, normalizada por su desviación estándar. Con base en estos genes discriminantes, se identificaron rutas moleculares características de cada tipo celular, y se aplicó FGSEA (fast gene set enrichment analysis) sobre los genes ordenados por su contribución a cada componente, utilizando como “*pathways*” las listas de genes específicos definidas previamente para cada tipo celular.

Para la asignación final de identidad celular, se empleó una jerarquía de reglas:

(i) cuando la correlación de Spearman coincidía con los resultados de FGSEA o del análisis de Z-scores, se mantenía dicha asignación;

## 8 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

---

(ii) si FGSEA y Z-scores coincidían y mostraban enriquecimiento significativo ( $p_{adj} \leq 0.05$ ;  $NES \geq 2$ ), se priorizaba la anotación basada en FGSEA;

(iii) en ausencia de concordancia entre métricas, se utilizaba la correlación de Spearman como criterio de respaldo.

Además, se implementó un procedimiento para resolver duplicidades en la asignación de identidades celulares. Cuando múltiples componentes eran asignados al mismo tipo celular, se reasignaban según la siguiente mayor correlación de Spearman.

### **4.2.5 Evaluación del desempeño de los algoritmos de deconvolución**

La comparación de métodos de deconvolución se realizó evaluando la concordancia entre proporciones estimadas y fracciones reales mediante tres métricas: coeficiente de correlación de Pearson, MAE y RMSE. Las métricas se calcularon a nivel global y por tipo celular.

### **4.2.6 Flujo computacional**

Todo el análisis fue implementado como un flujo reproducible en Snakemake, integrando desde la construcción de los objetos Seurat hasta la generación de las figuras finales. Este flujo de trabajo fue ejecutado en un entorno Linux, donde cada paso se implementó mediante scripts en R. Para asegurar la reproducibilidad de versiones y dependencias, se empleó un ambiente de conda que contenía tanto R como los paquetes necesarios.

## 4.3 Resultados

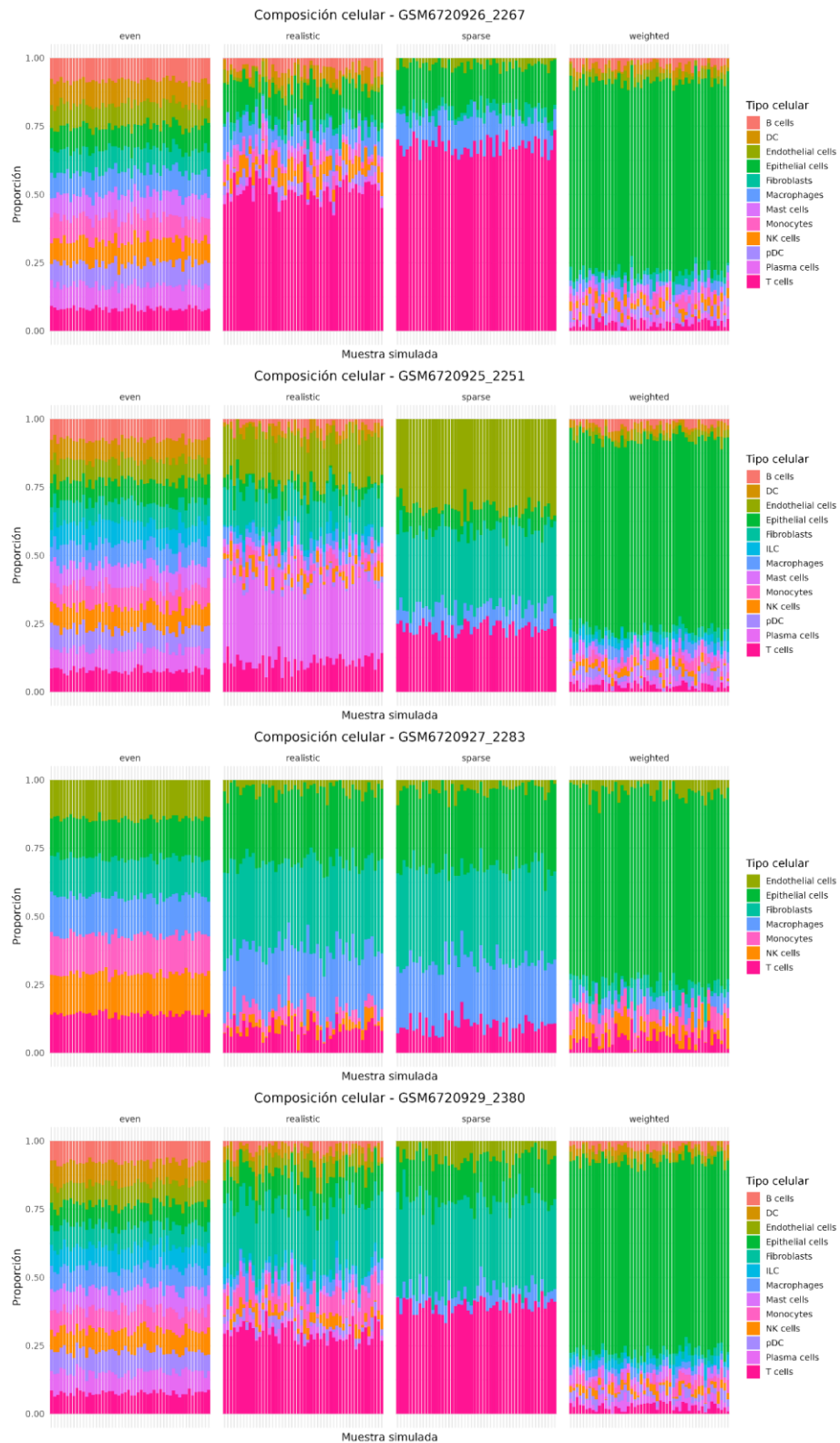
### 4.3.1 Generación de *pseudobulks*

Para evaluar el desempeño de los métodos de deconvolución, se generó un conjunto de “*pseudobulks*” a partir de datos scRNA-seq reales utilizando cuatro escenarios de mezcla: *even*, *realistic*, *sparse* y *weighted* (Figura 4-1). Cada “*pseudobulk*” se construyó a partir de un subconjunto de células asignadas proporcionalmente según el escenario. Cada escenario representó una mezcla entre 5 y 13 tipos celulares.

En los escenarios *even* y *weighted*, los *pseudobulks* incluyeron entre 12 y 13 poblaciones celulares, con presencia de células inmunes (*B cells*, *T cells*, *NK cells*, *Macrophages*, *Plasma cells*), así como células estructurales (*Fibroblasts*, *Endothelial cells*, *Epithelial cells*). El escenario *realistic* conservó esta complejidad, pero se ajustó a las proporciones derivadas de la distribución de los datos originales. Por último, el escenario “*sparse*” redujo la composición a un subconjunto restringido de 5 tipos celulares: *Endothelial cells*, *Fibroblasts*, *Macrophages*, *T cells* y *Epithelial cells*. El resumen de la conformación de los distintos “*pseudobulks*” se puede observar en la Tabla 4-1. Esta configuración de escenarios permite contrastar el desempeño en distintos tipos de muestras complejas con distinta diversidad celular.

**Figura 4-1. Composición celular de 4 escenarios (*even*, *realistic*, *sparse* y *weighted*) generados a partir de *pseudobulks* simulados con SimBu para cada *dataset* analizado.** En cada escenario se representan 50 muestras simuladas (columnas), con la altura proporcional de cada barra. Cada color representa un tipo celular.

## 10 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq



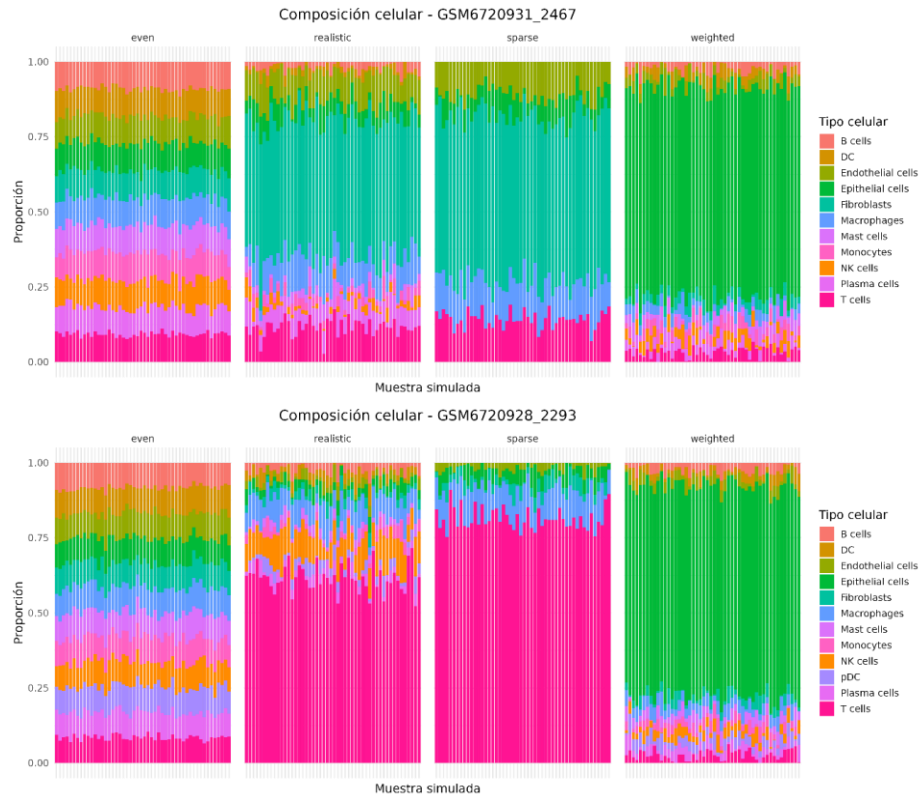


Tabla 4-1. Composición de *pseudobulks* simulados SimBu por dataset y escenario.

Dataset	Número de tipos celulares	Tipos celulares presentes
<b>GSM6720926_2267</b>	12 ( <i>even, realistic, weighted</i> ) / 5 ( <i>sparse</i> )	<i>NK cells, Macrophages, Epithelial cells, T cells, Monocytes, DC, B cells, Endothelial, Plasma cells, Fibroblasts, pDC, Mast cells</i>
<b>GSM6720925_2251</b>	13 ( <i>even, realistic, weighted</i> ) / 5 ( <i>sparse</i> )	<i>Plasma cells, B cells, Endothelial, T cells, Monocytes, Fibroblasts, NK cells, Epithelial cells, DC, ILC, pDC, Macrophages, Mast cells</i>
<b>GSM6720927_2283</b>	7 ( <i>even, realistic, weighted</i> ) / 5 ( <i>sparse</i> )	<i>Epithelial cells, Monocytes, Fibroblasts, T cells, Macrophages, Endothelial cells, NK cells</i>
<b>GSM6720929_2380</b>	13 ( <i>even, realistic, weighted</i> ) / 5 ( <i>sparse</i> )	<i>Fibroblasts, T cells, Monocytes, Endothelial, NK cells, Epithelial cells, Macrophages, B cells, Mast cells, ILC, pDC, DC, Plasma cells</i>
<b>GSM6720931_2467</b>	11 ( <i>even, realistic, weighted</i> ) / 5 ( <i>sparse</i> )	<i>Fibroblasts, Endothelial, Macrophages, Monocytes, B cells, T cells, Epithelial, Plasma cells, Mast cells, NK cells, DC</i>

- 12 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

<b>GSM6720928_2 293</b>	12 ( <i>even, realistic, weighted</i> ) / 5 ( <i>sparse</i> )	<i>Macrophages, T cells, NK cells, Epithelial cells, pDC, DC, Endothelial, B cells, Fibroblasts, Plasma cells, Monocytes, Mast cells</i>
-----------------------------	---	--

### 4.3.2 Deconvolución

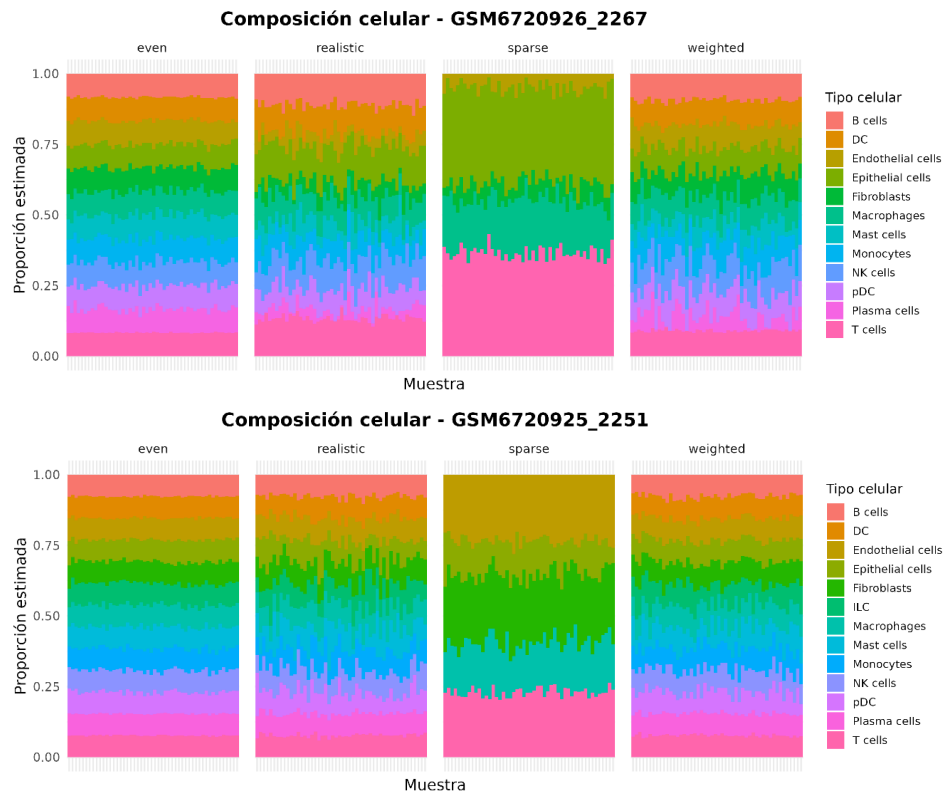
El resultado de los diferentes métodos de deconvolución produjo matrices de proporciones celulares por muestra y tipo celular para cada uno de los escenarios en los seis *datasets* analizados. Estas matrices se encuentran representadas en gráficos de barras apiladas como se puede observar el ejemplo de los resultados de TOAST en la figura 4-2. El conjunto completo de gráficos está disponible en Anexos. Esta forma de visualización permite una comparación visual directa entre los distintos escenarios de mezcla (*even, realistic, sparse* y *weighted*), facilitando la evaluación de las estimaciones generadas por cada método.

Se observa que en el escenario *even* los métodos TOAST, CIBERSORTx y CDseqR logran estimar correctamente las proporciones iguales de acuerdo con el diseño del *pseudobulk*. El método Linseed muestra mayor variación en la identificación de estas proporciones iguales, ya que algunos tipos celulares parecen tener predominancia en algunas muestras. En el escenario *sparse*, los *pseudobulks* estuvieron compuestos exclusivamente por cinco tipos celulares: *Endothelial cells, Epithelial cells, Fibroblasts, Macrophages, T cells*. Los métodos reference-free (RF) como Linseed y CDseqR identifican correctamente el número de células presentes, pero no asigna de forma directa un tipo específico. Por tanto, se requiere un paso posterior de anotación de los perfiles de expresión estimados.

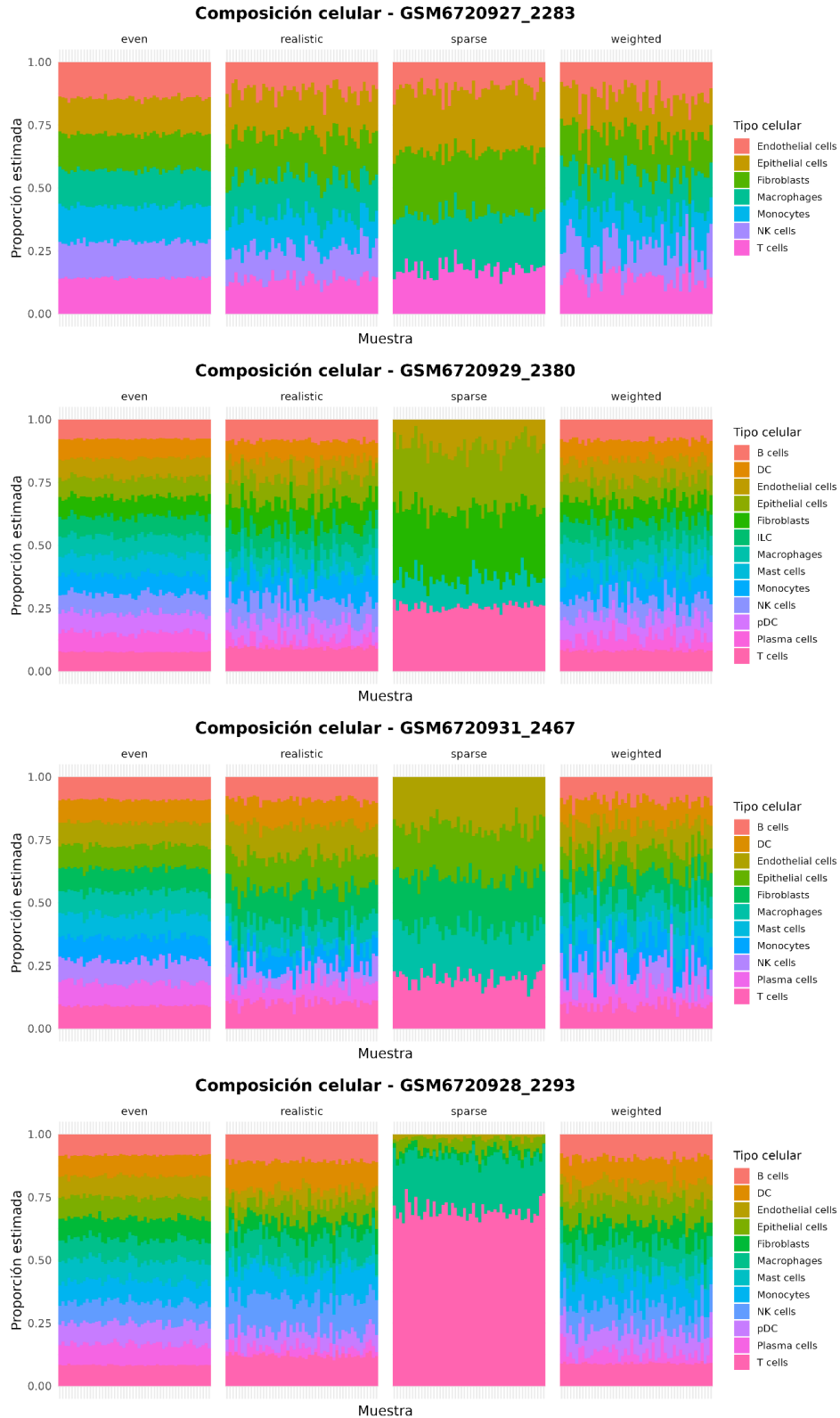
Como se muestra en el anexo A, las proporciones inferidas por los métodos RF (Linseed y CDseqR) no se asignan de manera clara a ninguno de los cinco linajes esperados. Este resultado impacta negativamente su desempeño general en este escenario *sparse*, ya que la calidad de la deconvolución depende no solo de la correcta estimación del número de componentes, sino también de la fidelidad con que estos pueden ser asociados a tipos celulares específicos. Al contrario, TOAST (método PRF) y CIBERSORTx (método RB) lograron identificar de forma correcta estos cinco tipos celulares en el escenario *sparse*.

En cuanto al escenario *weighted*, donde predominaba con un 70% el tipo celular *Epithelial cells* el único método que logró identificar esta alta proporción fue CIBERSORTx (Anexo C), a pesar de que Linseed, CDseqR y TOAST identificaban este tipo celular en ninguna muestra estimaron esa alta proporción.

**Figura 4-2. Proporciones celulares estimadas mediante deconvolución con TOAST en *pseudobulks* simulados bajo los escenarios *even*, *realistic*, *sparse* y *weighted*.** En cada escenario se representan las proporciones estimadas para 50 muestras (columnas), con la altura proporcional de cada barra indicando la fracción estimada para cada tipo celular. Se muestra la proporción celular para cada dataset.



14 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq



### 4.3.3 Evaluación del desempeño de los algoritmos de deconvolución

Además de la inspección visual de proporciones estimadas, se comparó el desempeño de cada método mediante métricas cuantitativas calculadas por tipo celular. Las métricas usadas fueron: correlación de Pearson entre proporciones estimadas y verdaderas, RMSE y MAE.

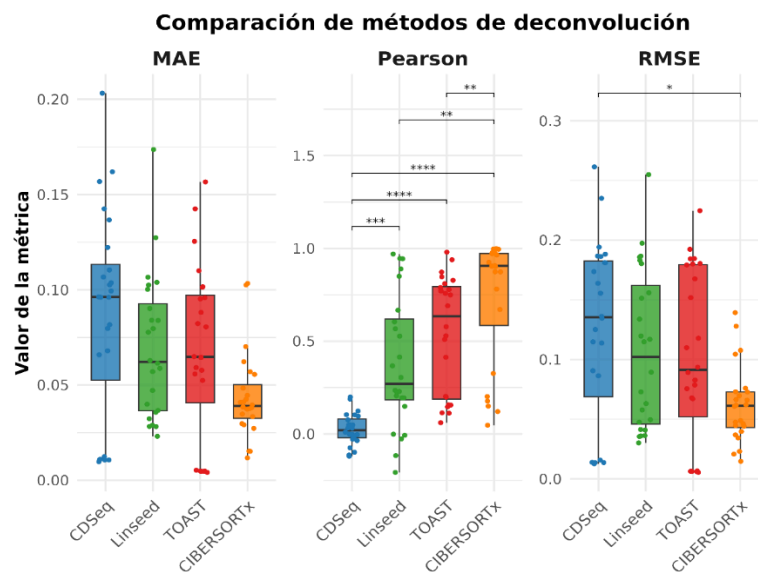
Primero se analizó el desempeño de los métodos de manera global, es decir, considerando todas las simulaciones en conjunto. El objetivo de este análisis fue obtener una visión preliminar sobre el desempeño de cada algoritmo estimando las proporciones celulares en términos de correlación y error. La comparación integrada de las métricas (Figura 4-3) indicó que CIBERSORTx presentó el desempeño global más alto (Pearson = 0.91, MAE = 0.039, RMSE = 0.061) seguido por TOAST (Pearson = 0.63, MAE = 0.065, RMSE = 0.091), mientras que Linseed (Pearson = 0.27, MAE = 0.062, RMSE = 0.102) y CDseqR (Pearson = 0.02, MAE = 0.096, RMSE = 0.135) mostraron resultados menos favorables con correlaciones cercanas a cero y errores más altos que los métodos previamente mencionados, evidenciando limitaciones claras para capturar las proporciones celulares en este conjunto de *pseudobulks*.

Si bien estos hallazgos evidencian diferencias claras entre el desempeño de los métodos evaluados, también destacan la importancia de realizar un análisis más específico por tipo celular. Las métricas globales, aunque útiles como resumen general, pueden ocultar fortalezas o limitaciones específicas de cada algoritmo en determinados linajes celulares. Por ello, a continuación se evaluó el comportamiento de cada algoritmo desagregado por tipo celular y escenario.

**Figura 4-3. Comparación del desempeño de métodos de deconvolución en métricas globales (Pearson, MAE y RMSE).** Diagrama de cajas de la distribución de valores de desempeño para CDSeqR, Linseed, TOAST y CIBERSORTx en términos de MAE, correlación de Pearson y RMSE. Cada caja representa la variabilidad del método a través de múltiples escenarios de pseudobulks simulados, los puntos corresponden a valores individuales.

16 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

---



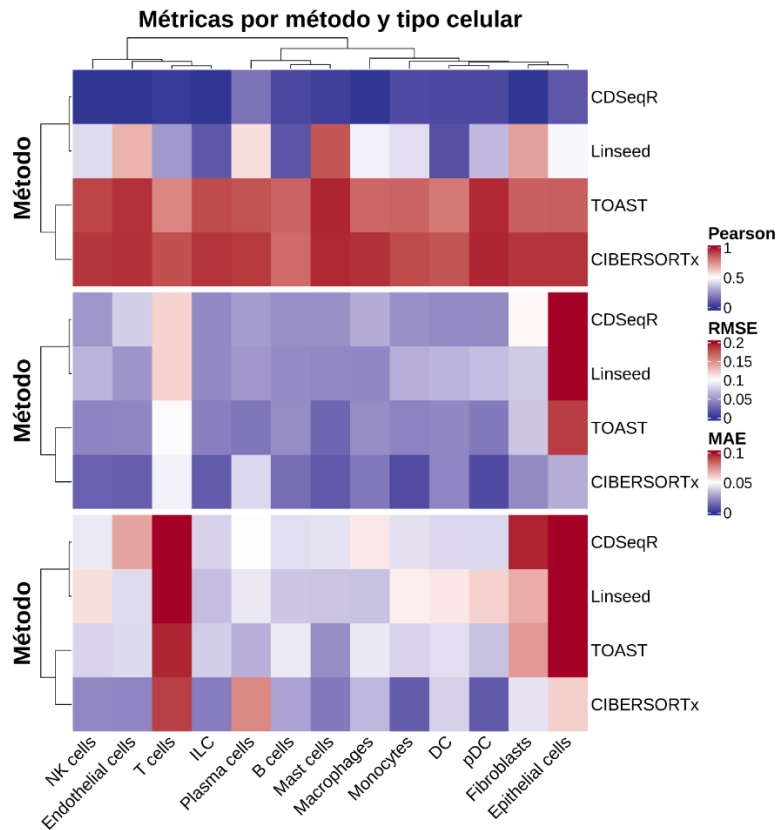
CIBERSORTx y TOAST obtuvieron los valores más altos de correlación de Pearson (entre 0.87 y 0.92, respectivamente) para la mayoría de los linajes, particularmente en células inmunes de alta frecuencia como *B cells* (entre 0.83 y 0.84), *NK cells* (entre 0.94 y 0.91) y *Macrophages* entre (0.95 y 0.84) (Figura 4-4, panel superior). En contraste, Linseed y CDSeqR tuvieron un menor rendimiento: Linseed mostró correlación de Pearson bajas o moderadas para las células *NK cells* (0.42), *Macrophages* (0.47) y *T cells* (0.26), mientras que CDSeqR presentó valores cercanos a cero en la mayoría de los linajes, como *B cells* (0.05), *T cells* (0.01) y *NK cells* (-0.06). Además, en tipos celulares menos frecuentes como *plasma cells* y *mast cells*, las correlaciones de Linseed fueron de 0.57 y 0.88, mientras que CDSeqR apenas alcanzó 0.16 y 0.02, respectivamente. Finalmente, en *Epithelial cells*, predominantes en los escenarios *weighted* y *sparse* de los pseudobulks, Linseed obtuvo 0.48 y CDSeqR solo 0.09, frente a los valores mucho más altos de TOAST (0.86) y CIBERSORTx (0.94).

CIBERSORTx mostró los valores más bajos tanto de RMSE y MAE en prácticamente todos los tipos celulares, con valores de RMSE entre 0.012 (*pDC*) y 0.093 (*T cells*) y MAE entre 0.010 (*pDC*) y 0.092 (*T cells*) confirmando su menor desviación respecto a las proporciones verdaderas (Figura 4-4). Mientras que TOAST y Linseed presentaron un desempeño intermedio con RMSE entre 0.043 en *Macrophages* y 0.201 en *Plasma cells*, y MAE entre

0.035 en *Macrophages* y 0.195 en *Plasma cells*. Sin embargo, el método TOAST mostro una distribución más homogénea entre los distintos tipos celulares, con valores de error que oscilaron en torno a 0.04–0.06 en linajes como *B cells* y *Monocytes*, y aumentos moderados en poblaciones como *T cells* (MAE = 0.11, RMSE = 0.12) y *Epithelial cells* (MAE = 0.19, RMSE = 0.20). En contraste, Linseed presentó un patrón más heterogéneo, con errores relativamente bajos en linajes como *Macrophages* (MAE = 0.036, RMSE = 0.043) y *Mast cells* (MAE = 0.036, RMSE = 0.044), pero incrementos marcados en *T cells* (MAE = 0.11, RMSE = 0.12) y *Epithelial cells* (MAE = 0.19, RMSE = 0.20). Finalmente, CDSeqR registró los valores de error más altos de manera sistemática, particularmente en *Epithelial cells* (MAE = 0.20, RMSE = 0.20), *Fibroblasts* (MAE = 0.097, RMSE = 0.103) y *T cells* (MAE = 0.11, RMSE = 0.12), lo que evidencia una menor capacidad de ajuste en la mayoría de los escenarios evaluados.

**Figura 4-4. Comparación del desempeño de métodos de deconvolución en métricas por tipo celular (RMSE, MAE y Pearson).** *Mapa de calor que muestran el desempeño promedio de CDSeqR, Linseed, TOAST y CIBERSORTx en distintos tipos celulares, evaluado mediante correlación de Pearson, RMSE y MAE. Las filas corresponden a métodos de deconvolución y las columnas a tipos celulares. Los gradientes de color indican concordancia (Pearson) o magnitud del error (RMSE, MAE).*

18 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq



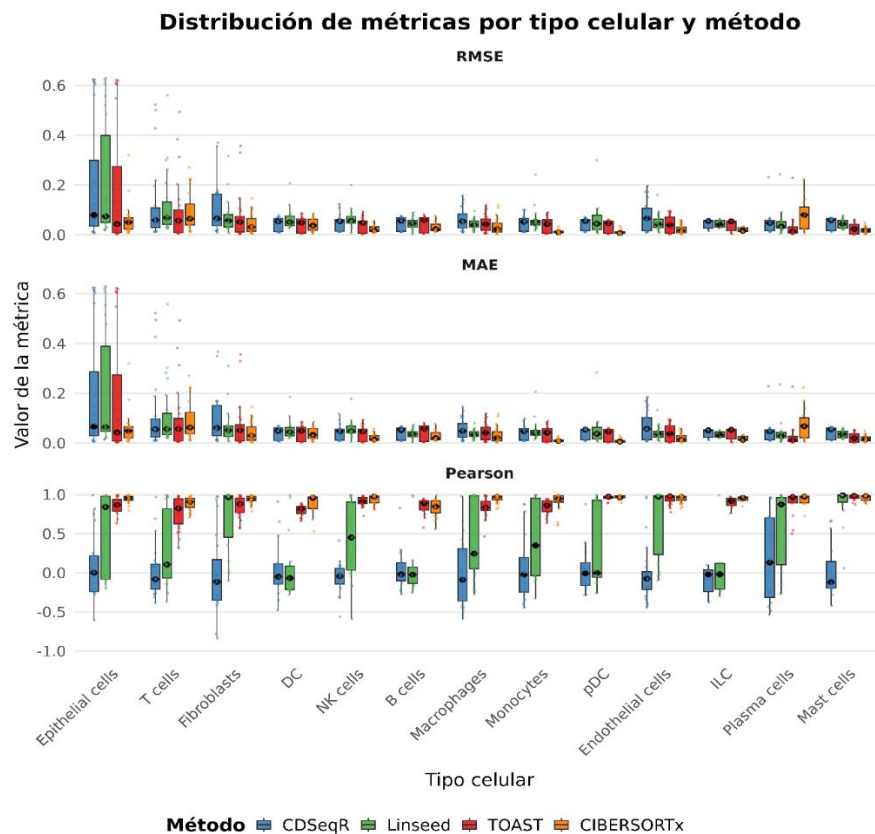
Por otra parte, la representación en el diagrama de cajas y bigotes observada en la figura 4-5 permite otra perspectiva de análisis para la comparación del desempeño de los métodos en relación con los tipos celulares, más allá de sus métricas promedio. En línea con los patrones observados en el mapa de calor, CIBERSORTx no solo presentó el mejor desempeño central, sino también la menor dispersión en RMSE, MAE y correlación de Pearson, lo que indica una estimación consistente a través de replicados y escenarios. TOAST mostró un comportamiento igualmente consistente, aunque con mayor variabilidad en linajes inmunes como *T cells* y *Fibroblasts*, lo que sugiere una mayor fluctuación de la estimación en los distintos escenarios.

En contraste, Linseed mostró una gran variabilidad en varios tipos celulares, con resultados de desempeño intermedios pero acompañados de varios valores extremos, lo que indica que sus estimaciones son menos consistentes. Por su parte, CDseqR mostró medianas

desfavorables con variabilidad alta entre réplicas, particularmente en *Epithelial cells* y *T cells*.

De manera general, CIBERSORTx y TOAST se destacan como los algoritmos más precisos y consistentes, alcanzando altas correlaciones de Pearson ( $r=0.9$ ) y bajos valores de error (MAE y RMSE) entre tipos celulares y los diferentes escenarios evaluados. Por el contrario, Linseed y CDseqR presentan desempeños más heterogéneo: aunque en algunos linajes logran métricas intermedias, sus resultados muestran mayor dispersión y menor reproducibilidad, lo que refleja una menor robustez frente a la complejidad y variabilidad de los *pseudobulks* simulados.

**Figura 4-5. Comparación del desempeño de métodos de deconvolución en función de métricas (Pearson, MAE y RMSE) por tipo celular.** Distribución de valores de RMSE, MAE y correlación de Pearson para cada tipo celular, comparando cuatro métodos de deconvolución: CDseqR, Linseed, TOAST y CIBERSORTx. Cada caja resume la variabilidad entre simulaciones, mientras que los puntos representan valores individuales.



## 20 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

---

Para ilustrar de manera directa la concordancia entre proporciones reales y estimadas, se generaron gráficos de dispersión *truth vs estimated* para cada método en los distintos escenarios de *pseudobulks* (Figuras 4-6 a 4-9). En estos paneles, un mejor desempeño se refleja en la concentración de los puntos alrededor de la diagonal, acompañada de valores altos de correlación y valores de error RMSE y MAE bajos. Esto con el objetivo de complementar el análisis del desempeño de los métodos bajo los distintos escenarios de simulación.

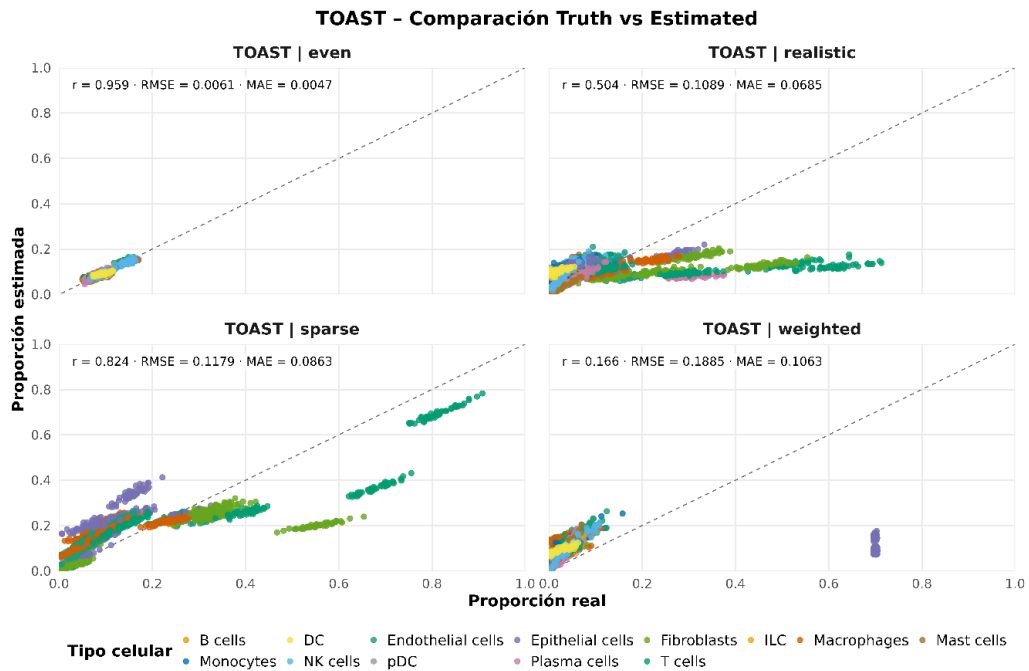
La comparación entre valores reales y estimados para el método TOAST (Figura 4-6) muestra una dependencia del desempeño con respecto al escenario de mezcla. En el escenario *even*, las estimaciones mostraron similitud con los valores reales, ya que los puntos están distribuidos estrechamente a lo largo de la diagonal de identidad ( $r = 0.96$ ; RMSE = 0.0061; MAE = 0.0047), lo que indica una asignación casi exacta de las fracciones celulares balanceadas.

En el escenario *sparse*, que incluye poblaciones celulares dominantes, TOAST mantuvo un nivel aceptable de ajuste ( $r = 0.82$ ), aunque con una mayor dispersión de puntos y una tendencia sistemática a subestimar proporciones elevadas como las *Endothelial cells* y *Fibroblasts*, reflejada en un aumento de los errores (RMSE = 0.118; MAE = 0.086).

En contraste, los escenarios *realistic* y *weighted* evidenciaron una disminución notable en el rendimiento del método. En el escenario *realistic*, las proporciones estimadas tendieron a concentrarse en valores bajos, con una correlación moderada ( $r = 0.50$ ) y valores de errores incrementados (RMSE = 0.109; MAE = 0.069). En el escenario *weighted*, el desempeño fue inferior: la mayoría de las estimaciones colapsaron hacia valores cercanos a cero, sin seguir el patrón esperado ( $r = 0.17$ ), lo que sugiere una limitación importante del método para manejar distribuciones altamente sesgadas hacia una o pocas poblaciones, dado que no identificó a las *Epithelial cells*, que en este escenario era el tipo celular predominante.

En conjunto, estos resultados indican que TOAST presenta un desempeño robusto en contextos con proporciones balanceadas o número reducido de linajes, pero pierde precisión en escenarios con fuerte desbalance celular, particularmente cuando una población domina ampliamente la mezcla.

**Figura 4-6. Comparación de proporciones celulares reales vs. estimadas con TOAST en cuatro escenarios de simulación (*even*, *realistic*, *sparse* y *weighted*).** Relación entre proporciones celulares reales y estimadas mediante TOAST en pseudobulks simulados bajo los escenarios *even*, *realistic*, *sparse* y *weighted*. Cada punto corresponde a un tipo celular en una muestra simulada. Se incluyen los valores de correlación de Pearson (*r*), RMSE y MAE para cada escenario.



El desempeño de Linseed varió según el escenario de mezcla (Figura 4-7). En el escenario *even*, las estimaciones presentaron una correlación relativamente baja con respecto a las proporciones reales ( $r = 0.49$ ), a pesar de registrar valores de error reducidos ( $RMSE = 0.038$ ;  $MAE = 0.030$ ). Aunque los puntos se distribuyeron en torno a la diagonal, la dispersión observada indica una pérdida de ajuste que sugiere que, incluso en condiciones balanceadas, el método no alcanza la precisión lograda por enfoques basados en referencia.

## 22 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

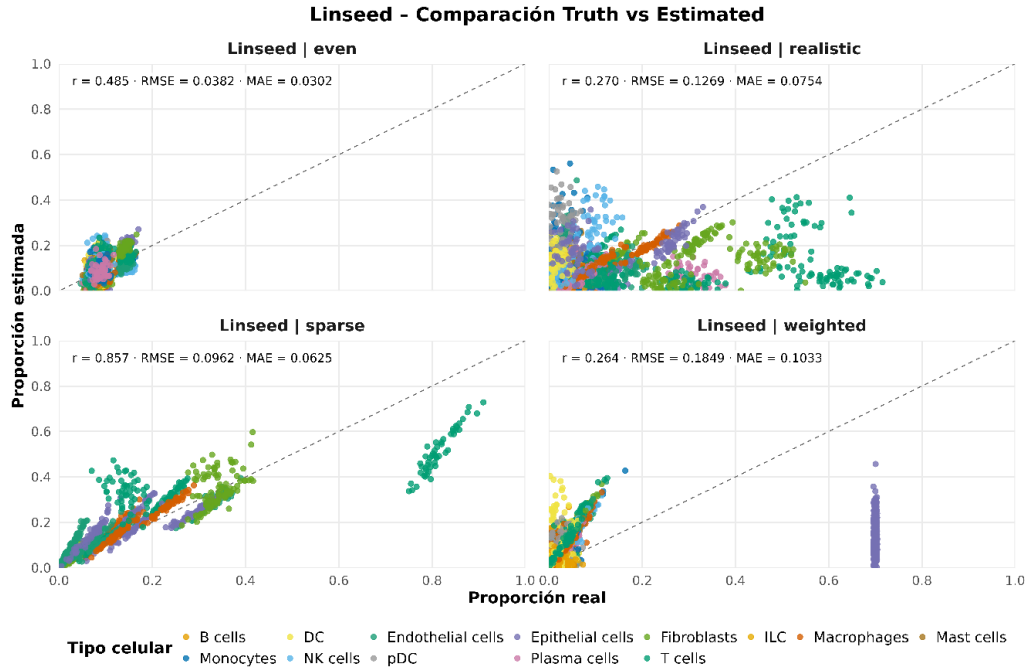
---

En el escenario *sparse*, donde solo unas pocas poblaciones dominan la mezcla, Linseed alcanzó su mejor rendimiento ( $r = 0.86$ ), con una nube de puntos más estrechamente alineada con la diagonal y errores comparativamente bajos (RMSE = 0.096; MAE = 0.063). Este resultado resalta la capacidad del método para capturar proporciones cuando la estructura de la mezcla es más simple y las poblaciones están bien definidas geoméricamente.

Por el contrario, en los escenarios *realistic* y *weighted*, el rendimiento fue considerablemente inferior. En *realistic*, la correlación fue baja ( $r = 0.27$ ), con alta dispersión y una tendencia sistemática a subestimar proporciones elevadas (RMSE = 0.127; MAE = 0.075). Asimismo, en el escenario *weighted*, las estimaciones se concentraron en valores bajos y poco diferenciados, reflejando una correlación baja ( $r = 0.26$ ) y valores de error más altos (RMSE = 0.185; MAE = 0.103).

En conjunto, estos resultados indican que el desempeño de Linseed está fuertemente influido por la complejidad del escenario de mezcla. Si bien puede ofrecer resultados competitivos en configuraciones simples como *sparse*, su precisión y consistencia disminuyen en contextos más realistas o dominados por ciertas poblaciones, lo que limita su aplicabilidad en condiciones de mayor heterogeneidad.

**Figura 4-7. Comparación de proporciones celulares reales vs. estimadas con Linseed en cuatro escenarios de simulación (*even*, *realistic*, *sparse* y *weighted*).** Relación entre proporciones celulares reales y estimadas mediante Linseed en pseudobulks simulados bajo los escenarios *even*, *realistic*, *sparse* y *weighted*. Cada punto corresponde a un tipo celular en una muestra simulada. Se incluyen los valores de correlación de Pearson ( $r$ ), RMSE y MAE para cada escenario.



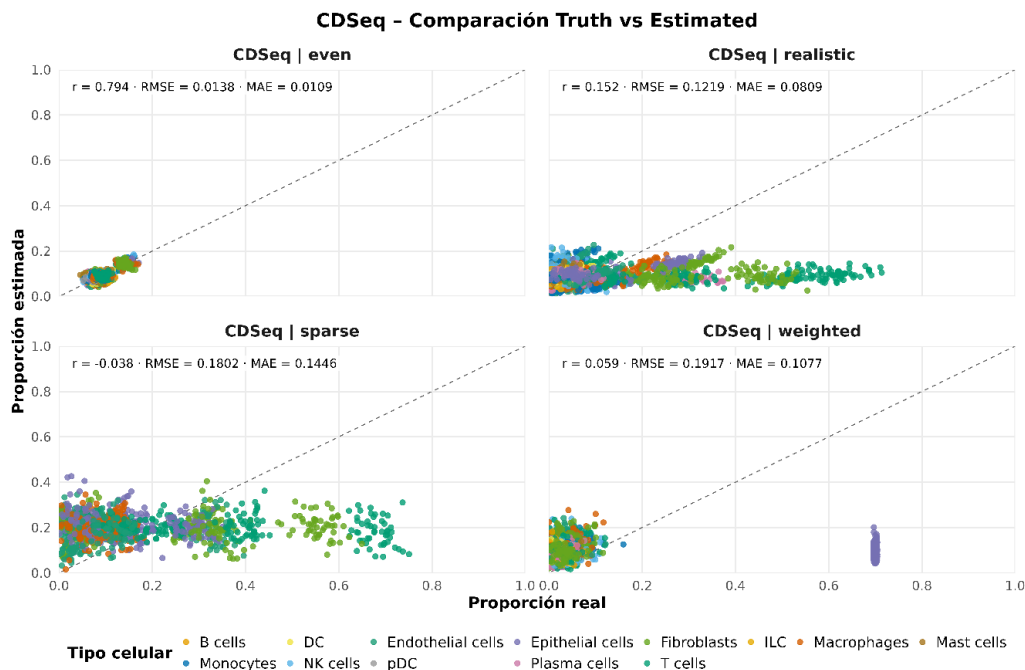
El desempeño de CDSeqR presentó limitaciones evidentes en la asignación de proporciones celulares (Figura 4-8). En el escenario *even*, el método alcanzó un ajuste moderado, con una correlación aceptable ( $r = 0.79$ ) y errores relativamente bajos (RMSE = 0.014; MAE = 0.011). Aunque los puntos se agruparon en torno a la diagonal, la dispersión fue mayor en comparación con los métodos basados en referencia, lo que refleja una menor precisión incluso bajo condiciones balanceadas.

Sin embargo, el rendimiento se disminuyó en los escenarios *realistic*, *sparse* y *weighted*. En el escenario *realistic*, la correlación disminuyó a  $r = 0.15$ , con una dispersión notable y un aumento en los valores de error RMSE(0.122) y MAE ( 0.081). En el escenario *sparse*, la correlación fue negativa ( $r = -0.04$ ), lo que indica correlación negativa entre las proporciones reales y las estimadas, acompañada de un incremento sustancial en los errores (RMSE = 0.180; MAE = 0.145). De forma similar, en *weighted* se observó una correlación cercana a cero ( $r = 0.06$ ) y valores error RMSE y MAE elevados (0.192 y 0.108, respectivamente), con predicciones claramente desviadas respecto a los valores esperados.

## 24 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

En conjunto, los resultados indican que CDseqR ofrece un rendimiento aceptable únicamente en condiciones balanceadas, pero pierde casi por completo su capacidad de estimación en escenarios más realistas o desbalanceados.

**Figura 4-8. Comparación de proporciones celulares reales vs. estimadas con CDSeqR en cuatro escenarios de simulación (*even*, *realistic*, *sparse* y *weighted*).** Relación entre proporciones celulares reales y estimadas mediante CDseqR en pseudobulks simulados bajo los escenarios *even*, *realistic*, *sparse* y *weighted*. Cada punto corresponde a un tipo celular en una muestra simulada. Se incluyen los valores de correlación de Pearson ( $r$ ), RMSE y MAE para cada escenario.



Las comparaciones entre proporciones reales y estimadas con CIBERSORTx (Figura 4-9) evidenciaron, el mejor desempeño entre todos los métodos evaluados. En el escenario *even*, las estimaciones se alinearon con los valores reales. Se observa una distribución estrechamente a lo largo de la diagonal de identidad y una correlación alta ( $r = 0.98$ ;  $RMSE = 0.004$ ;  $MAE = 0.003$ ). En el escenario *realistic*, CIBERSORTx mantuvo un rendimiento notable ( $r = 0.90$ ), con valores de error bajos ( $RMSE = 0.061$ ;  $MAE = 0.044$ ). Sin embargo,

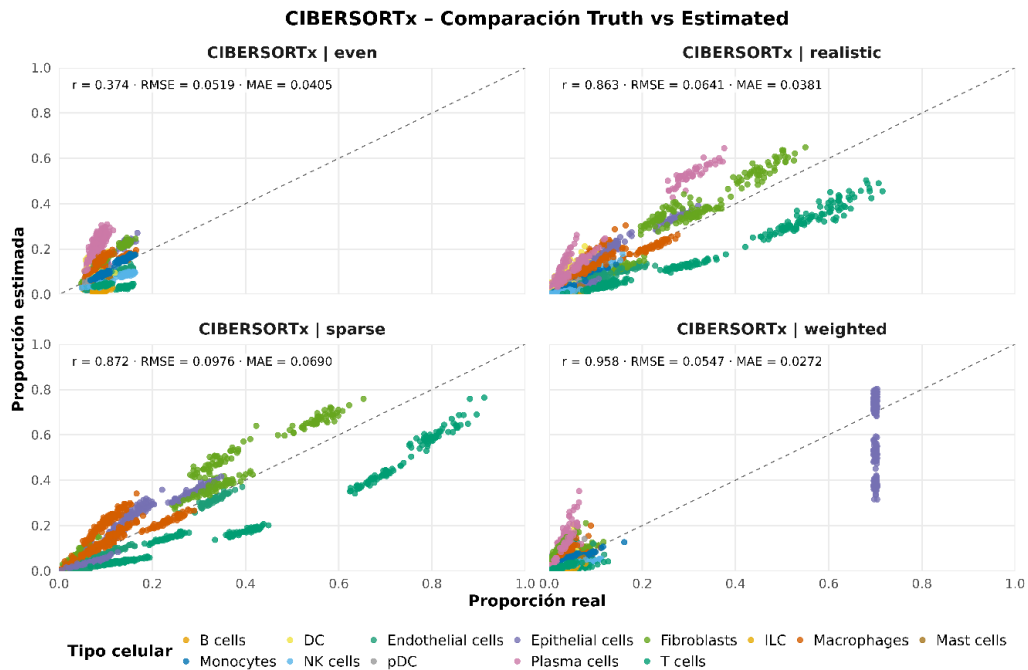
mostró una ligera dispersión en las poblaciones de baja frecuencia (*Mast cells*, *Plasma cells*, *ILC* y *pDC*).

En el escenario *sparse*, el método también mostró una correlación elevada ( $r = 0.93$ ), con estimaciones cercanas a la diagonal y errores moderados (RMSE = 0.081; MAE = 0.055), lo que indica una buena capacidad para asignar proporciones en contextos de baja complejidad celular. Finalmente, en el escenario *weighted*, donde predominan una o pocas poblaciones, CIBERSORTx conservó un buen desempeño ( $r = 0.87$ ), aunque con mayor dispersión en el rango de proporciones celulares bajas, lo que sugiere cierta dificultad para estimar con precisión valores extremos en poblaciones dominantes.

En conjunto, los resultados posicionan a CIBERSORTx como el método más robusto y consistente del conjunto evaluado, destacando por su precisión tanto en escenarios balanceados como en configuraciones más realistas o desbalanceadas.

**Figura 4-9. Comparación de proporciones celulares reales vs. estimadas con CIBERSORTx en cuatro escenarios de simulación (*even*, *realistic*, *sparse* y *weighted*).** Relación entre proporciones celulares reales y estimadas mediante CIBERSORTx en pseudobulks simulados bajo los escenarios *even*, *realistic*, *sparse* y *weighted*. Cada punto corresponde a un tipo celular en una muestra simulada. Se incluyen los valores de correlación de Pearson ( $r$ ), RMSE y MAE para cada escenario.

## 26 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq



En conjunto, esta comparación muestra que los métodos basados en referencia (CIBERSORTx y TOAST) mantienen un mejor desempeño global y mayor consistencia, mientras que los enfoques RF (Linseed y CDSeqR) exhiben un comportamiento con mayor dependencia de la simplicidad del escenario y con un deterioro notable en configuraciones realistas o dominadas por pocas poblaciones.

## 4.4 Discusión

La evaluación comparativa de métodos de deconvolución es esencial para orientar su uso en estudios de transcriptómica a partir de RNA-seq, dado que las características de los datos, la composición celular y la heterogeneidad en las muestras de cáncer pueden afectar de manera sustancial su desempeño. En este trabajo analizamos distintos algoritmos aplicados a pseudobulks generados a partir de datos de scRNA-seq, con el fin de caracterizar su robustez y estabilidad en escenarios con variabilidad controlada.

En el presente estudio se evidenció diferencias consistentes en el desempeño y estabilidad de los métodos RB, RF y PRF. CIBERSORTx se destacó como el método más robusto, con correlaciones elevadas y bajos valores de error en todos los escenarios evaluados. TOAST mostró un rendimiento intermedio, con buen ajuste en configuraciones balanceadas, pero una pérdida de precisión en mezclas celulares con composiciones sesgadas. Mientras que con el método Linseed se obtuvo resultados aceptables únicamente en contextos simples como los escenarios *sparse*, mientras que su desempeño disminuyó notablemente en escenarios más complejos o dominados por poblaciones celulares específicas. CDSeqR, por su parte, mantuvo un ajuste razonable solo bajo condiciones balanceadas, con un deterioro sustancial en el resto de los escenarios. En conjunto, estos resultados sugieren que los métodos RB superan a los enfoques RF, lo cual es congruente con lo reportado en la literatura<sup>43</sup>.

Estudios como el *DREAM Challenge* han destacado a CIBERSORTx como el mejor método en tareas de deconvolución a nivel grueso, evidenciando su solidez frente a diversos factores experimentales, incluidos los protocolos de captura y la heterogeneidad de los perfiles de referencia<sup>12</sup>. Aunque algunas fuentes mencionan limitaciones en contextos con árboles celulares muy complejos (>10 tipos), nuestros resultados en *pseudobulks* no mostraron indicios claros de estos inconvenientes<sup>13</sup>.

En contraste, la evaluación de los métodos reference-free reveló un panorama más heterogéneo. TOAST al ser un método PRF mostró un rendimiento intermedio: se desempeñó bien en configuraciones balanceadas, pero perdió precisión en escenarios desiguales. Su utilidad radica en contextos donde los paneles de referencia no están disponibles, o las muestras son limitadas en tamaño. Entre sus fortalezas destaca la capacidad de integrar marcadores específicos, así como un mecanismo iterativo de selección de características que puede mejorar la estimación. Sin embargo, su diseño original enfocado en datos de microarreglos podría explicar las dificultades observadas al aplicarlo en *pseudobulks* generados desde scRNA-seq<sup>48,49</sup>. Esto podría explicarse debido a que los datos de RNA-seq, especialmente los derivados de scRNA-seq, son intrínsecamente más dispersos y presentan una alta proporción de ceros, lo cual puede no ser manejado eficientemente por un modelo matemático optimizado para distribuciones de datos continuas y menos dispersas<sup>54</sup>.

## 28 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

---

Por otra parte, los enfoques RF enfrentan desafíos significativos en contextos donde el conocimiento a priori es limitado o poco confiable, suelen ser más variables, requieren más muestras, y pueden fallar en escenarios clínicos pequeños<sup>9</sup>. En este trabajo, el método Linseed mostró un desempeño competitivo únicamente en configuraciones simples donde los tipos celulares están claramente definidos. Sin embargo, su precisión se deterioró a medida que aumentó la complejidad de las mezclas. Aunque ha sido destacado en la literatura como uno de los métodos RF más precisos, nuestros resultados muestran limitaciones importantes bajo condiciones más cercanas a escenarios biológicos reales<sup>43</sup>. Linseed, siendo un método de deconvolución de tipo "scoring" o basado en la propiedad de linealidad mutua de genes específicos de un tipo celular, inherentemente encuentra dificultades en escenarios complejos, sobre todo cuando se aumenta la cantidad de tipos celulares<sup>43</sup>. En la publicación de Linseed (2019), se menciona una limitación donde si la variabilidad en las proporciones de las subpoblaciones es menor que su similitud transcripcional, o si dos tipos celulares covarían constantemente en todas las muestras, el método no podrá discriminarlos y los tratará como un único "supertipo"<sup>51</sup>.

En adición, CDseqR, mostró un ajuste aceptable únicamente en condiciones balanceadas. Su rendimiento se redujo drásticamente en escenarios con alta heterogeneidad o dominancia de uno o pocos tipos celulares. Si bien su enfoque integral, que permite estimar tanto proporciones como perfiles de expresión sin necesidad de referencias, constituye una ventaja teórica, su robustez se ve claramente comprometida en entornos con complejidad estructural elevada, por ejemplo, en escenarios muy heterogéneos o con tipos celulares dominantes, la simplificación automática que es capaz de realizar CDseqR con perfiles de expresión génica (GEPs) altamente correlacionados, podría llevar a una pérdida de la especificidad y precisión de la estimación. Por otra parte, para ambos métodos RF la discrepancia puede deberse a: 1. El diseño de los *pseudobulks* empleados, ya que es posible que al construir los pseudobulks se introduzca un sesgo de contenido de mRNA vs tamaño celular<sup>55</sup>. 2. El número de tipos celulares presentes ( $K$ ), debido a que si  $K$  es incorrecto (demasiado alto o bajo), la deconvolución puede generar señales artificiales o fusionar tipos celulares biológicamente distintos<sup>57</sup>. 3. La estrategia de anotación de los resultados porque si la anotación es incompatible con la complejidad real del tejido o si los

componentes biológicos no se mapean unívocamente a etiquetas predefinidas, la precisión percibida del método disminuirá en escenarios complejos o heterogéneos<sup>57</sup>.

Sin embargo, los métodos RF destacan por su flexibilidad y capacidad de descubrimiento, ya que, al no depender de matrices de referencia predefinidas, pueden adaptarse mejor a la complejidad celular de los tejidos e incluso identificar nuevos tipos celulares o marcadores específicos del contexto, como ocurre en el microambiente tumoral<sup>10</sup>. De allí el interés de nuestro trabajo de evaluar su desempeño. Ejemplos como el *DREAM Challenge* han mostrado que enfoques no supervisados, como ICA, pueden igualar o incluso superar a los métodos basados en referencia bajo condiciones subóptimas<sup>57</sup>.

La deconvolución de datos de expresión génica tiene aplicaciones clínicas y prácticas fundamentales, especialmente cuando se aprovechan grandes repositorios genómicos como TCGA o GEO<sup>13,58</sup>. Estos enfoques han facilitado la inferencia de perfiles de expresión génica específicos por tipo celular y, de manera crucial, la correlación de fracciones celulares como hepatocitos, colangiocitos, células B maduras o células T reguladoras con desenlaces clínicos como la supervivencia global (OS) y la supervivencia libre de enfermedad (DFS), particularmente en cohortes como TCGA-LIHC<sup>13</sup>.

En conjunto, la capacidad de estos métodos para descomponer datos *bulk* RNA-seq y vincular sus resultados con información clínica y genómica ofrece una vía poderosa para el descubrimiento de nuevas dianas terapéuticas y una comprensión más profunda de la biología del cáncer, haciendo uso de los recursos públicos en los repositorios de datos genómicos.

CDseqR

# **5 Capítulo 2: Estimación de proporciones celulares mediante deconvolución y su asociación con la supervivencia en cáncer de ovario seroso de alto grado a partir de datos de TCGA**

## **5.1 Introducción**

En este capítulo se realiza la estimación de proporciones celulares en muestras de HGSOC y su posible asociación con la supervivencia, utilizando datos de TCGA. El HGSOC es un tipo de cáncer altamente complejo y heterogéneo, caracterizado por mal pronóstico y por un microambiente tumoral (TME) que desempeña un papel clave en la progresión de la enfermedad y la respuesta terapéutica. Identificar la composición celular del TME es fundamental para la identificación de nuevos biomarcadores y posibles dianas terapéuticas<sup>28,45</sup>.

En este contexto, el objetivo principal del capítulo es aplicar los métodos de deconvolución previamente seleccionados a las cohortes de pacientes con HGSOC disponibles en el proyecto TCGA-OV. Se procederá a estimar las proporciones de los distintos tipos celulares presentes en el TME de estas muestras y se analizará su asociación con los desenlaces de supervivencia de las pacientes. La correlación entre la composición del TME y los resultados clínicos permitirá explorar el potencial valor pronóstico de las proporciones celulares, así como su posible utilidad en el contexto de la medicina de precisión para el HGSOC.

## 5.2 Metodología

### 5.2.1 Procesamiento de datos

Los datos de expresión génica de cáncer de ovario seroso de alto grado fueron descargados del repositorio TCGA a través del paquete TCGAbiolinks<sup>64</sup>. Se obtuvieron matrices de conteos de RNA-seq procesadas con el flujo STAR-Counts, junto con metadatos clínicos asociados. Las muestras fueron filtradas para incluir únicamente tumores serosos primarios con información de supervivencia válida, pacientes mayores de 20 años, estadios clínicos FIGO IIIC o IV, raza blanca y tiempo de seguimiento superior a 365 días.

Los conteos crudos fueron procesados utilizando edgeR<sup>65</sup> y biomaRt<sup>66</sup>. Los identificadores Ensembl fueron convertidos a símbolos HGNC, colapsando genes duplicados por suma de conteos. La normalización se realizó mediante el método TMM (*Trimmed Mean of M-values*) y se aplicó un filtrado de baja expresión reteniendo genes con al menos 1 CPM en un mínimo de 10 muestras. El control de calidad incluyó la inspección del número de genes expresados por muestra y diagramas de cajas y bigotes. Finalmente, se exportaron matrices de expresión filtrada en formato crudo y transformado (logCPM) para los análisis posteriores.

### 5.2.2 Deconvolución

La deconvolución de las muestras de HGSOC obtenidas de TCGA se realizó utilizando Linseed, CDseqR, CIBERSORTx y TOAST manteniendo para cada algoritmo los parámetros descritos en el capítulo anterior, con el objetivo de mantener la comparabilidad entre los métodos.

### 5.2.3 Análisis de supervivencia

Las proporciones celulares estimadas mediante métodos de deconvolución fueron integradas con los datos clínicos de pacientes de la cohorte TCGA-OV. Para cada tipo celular, se aplicó la prueba no paramétrica de Wilcoxon con el objetivo de comparar las distribuciones entre pacientes vivos y fallecidos. Los valores de  $p$  obtenidos fueron

## 32 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

---

ajustados mediante el método de corrección FDR, con el fin de identificar diferencias estadísticamente significativas en función del estado vital.

Posteriormente, las muestras fueron estratificadas en cuartiles según la proporción estimada para cada tipo celular. Se generaron curvas de supervivencia global y se aplicaron pruebas de *log-rank* para comparar la distribución de supervivencia entre los grupos. Además, se realizaron comparaciones específicas entre el primer (Q1) y el cuarto cuartil (Q4) para cada tipo celular.

## 5.3 Resultados

### 5.3.1 Deconvolución

A partir del repositorio TCGA se descargaron inicialmente datos de 421 muestras correspondientes a cáncer de ovario. Luego se seleccionaron únicamente aquellos casos con diagnóstico confirmado de carcinoma seroso y tumor primario, lo que redujo el conjunto a 276 muestras. De estas, 275 presentaban información completa de supervivencia, luego se aplicó un filtro adicional por edad y se excluyeron los casos sin dato reportado o con pacientes menores de 20 años al momento del diagnóstico (para excluir casos atípicos de CO pediátricos), se conservaron 270 casos. Posteriormente, la cohorte fue restringida a pacientes con enfermedad en estadios avanzados (FIGO IIIC/IV), etnia blanca (raza blanca, TCGA) y una supervivencia global mayor a 365 días. Este último filtro se aplicó con el fin de excluir casos con muertes tempranas, que pueden deberse a complicaciones perioperatorias, comorbilidades o factores ajenos al curso natural de la enfermedad. Esto condujo a una cohorte final de 150 muestras con datos clínicos y de expresión válidos (Anexo D). Las matrices de expresión génica y los metadatos clínicos correspondientes fueron consolidados y exportados para su uso en los análisis de deconvolución posteriores.

La aplicación de los cuatro métodos de deconvolución a las 150 muestras de TCGA-OV mostró discrepancias en las composiciones celulares estimadas por cada método (Figura 5-1). CIBERSORTx estimó un microambiente tumoral dominado por células epiteliales

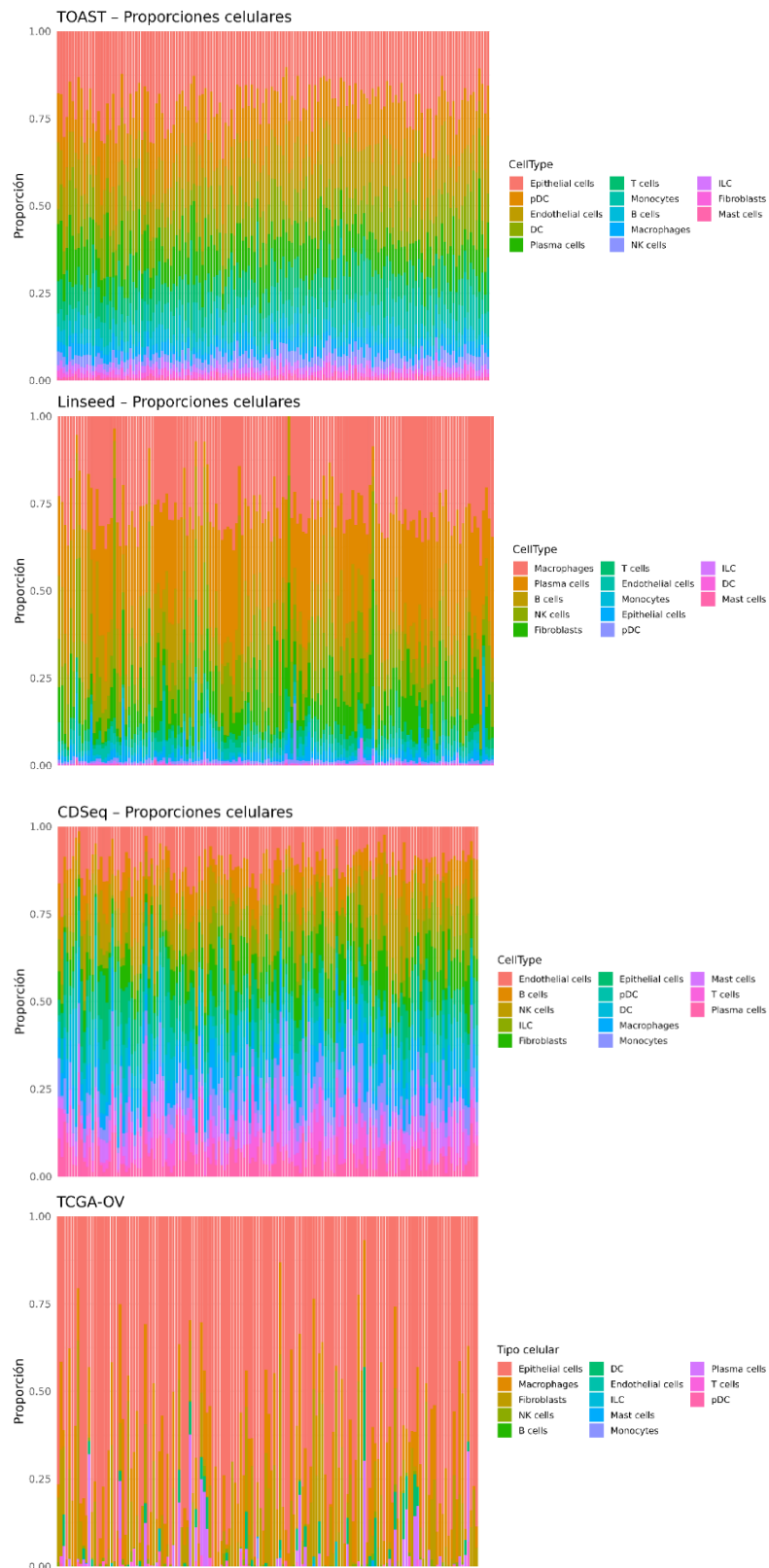
(~68%), con una contribución relevante de fibroblastos (~11%) y macrófagos (~16%), mientras que las restantes poblaciones celulares se encontraron en menor proporción o ausentes.

TOAST, en contraste, ofreció un panorama más heterogéneo, con una menor proporción de células epiteliales (~21%) y una mayor representación de células inmunes y vasculares, incluyendo pDC (~15%), células endoteliales (~14%), *T cells* (~8%) y DC (~10%). Este patrón sugiere una mayor sensibilidad del algoritmo a señales menos dominantes y podría reflejar una estimación más diversificada del microambiente tumoral, aunque con potencial riesgo de sobreestimación de tipos celulares minoritarios.

Linseed, por su parte, no logró capturar adecuadamente la heterogeneidad esperada: las estimaciones fueron prácticamente uniformes entre muestras y linajes, con escasa diferenciación en la composición celular, lo que indica una limitada capacidad de resolución en este tipo de datos *bulk*. Finalmente, CDSeqR retornó proporciones cercanas a cero para la mayoría de los tipos celulares, con escasa variabilidad entre muestras, lo que sugiere una falla de convergencia o una inadecuación del modelo al contexto del dataset de TCGA-OV.

**Figura 5-1. Proporciones celulares estimadas mediante deconvolución con TOAST, Linseed, CDseqR y CIBERSORTx en muestras de pacientes obtenidas de TCGA.** Proporciones celulares estimadas con TOAST, Linseed, CDSeqR y CIBERSORTx a partir de 150 muestras de pacientes con cáncer de ovario (TCGA-OV). Cada barra corresponde a una muestra individual, y la altura proporcional de los segmentos representa la fracción estimada de cada tipo celular en la mezcla.

34 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq



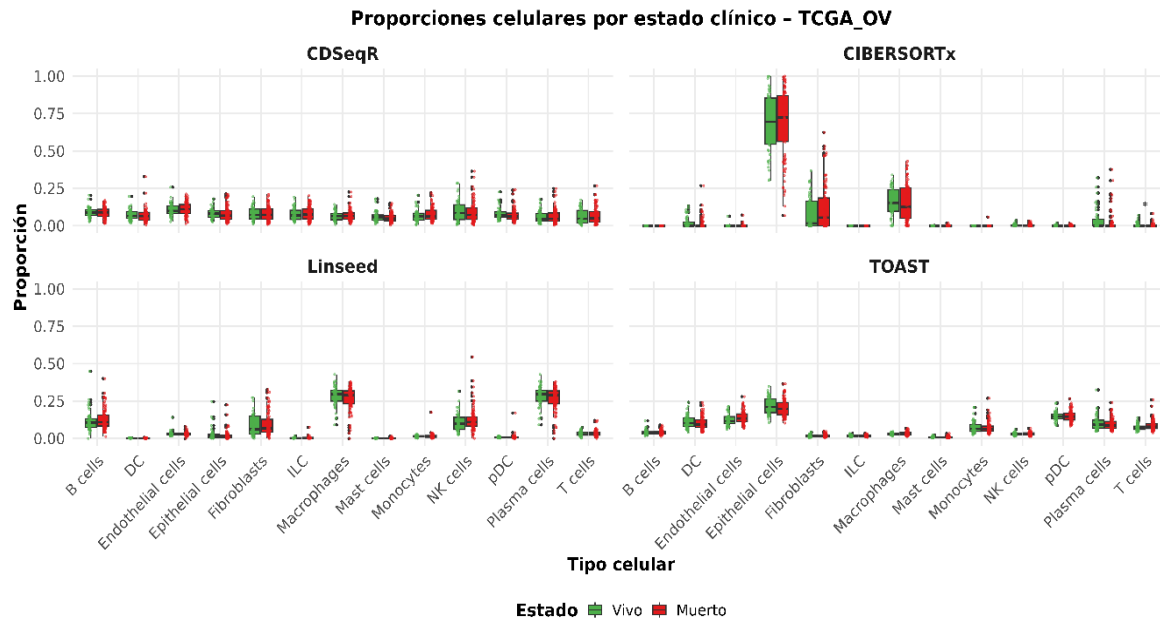
Estas diferencias entre métodos en la estimación de composiciones celulares impactan directamente el análisis clínico. En cáncer de ovario, se ha asociado la abundancia de fibroblastos y macrófagos con un microambiente inmunosupresor y peor pronóstico, mientras que la infiltración de linajes inmunes como células T o pDC puede correlacionarse con una mejor respuesta terapéutica<sup>26,28</sup>. Por ello, más allá de caracterizar la arquitectura del microambiente tumoral, es fundamental explorar cómo estas proporciones estimadas se relacionan con la supervivencia de las pacientes. Este vínculo se analiza en la siguiente sección mediante curvas de Kaplan-Meier, utilizando las proporciones inferidas por cada método de deconvolución.

### 5.3.2 Análisis de supervivencia

Al comparar las proporciones celulares inferidas entre pacientes vivos y fallecidos, se observa lo mencionado anteriormente: CIBERSORTx y TOAST estimaron una mayor composición de *Epithelial cells* (Figura 5-2). Sin embargo, para la mayoría de los tipos celulares, estos métodos no mostraron diferencias estadísticamente significativas entre ambos grupos. Tras la corrección por comparaciones múltiples, se identificó una diferencia estadísticamente significativa con CIBERSORTx en la proporción de *T cells*.

**Figura 5-2. Comparación de proporciones celulares estimadas por TOAST, Linseed, CDseqR y CIBERSORTx entre pacientes vivos (verde) y fallecidos (rojo) de la cohorte obtenida de TCGA-OV.** Cada panel corresponde a un método de deconvolución, y las cajas representan la distribución de fracciones estimadas para cada tipo celular.

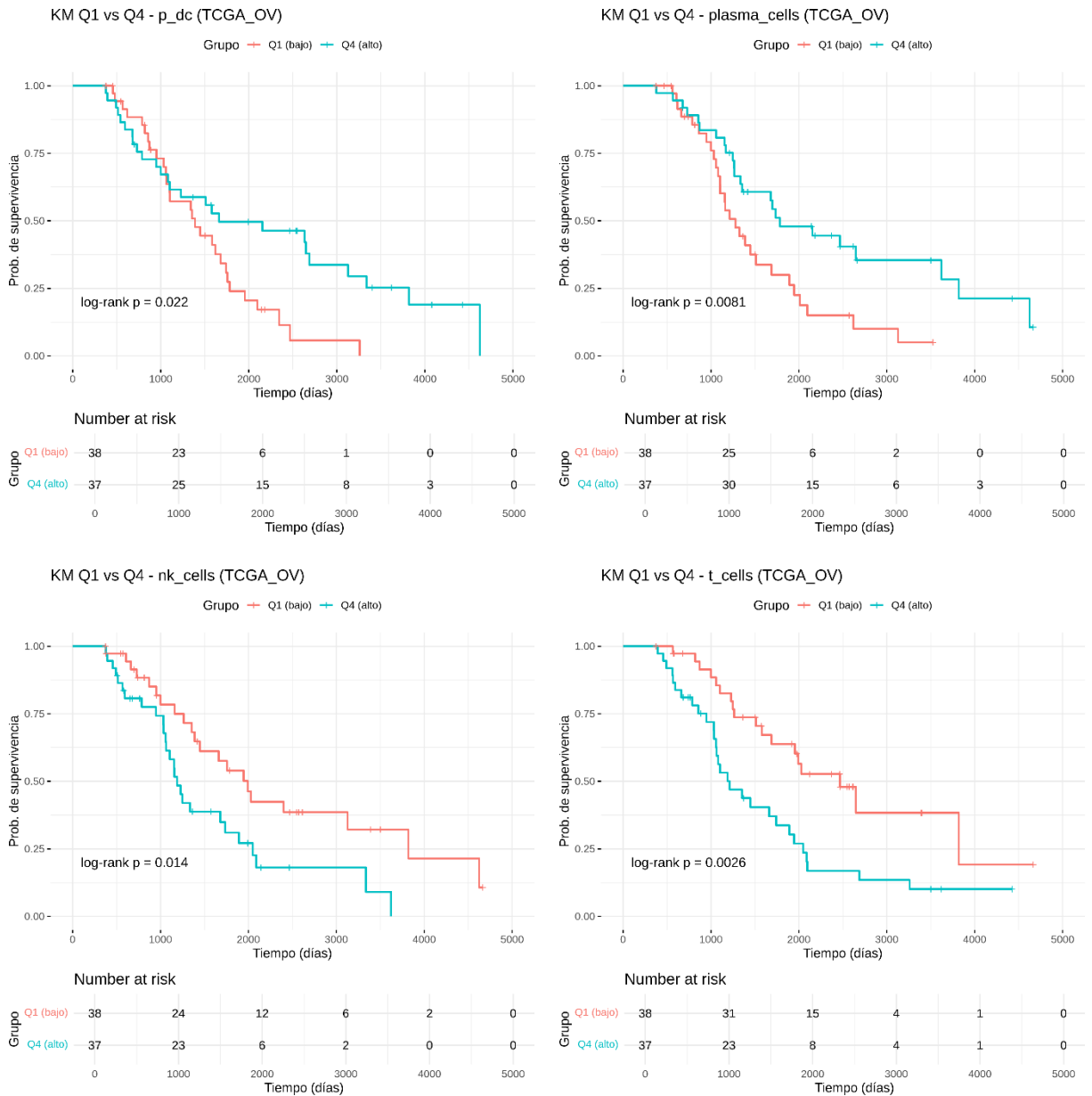
36 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq



Debido a que la comparación directa de proporciones celulares entre pacientes vivos y fallecidos no reveló diferencias significativas en la mayoría de los métodos, se optó por realizar análisis de Kaplan-Meier (KM) estratificando las muestras según los cuartiles extremos (Q1 vs Q4) de abundancia celular. Los resultados muestran que las asociaciones más consistentes provinieron de los métodos basados en referencia, mientras que Linseed y CDseqR no identificaron ningún linaje con diferencias significativas en las curvas de supervivencia (Tabla 5-1), en línea con sus limitaciones previamente observadas en la estimación de proporciones celulares.

El análisis de Kaplan-Meier mostró diferencias significativas entre la composición inmune estimada por TOAST y la supervivencia global de las pacientes (Figura 5-3). En particular, una mayor proporción de *Plasma cells* y pDC se asoció con una supervivencia más prolongada ( $\Delta = +505$  y  $+273$  días, respectivamente), mientras que un aumento en *T cells* y *NK cells* se vinculó con un peor desenlace clínico ( $\Delta = -1,254$  y  $-804$  días, respectivamente). Estos hallazgos sugieren que, a pesar de su variabilidad en la estimación de proporciones, TOAST fue capaz de capturar señales funcionales relevantes del microambiente tumoral, distinguiendo linajes asociados tanto a mejor como a peor pronóstico en cáncer de ovario.

**Figura 5-3. Curvas de Kaplan–Meier para tipos celulares con asociaciones significativas de supervivencia estimadas por TOAST en TCGA-OV.** Curvas de Kaplan–Meier de la cohorte TCGA-OV estratificadas por cuartiles (Q1 = baja, Q4 = alta proporción) para los tipos celulares cuya proporción estimada mediante TOAST mostró una asociación significativa con la supervivencia global. Se muestran los valores de  $p$  de la prueba *log-rank* y el número de pacientes en riesgo a lo largo del tiempo.

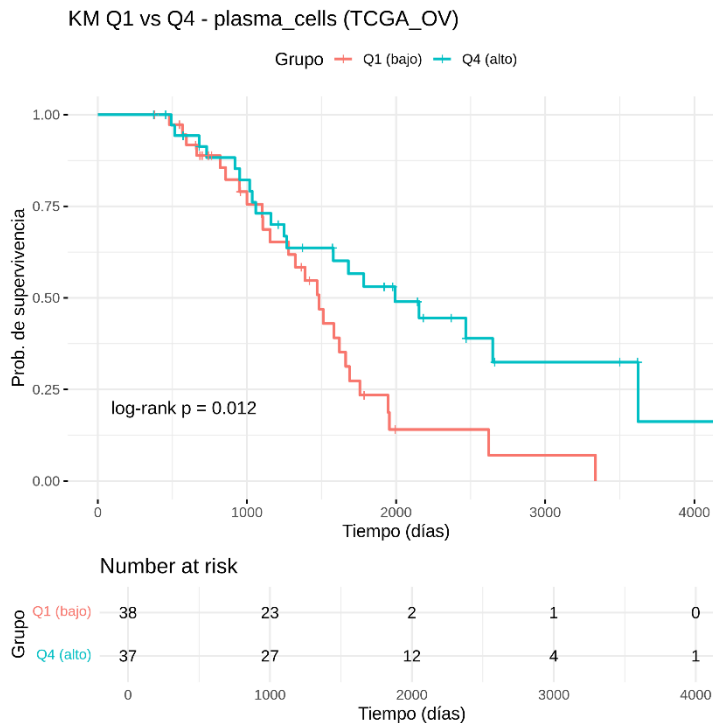


38 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

---

En el caso de CIBERSORTx (Figura 5-4), el único linaje con asociación significativa fue el de las *Plasma cells*, donde los pacientes del cuartil alto (Q4) mostraron una mediana de supervivencia aproximadamente 500 días superior a los del cuartil bajo (Q1; *log-rank p* = 0.012). Siendo este resultado consistente con obtenido para el mismo tipo celular al hacer la deconvolución con TOAST.

**Figura 5-4. Curva de Kaplan–Meier para el tipo celular con asociación significativa de supervivencia estimado por CIBERSORTx en TCGA-OV.** Curva de Kaplan–Meier de la cohorte TCGA-OV estratificada por cuartiles (Q1 = baja, Q4 = alta proporción) para el tipo celular cuya proporción estimada mediante CIBERSORTx mostró una asociación significativa con la supervivencia global. Se muestran los valores de *p* de la prueba *log-rank* y el número de pacientes en riesgo a lo largo del tiempo.



**Tabla 5-1. Resultados del análisis de supervivencia según proporciones celulares inferidas por distintos métodos de deconvolución en TCGA-OV.** Resultados del análisis de supervivencia global mediante curvas de Kaplan–Meier estratificadas por cuartiles extremos (Q1 vs. Q4) de proporciones celulares estimadas con CDSeqR, CIBERSORTx, Linseed y TOAST en la cohorte TCGA-OV. Se muestran los valores de *p* (*log rank*) para las comparaciones entre grupos.

Método	Tipo celular	p (log-rank)	Mediana Q1	Mediana Q4	Δ mediana (días)
CDSeqR	ilc	0.0605	1,784	1,354	-430
	monocytes	0.2507	1,579	1,336	-243
	plasma cells	0.2983	1,620	1,946	326
	p_dc	0.3418	1,324	1,511	187
	<i>Fibroblasts</i>	0.3614	1,742	1,229	-513
	nk cells	0.4360	1,736	1,579	-157
	b cells	0.4384	1,448	1,484	36
	endothelial cells	0.5083	1,492	1,680	188
	t cells	0.5773	1,355	1,680	325
	<i>Macrophages</i>	0.7707	1,583	1,354	-229
	mast cells	0.8101	1,579	1,511	-68
	epithelial cells	0.8132	1,336	1,389	53
dc	0.9284	1,511	1,484	-27	
CIBERSORTx	plasma cells	0.0124*	1,484	1,993	509
	dc	0.0726	1,583	2,028	445
	t cells	0.0761	1,583	2,089	506
	nk cells	0.3527	1,742	1,336	-406
	epithelial cells	0.4894	1,891	1,511	-380
	<i>Fibroblasts</i>	0.7041	1,389	1,680	291
	<i>Macrophages</i>	0.7505	1,620	1,470	-150
Linseed	endothelial cells	0.0634	1,354	1,699	345
	epithelial cells	0.0847	1,354	1,784	430
	p_dc	0.1618	1,492	2,634	1,142
	<i>Macrophages</i>	0.2480	1,736	1,354	-382
	plasma cells	0.2480	1,736	1,354	-382
	<i>Fibroblasts</i>	0.4100	1,355	1,680	325
	b cells	0.4436	1,470	1,736	266
	monocytes	0.6605	1,662	1,229	-433
	nk cells	0.7314	1,993	1,757	-236
t cells	0.8850	1,662	1,341	-321	
TOAST	t cells	0.0026**	2,467	1,213	-1,254
	plasma cells	0.0081**	1,279	1,784	505
	nk cells	0.0141*	1,993	1,189	-804
	p_dc	0.0223*	1,389	1,662	273
	ilc	0.0849	1,354	1,757	403
	<i>Macrophages</i>	0.1651	1,742	1,389	-353
	endothelial cells	0.3732	1,993	1,736	-257
	<i>Fibroblasts</i>	0.4100	1,355	1,688	333
monocytes	0.4970	1,757	1,492	-265	

40 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

---

	epithelial_cells	0.6313	1,662	1,757	95
	dc	0.6315	1,688	1,448	-240
	mast_cells	0.6498	1,484	1,213	-271
	b_cells	0.6559	1,511	1,470	-41

\* p < 0.05, \*\* p < 0.01

## 5.4 Discusión

EL CO es una patología que se diagnostica de manera tardía debido a la inespecificidad de sus síntomas, presenta alta recurrencia y el 70% de los pacientes generan resistencia a los tratamientos convencionales, por cual tiene una alta tasa de mortalidad<sup>24</sup>. Entender complejidad del microambiente celular y transcripcional de este cáncer permite identificar posibles marcadores pronósticos con el propósito de mejorar la supervivencia<sup>25,26,28</sup>. Por esta razón el enfoque de este trabajo fue estudiar la relación entre la composición celular del TME y la supervivencia del CO. A pesar de que no todos los métodos de deconvolución lograron detectar asociaciones claras con la supervivencia, TOAST y CIBERSORTx sí identificaron señales con posible relevancia clínica. Esto refuerza la idea de que, a pesar de las limitaciones de estos métodos, la estimación de proporciones celulares a partir de RNA-seq de tejido completo puede aportar información valiosa sobre el pronóstico, especialmente cuando se emplean herramientas con capacidad de capturar con mayor fidelidad la diversidad funcional del TME.

Uno de los resultados más consistentes fue la asociación positiva entre la infiltración de células plasmáticas y una mayor supervivencia, observada tanto con TOAST como con CIBERSORTx. Este hallazgo está en línea con lo reportado en la literatura, donde las células B y plasmáticas han sido vinculadas con una respuesta inmune antitumoral efectiva. Por ejemplo, la presencia de estructuras linfoides terciarias (TLS) en el HGSOC se ha asociado con mejor pronóstico, ya que promueven la maduración de células B y la activación de células T citotóxicas<sup>27,28</sup>.

Por otra parte, los resultados obtenidos con TOAST para otros linajes, fueron sorprendentes ya que las células T y NK se asociaron con peor pronóstico, lo que contrasta con gran parte de la evidencia previa. Tradicionalmente, la infiltración de células T CD8+ y CD4+ se ha

relacionado con mejor supervivencia en el CO, al igual que la actividad citotóxica de las células NK<sup>24,25,67</sup>. Sin embargo, estos resultados podrían reflejar fenómenos biológicos más complejos. En el CO, el TME suele estar fuertemente inmunosuprimido, favoreciendo el agotamiento de células T por la acción de Tregs, MDSCs y TAMs, así como por la presencia de las citocinas IL-10 y TGF- $\beta$ <sup>24,67</sup>. Además, se ha reportado que una alta densidad de células T CD8+ no siempre es sinónimo de buen pronóstico, sobre todo si están co-localizadas con macrófagos PD-L1+, lo que podría promover su disfunción. Este tipo de interacciones, clave para entender el efecto real de los infiltrados inmunes, escapa a la estimación de los métodos de deconvolución sobre datos *bulk*<sup>68</sup>.

En el caso de las células NK, la asociación negativa podría explicarse por la inmunosupresión de su función en el TME del CO<sup>24</sup>. Aunque generalmente las células NK están presentes en grandes cantidades, especialmente en el líquido ascítico, muchas de ellas pierden su capacidad citotóxica debido a señales inmunosupresoras locales<sup>24,25</sup>. Además, ciertos subtipos de células NK incluso han sido asociados con funciones pro-tumorales, como la promoción de angiogénesis, lo que podría explicar que su abundancia esté relacionada con un peor desenlace clínicos<sup>25</sup>.

Por otro lado, las células dendríticas plasmocitoides (pDC) mostraron una relación positiva con la supervivencia, lo cual es coherente con su rol en la activación de respuestas inmunes adaptativas<sup>7,28</sup>. Aunque algunas funciones de las pDCs pueden verse afectadas en el TME, su presencia se ha asociado a un mayor potencial inmunogénico en HGSOE.

Como se mencionó anteriormente, solo los métodos PRF y RB lograron identificar proporciones celulares con impacto potencial en la supervivencia de las pacientes, mientras que los métodos RF no mostraron el mismo desempeño. Esta diferencia puede explicarse por la capacidad de los enfoques PRF y RB de aprovechar información biológica previa, en particular la selección de marcadores específicos de tipo celular, lo que refuerza la precisión de la deconvolución<sup>11</sup>. En contraste, Linseed y CDSeqR, al operar sin referencias externas, reproducen las limitaciones señaladas en el capítulo anterior: en escenarios de alta complejidad biológica tienden a mezclar señales o fusionar linajes debido a la colinealidad en los patrones de expresión de ciertos genes, interpretándolos como un único tipo celular<sup>51</sup>. Este fenómeno reduce su sensibilidad para detectar

42 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

---

diferencias en proporciones celulares que podrían asociarse de manera significativa con la supervivencia.

Por otra parte, CIBERSORTx y TOAST a pesar de que fueron ejecutados usando como referencia el mismo *dataset* de referencia de célula única mostraron resultados diferentes y asociación de tipos celulares con la supervivencia diferentes, compartiendo solo la asociación significativa de las plasma cells. CIBERSORTx se apoya en v-SVR, un enfoque robusto al ruido que selecciona automáticamente genes relevantes y maneja la multicolinealidad entre tipos celulares con perfiles similares<sup>11</sup>. En contraste, TOAST recurre a un proceso iterativo de selección de características basado en NMF para posteriormente llevar a cabo una deconvolución RF<sup>48,49</sup>. Con base en esta diferencia en los algoritmos, CIBERSORTx y TOAST identifican proporciones y tipos celulares diferentes, el primero puede identificar tipos celulares con firmas génicas más robustas en el perfil de referencia, mientras que el segundo al ser PRF puede ser más flexible en la estimación e identificar tipos y proporciones celulares diferentes que puede que no sean tan explícitas en el perfil de referencia.

Los resultados de este estudio evidencian que los métodos de deconvolución evaluados comparten una limitación en no poder captar la organización espacial del microambiente tumoral ni distinguir con precisión entre subpoblaciones funcionalmente distintas dentro de un mismo linaje celular. Por ejemplo, cuando TOAST asocia una mayor proporción de células T o NK con peor supervivencia, es importante considerar que estas categorías engloban tanto células efectoras (T CD8+ activadas o NK citotóxicas) como poblaciones inmunosupresoras o funcionalmente exhaustas (Tregs o NK reprimidas). Esta falta de sensibilidad podría explicar las asociaciones que, en un primer análisis, resultan inesperadas o contraintuitivas.

En resumen, la deconvolución de datos de RNA-seq a partir de tejido completo representa una herramienta útil para explorar la composición del TME y generar hipótesis sobre su impacto clínico. Aunque presenta ciertas limitaciones inherentes, su utilidad podría aumentar cuando se integre con datos multiómicos, análisis espaciales como la

transcriptómica espacial, y datos clínicos que permitan enriquecer las hipótesis obtenidas a partir de la deconvolución.

## 6 Conclusiones y recomendaciones

### 6.1 Conclusiones

Este estudio permitió comparar varios métodos de deconvolución de distintos tipos, aplicándolos tanto a *pseudobulks* generados a partir de datos de scRNA-seq como a muestras reales de RNA-seq de la cohorte TCGA-OV de cáncer de ovario seroso de alto grado.

En la primer parte, al evaluar los métodos sobre *pseudobulks* con proporciones conocidas, se observaron diferencias claras en el rendimiento. CIBERSORTx fue el más consistente, mostrando las mejores correlaciones y los menores errores en casi todos los escenarios simulados. TOAST tuvo un desempeño aceptable, especialmente cuando las proporciones celulares estaban equilibradas, pero su precisión disminuyó en mezclas más desiguales. Por otro lado, los métodos sin referencia (Linseed y CDSeqR) enfrentaron mayores dificultades: Linseed funcionó bien solo en contextos simples, mientras que CDSeqR tuvo problemas constantes para recuperar las proporciones reales. En particular, la dependencia de estos métodos RF de la geometría del simplex y de la convergencia sin referencias externas explica en gran medida estas limitaciones en escenarios con alta colinealidad, como ocurre en tumores sólidos heterogéneos.

Al aplicar estos métodos a las muestras *bulk* de TCGA-OV, los algoritmos basados en referencia volvieron a destacar. CIBERSORTx identificó un entorno tumoral dominado por células epiteliales, fibroblastos y macrófagos, mientras que TOAST reflejó una mayor diversidad de tipos celulares inmunes y vasculares. En cambio, Linseed y CDSeqR no lograron reproducir la heterogeneidad del microambiente ovárico. Esto debido a que métodos como CIBERSORTx y TOAST se apoyan en referencias previas o conocimiento a priori que actúan como guías durante la estimación, lo cual les permite generar resultados

más estables y coherentes. Mientras que los métodos RF en este contexto con datos reales siguen mostrando las limitaciones observadas en los *pseudobulks*.

En cuanto a la supervivencia, los métodos basados en referencia detectaron asociaciones relevantes. Con TOAST se encontró que una mayor presencia de células plasmáticas y pDC se relacionaba con una mejor supervivencia, mientras que niveles altos de células T y NK se asociaron con peores desenlaces. De manera similar, CIBERSORTx también resaltó el efecto positivo de las células plasmáticas. Linseed y CDSeqR no identificaron asociaciones significativas. El hallazgo consistente de la asociación positiva de las células plasmáticas refuerza su potencial como marcador pronóstico en HGSOC, en concordancia con estudios que vinculan la respuesta inmune humoral con mejores desenlaces en diversos tumores sólidos. En contraste, la asociación adversa observada para las células T y NK podría reflejar estados disfuncionales o de agotamiento inmunológico, lo que subraya la importancia de caracterizar funcionalmente estas subpoblaciones, más allá de su simple abundancia relativa. .

En conjunto, este trabajo sugiere que la deconvolución computacional es una herramienta poderosa para investigar el microambiente tumoral y su impacto clínico. Sin embargo, también deja claro que sus estimaciones deben interpretarse con cuidado, y preferiblemente no de manera aislada sino procurando su integración con datos adicionales para una mayor confianza en los hallazgos y avanzar hacia una medicina de precisión más informada y efectiva.

## 6.2 Recomendaciones

Se sugiere profundizar en la caracterización funcional de subpoblaciones inmunes (tanto para métodos RB como RF) mediante la integración de marcadores específicos que permitan distinguir entre estados o subpoblaciones celulares efectoras, reguladoras o disfuncionales, los cuales podrían tener implicancias pronósticas más específicas. De manera complementaria, los métodos PRF ofrecen la posibilidad de descubrir nuevos estados celulares o firmas de expresión génica (meta-programas) dentro del microambiente tumoral, los cuales no han sido previamente catalogados pero podrían tener un fuerte impacto pronóstico. Estos “fenotipos novedosos” representarían

46 Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq

---

potenciales estados de activación única o interacciones celulares clave para la progresión del cáncer, y su identificación puede guiar hipótesis traslacionales de alto valor, por lo cual este profundizar en este tipo de enfoques surge como una oportunidad interesante desde la bioinformática aplicada a la oncología.

Dado que el cáncer presenta microambiente celular heterogéneo su análisis debería realizarse con métodos híbridas que combinen elementos de métodos RB y RF (similar a lo que propone el método PRF TOAST), para mantener una estructura guiada por el conocimiento previo, sin perder la flexibilidad necesaria para adaptarse a la complejidad de los datos, esto podría mejorar la robustez de las estimaciones al minimizar tanto el sesgo hacia la referencia como la inestabilidad de los modelos completamente no supervisados.

En relación con lo anterior, en lugar de descartar los métodos RF, se recomienda reformularlos hacia enfoques PRF que incorporen conocimiento previo de manera flexible. Esto puede lograrse mediante *sparse regularization*, que obligue a que cada firma celular se defina por un conjunto reducido de genes marcadores, y regularización basada en grafos de co-expresión guiados por anotaciones (GO, KEGG, Reactome), que refuercen la coherencia funcional de las firmas inferidas. Adicionalmente, el uso de inicialización guiada por marcadores canónicos (*Marker-guided initialization*) y criterios de consenso entre múltiples ejecuciones (*Consensus-based stability criteria*) puede mejorar la estabilidad de los resultados del algoritmo. La implementación de estos marcos híbridos permitiría transformar a los métodos RF de herramientas exploratorias poco confiables en algoritmos capaces de descubrir fenotipos celulares novedosos y programas de activación con relevancia pronóstica, cerrando la brecha entre descubrimiento puro y cuantificación validada.

Para futuras investigaciones, se recomienda complementar este enfoque con modelos de supervivencia ajustados por covariables clínicas, como la edad, el estadio FIGO, el estatus de BRCA, entre otras. Asimismo, el uso de modelos predictivos basados en machine learning, como *random forest survival*, *elastic net Cox* o enfoques de *deep learning*

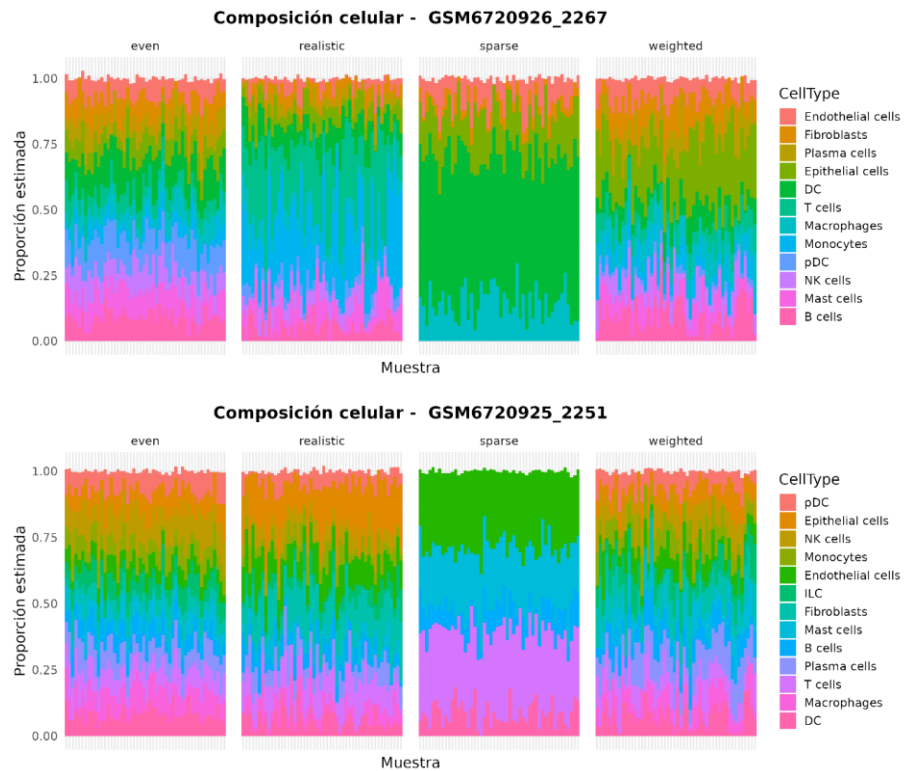
aplicados a datos ómicos con el ánimo de construir modelos pronósticos con mayor nivel de confianza.

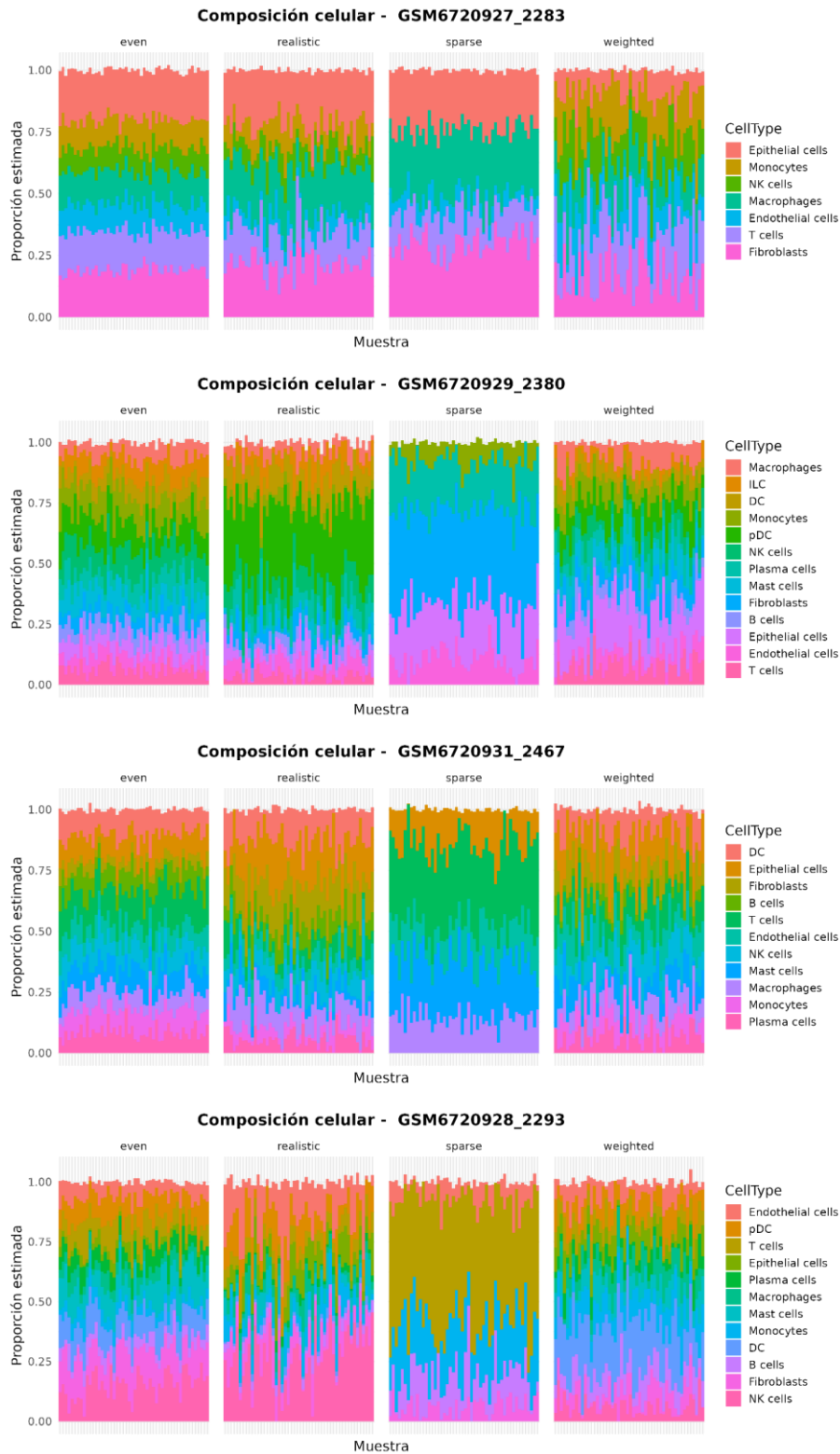
Por último, la combinación de las estimaciones de deconvolución con datos multiómicos y clínicos, incluyendo mutaciones, perfiles proteómicos, metilación e imágenes histológicas, puede enriquecer la interpretación biológica y mejorar su aplicabilidad para predecir desenlaces clínicos y guiar decisiones terapéuticas más precisas.



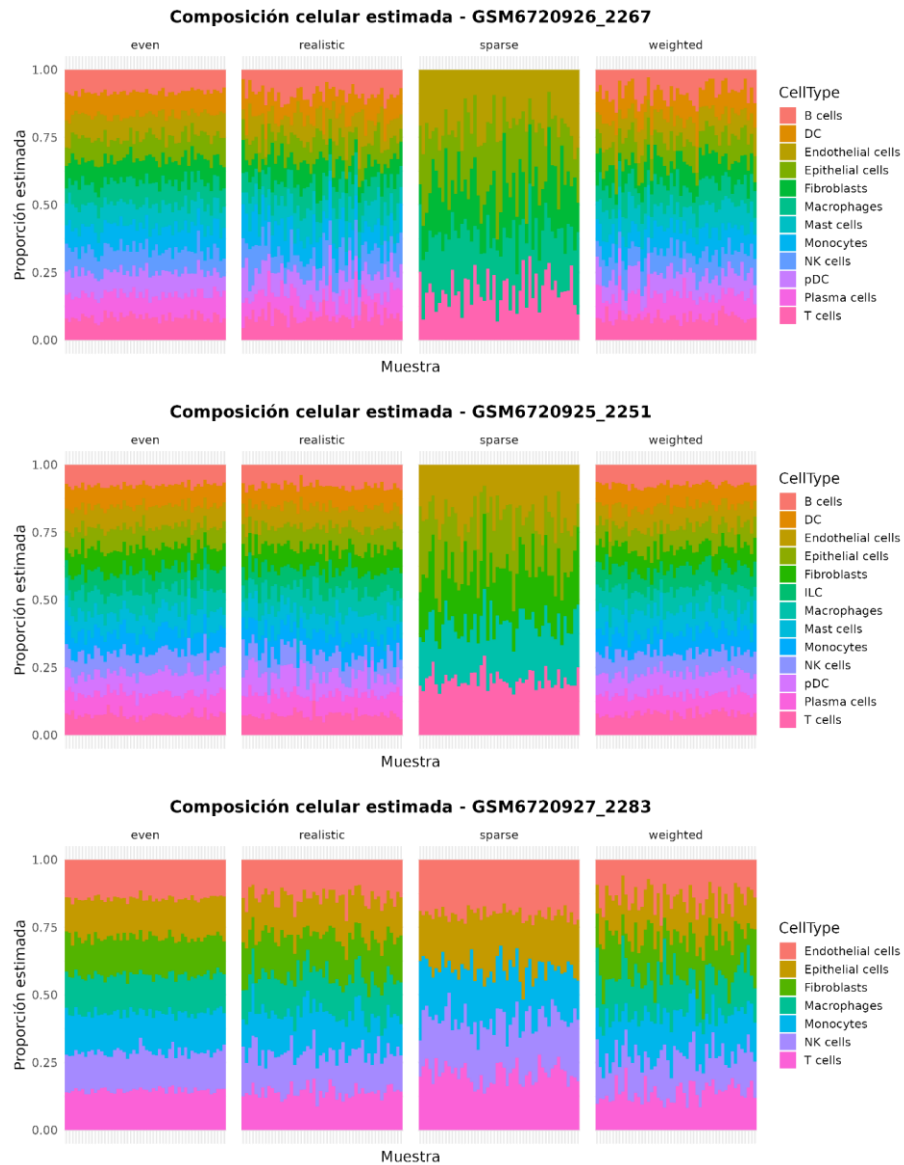
# A. Anexo:

Anexo A. Proporciones celulares estimadas mediante deconvolución con Linseed en *pseudobulks* simulados bajo los escenarios *even*, *realistic*, *sparse* y *weighted*.

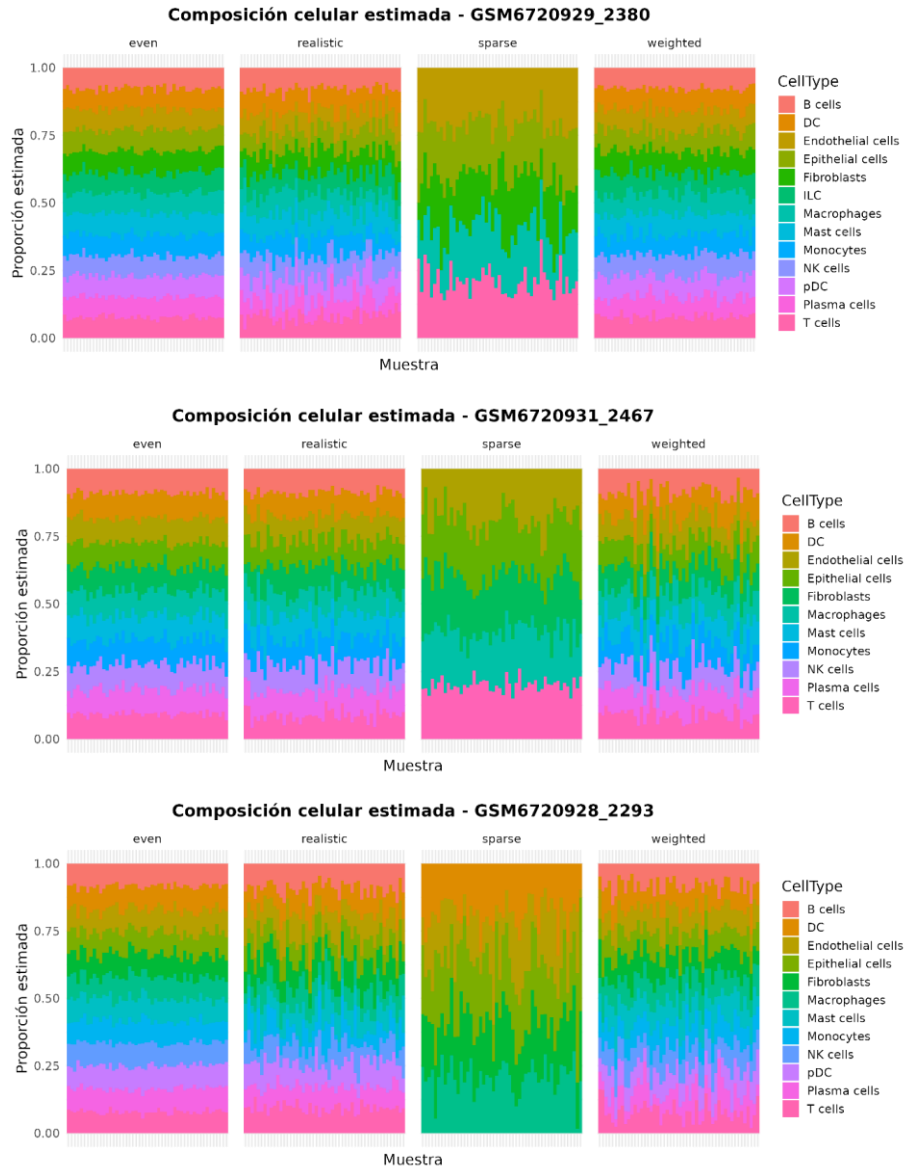




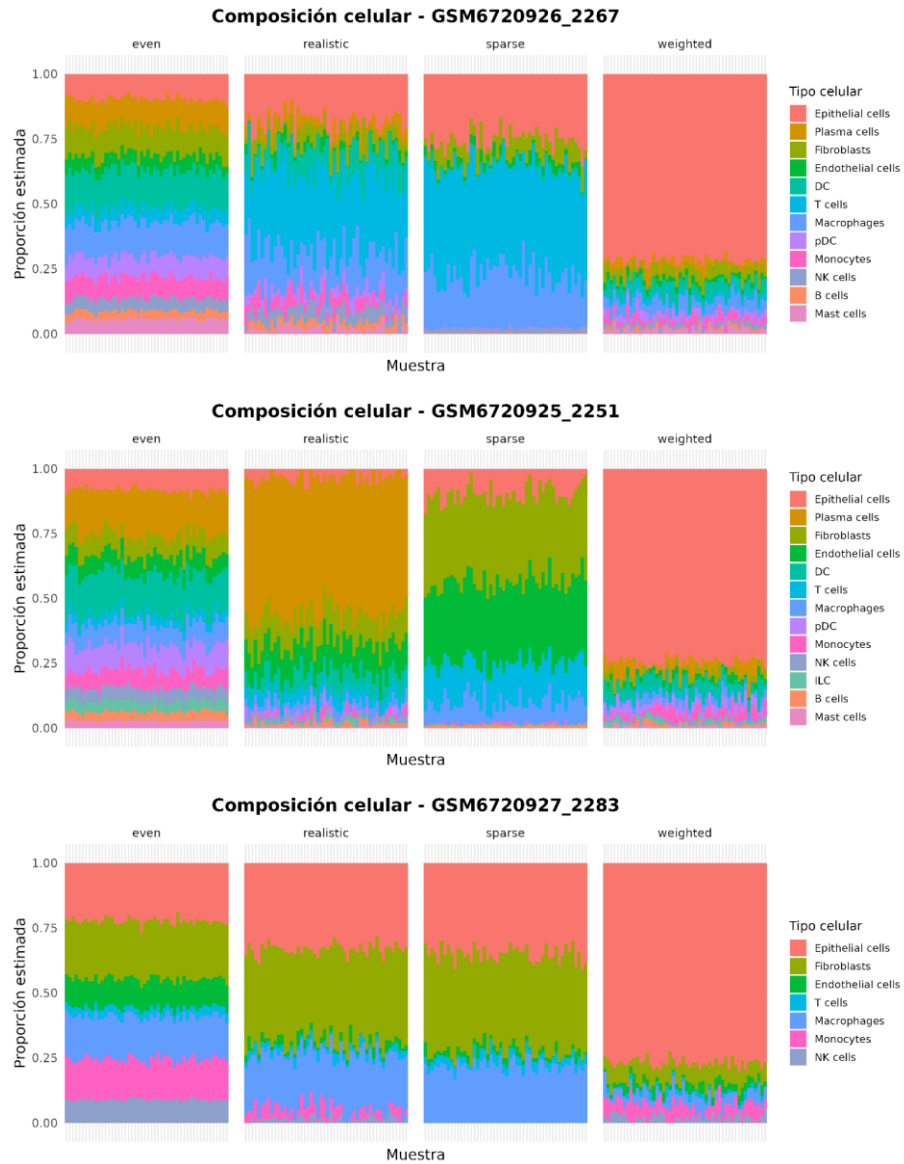
**Anexo B. Proporciones celulares estimadas mediante deconvolución con CDseqR en *pseudobulks* simulados bajo los escenarios *even*, *realistic*, *sparse* y *weighted*.**

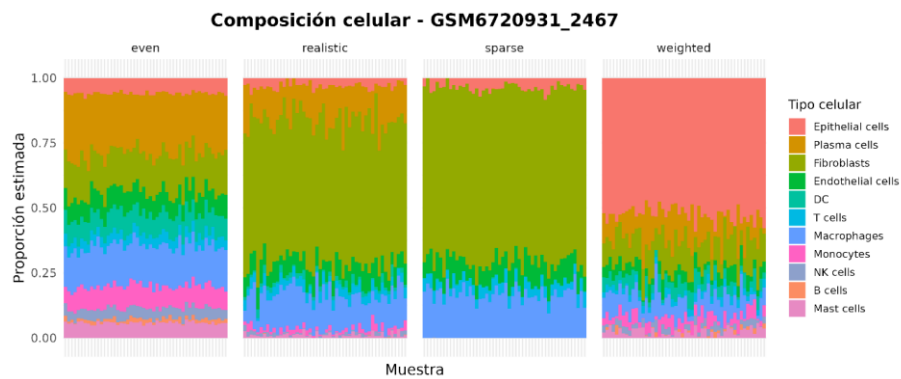
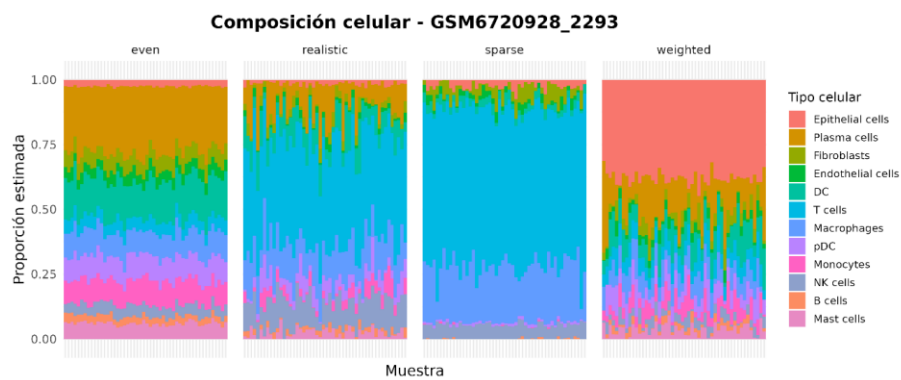
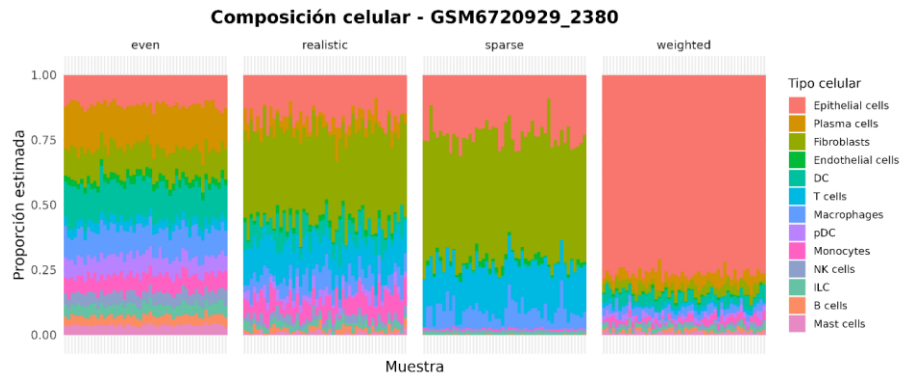


Comparación del desempeño de métodos de deconvolución para la identificación de la composición celular y su asociación con la supervivencia en muestras de cáncer de ovario seroso de alto grado a partir de datos de RNA-seq



**Anexo C. Proporciones celulares estimadas mediante deconvolución con CIBERSORTx en *pseudobulks* simulados bajo los escenarios *even*, *realistic*, *sparse* y *weighted*.**





**Anexo D. Descripción de la cohorte TCGA-OV filtrada.**

<b>Variable</b>	<b>Valor</b>
<b>Número total de muestras</b>	150
<b>Edad (media <math>\pm</math> DE)</b>	60.2 $\pm$ 11.0
<b>Edad (rango)</b>	40 – 88
<b>Supervivencia global (mediana, días)</b>	1330
<b>Seguimiento (rango, días)</b>	374 – 4665
<b>Eventos (fallecidos)</b>	100
<b>Censurados (vivos)</b>	50
<b>Estadio FIGO IIIC</b>	131
<b>Estadio FIGO IV</b>	19
<b>Sexo femenino</b>	150
<b>Sexo masculino</b>	0
<b>Raza blanca</b>	150

## Bibliografía

1. Kotnik EN, Mullen MM, Spies NC, Li T, Inkman M, Zhang J, et al. Genetic characterization of primary and metastatic high-grade serous ovarian cancer tumors reveals distinct features associated with survival. *Commun Biol*. 3 de julio de 2023;6(1):688.
2. Nallasamy P, Nimmakayala RK, Parte S, Are AC, Batra SK, Ponnusamy MP. Tumor microenvironment enriches the stemness features: the architectural event of therapy resistance and metastasis. *Molecular Cancer*. 22 de diciembre de 2022;21(1):225.
3. Yenyuwadee S, Aliazis K, Wang Q, Christofides A, Shah R, Patsoukis N, et al. Immune cellular components and signaling pathways in the tumor microenvironment. *Seminars in Cancer Biology*. 1 de noviembre de 2022;86:187-201.
4. Yang Y, Yang Y, Yang J, Zhao X, Wei X. Tumor Microenvironment in Ovarian Cancer: Function and Therapeutic Strategy. *Frontiers in Cell and Developmental Biology* [Internet]. 2020 [citado 26 de noviembre de 2023];8. Disponible en: <https://www.frontiersin.org/articles/10.3389/fcell.2020.00758>
5. Pernot S, Evrard S, Khatib AM. The Give-and-Take Interaction Between the Tumor Microenvironment and Immune Cells Regulating Tumor Progression and Repression. *Frontiers in Immunology* [Internet]. 2022 [citado 26 de noviembre de 2023];13. Disponible en: <https://www.frontiersin.org/articles/10.3389/fimmu.2022.850856>
6. Bożyk A, Wojas-Krawczyk K, Krawczyk P, Milanowski J. Tumor Microenvironment—A Short Review of Cellular and Interaction Diversity. *Biology*. junio de 2022;11(6):929.
7. Jiang Y, Wang C, Zhou S. Targeting tumor microenvironment in ovarian cancer: Premise and promise. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*. abril de 2020;1873(2):188361.
8. Integrated Genomic Analyses of Ovarian Carcinoma. *Nature*. 29 de junio de 2011;474(7353):609-15.

9. Im Y, Kim Y. A Comprehensive Overview of RNA Deconvolution Methods and Their Application. *Mol Cells*. 28 de febrero de 2023;46(2):99-105.
10. Xu X, Li R, Mo O, Liu K, Li J, Hao P. Cell-type deconvolution for bulk RNA-seq data using single-cell reference: a comparative analysis and recommendation guideline. *Briefings in Bioinformatics*. 22 de noviembre de 2024;26(1):bbaf031.
11. Steen CB, Liu CL, Alizadeh AA, Newman AM. Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. En: Kidder BL, editor. *Stem Cell Transcriptional Networks* [Internet]. New York, NY: Springer US; 2020 [citado 8 de agosto de 2025]. p. 135-57. (Methods in Molecular Biology; vol. 2117). Disponible en: [http://link.springer.com/10.1007/978-1-0716-0301-7\\_7](http://link.springer.com/10.1007/978-1-0716-0301-7_7)
12. White BS, De Reyniès A, Newman AM, Waterfall JJ, Lamb A, Petitprez F, et al. Community assessment of methods to deconvolve cellular composition from bulk gene expression. *Nat Commun*. 27 de agosto de 2024;15(1):7362.
13. Zhang S, Bacon W, Peppelenbosch MP, Van Kemenade F, Stubbs AP. Deciphering Tumour Microenvironment of Liver Cancer through Deconvolution of Bulk RNA-Seq Data with Single-Cell Atlas. *Cancers*. 27 de diciembre de 2022;15(1):153.
14. Jin H, Liu Z. A comparative study of deconvolution methods for RNA-seq data under a dynamic testing landscape [Internet]. *Bioinformatics*; 2020 dic [citado 12 de febrero de 2024]. Disponible en: <http://biorxiv.org/lookup/doi/10.1101/2020.12.09.418640>
15. Caruso G, Weroha SJ, Cliby W. Ovarian Cancer: A Review. *JAMA* [Internet]. 21 de julio de 2025 [citado 13 de septiembre de 2025]; Disponible en: <https://doi.org/10.1001/jama.2025.9495>
16. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*. 2021;71(3):209-49.
17. Kuroki L, Guntupalli SR. Treatment of epithelial ovarian cancer. *BMJ*. 9 de noviembre de 2020;371:m3773.
18. Mhatre A, Koroth J, Manjunath M, Kumar S S, Gawari R, Choudhary B. Multi-omics analysis of the Indian ovarian cancer cohort revealed histotype-specific mutation and gene expression patterns. *Frontiers in Genetics* [Internet]. 2023 [citado 25 de noviembre de 2023];14. Disponible en: <https://www.frontiersin.org/articles/10.3389/fgene.2023.1102114>
19. Huang J, Chan WC, Ngai CH, Lok V, Zhang L, Lucero-Prisno DE, et al. Worldwide Burden, Risk Factors, and Temporal Trends of Ovarian Cancer: A Global Study. *Cancers (Basel)*. 29 de abril de 2022;14(9):2230.
20. Mazidimoradi A, Momenimovahed Z, Allahqoli L, Tiznobaik A, Hajinasab N, Salehiniya H, et al. The global, regional and national epidemiology, incidence,

- 
- mortality, and burden of ovarian cancer. *Health Sci Rep*. noviembre de 2022;5(6):e936.
21. Lawrenson K, Fonseca MAS, Liu AY, Dezem FS, Lee JM, Lin X, et al. A Study of High-Grade Serous Ovarian Cancer Origins Implicates the SOX18 Transcription Factor in Tumor Development. *Cell Reports*. 10 de diciembre de 2019;29(11):3726-3735.e4.
  22. Saida T, Tanaka YO, Matsumoto K, Satoh T, Yoshikawa H, Minami M. Revised FIGO staging system for cancer of the ovary, fallopian tube, and peritoneum: important implications for radiologists. *Jpn J Radiol*. 1 de febrero de 2016;34(2):117-24.
  23. Hong MK, Ding DC. Early Diagnosis of Ovarian Cancer: A Comprehensive Review of the Advances, Challenges, and Future Directions. *Diagnostics (Basel)*. 7 de febrero de 2025;15(4):406.
  24. Ghoneum A, Almousa S, Warren B, Abdulfattah AY, Shu J, Abouelfadl H, et al. Exploring the clinical value of tumor microenvironment in platinum-resistant ovarian cancer. *Seminars in Cancer Biology*. diciembre de 2021;77:83-98.
  25. Rodriguez G, Galpin K, McCloskey C, Vanderhyden B. The Tumor Microenvironment of Epithelial Ovarian Cancer and Its Influence on Response to Immunotherapy. *Cancers*. 24 de julio de 2018;10(8):242.
  26. Chen J, Yang L, Ma Y, Zhang Y. Recent advances in understanding the immune microenvironment in ovarian cancer. *Front Immunol*. 5 de junio de 2024;15:1412328.
  27. Lu H, Lou H, Wengert G, Paudel R, Patel N, Desai S, et al. Tumor and local lymphoid tissue interaction determines prognosis in high-grade serous ovarian cancer. *Cell Reports Medicine*. julio de 2023;4(7):101092.
  28. Yang L, Wang S, Zhang Q, Pan Y, Lv Y, Chen X, et al. Clinical significance of the immune microenvironment in ovarian cancer patients. *Mol Omics*. 2018;14(5):341-51.
  29. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med*. 22 de septiembre de 2016;375(12):1109-12.
  30. de Bruijn I, Kundra R, Mastrogiacomo B, Tran TN, Sikina L, Mazor T, et al. Analysis and Visualization of Longitudinal Genomic and Clinical Data from the AACR Project GENIE Biopharma Collaborative in cBioPortal. *Cancer Res*. 5 de septiembre de 2023;
  31. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. mayo de 2012;2(5):401-4.

32. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*. 2 de abril de 2013;6(269):p11.
33. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 1 de enero de 2002;30(1):207-10.
34. Beg A, Parveen R. Review of Bioinformatics Tools and Techniques to Accelerate Ovarian Cancer Research. *International Journal of Bioinformatics and Intelligent Computing*. 11 de febrero de 2022;1(1):01-10.
35. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. enero de 2009;4(1):44-57.
36. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res*. 6 de enero de 2023;51(D1):D587-92.
37. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. enero de 2009;37(1):1-13.
38. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Res*. 7 de enero de 2022;50(D1):D687-92.
39. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. mayo de 2000;25(1):25-9.
40. Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The Gene Ontology knowledgebase in 2023. *Genetics*. 4 de mayo de 2023;224(1):iyad031.
41. Huang K, Xiao C, Glass LM, Critchlow CW, Gibson G, Sun J. Machine learning applications for therapeutic tasks with genomics data. *Patterns*. 8 de octubre de 2021;2(10):100328.
42. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun*. 6 de noviembre de 2020;11(1):5650.
43. Nguyen H, Nguyen H, Tran D, Draghici S, Nguyen T. Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges. *Nucleic Acids Research*. 22 de mayo de 2024;52(9):4761-83.

44. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*. 15 de julio de 2019;35(14):i436-45.
45. Hippen AA, Omran DK, Weber LM, Jung E, Drapkin R, Doherty JA, et al. Performance of computational algorithms to deconvolve heterogeneous bulk ovarian tumor tissue depends on experimental factors. *Genome Biology*. 20 de octubre de 2023;24(1):239.
46. Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*. 1 de junio de 2018;34(11):1969-79.
47. Huuki-Myers LA, Montgomery KD, Kwon SH, Cinquemani S, Eagles NJ, Gonzalez-Padilla D, et al. Benchmark of cellular deconvolution methods using a multi-assay reference dataset from postmortem human prefrontal cortex. *bioRxiv*. 7 de abril de 2024;2024.02.09.579665.
48. Li Z, Wu H. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biol*. diciembre de 2019;20(1):190.
49. Li Z, Guo Z, Cheng Y, Jin P, Wu H. Robust partial reference-free cell composition estimation from tissue expression. Luigi Martelli P, editor. *Bioinformatics*. 1 de junio de 2020;36(11):3431-8.
50. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun*. 22 de enero de 2019;10(1):380.
51. Zaitsev K, Bambouskova M, Swain A, Artyomov MN. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat Commun*. 17 de mayo de 2019;10(1):2209.
52. Kang K, Huang C, Li Y, Umbach DM, Li L. CDSeqR: fast complete deconvolution for gene expression data from bulk tissues. *BMC Bioinformatics*. diciembre de 2021;22(1):262.
53. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol*. julio de 2019;37(7):773-82.
54. Maden SK, Kwon SH, Huuki-Myers LA, Collado-Torres L, Hicks SC, Maynard KR. Challenges and opportunities to computationally deconvolve heterogeneous tissue with varying cell sizes using single-cell RNA-sequencing datasets. *Genome Biol*. 14 de diciembre de 2023;24(1):288.

55. Dietrich A, Sturm G, Merotto L, Marini F, Finotello F, List M. *SimBu*: bias-aware simulation of bulk RNA-seq data with variable cell-type composition. *Bioinformatics*. 16 de septiembre de 2022;38(Supplement\_2):ii141-7.
56. Rovetta A. Raiders of the Lost Correlation: A Guide on Using Pearson and Spearman Coefficients to Detect Hidden Correlations in Medical Sciences. *Cureus*. 30 de noviembre de 2020;12(11):e11794.
57. Sompairac N. Unsupervised hierarchical deconvolution of gene expression data to unravel the tumor micro-environment complexity [Internet]. Université Paris; Disponible en: <https://theses.hal.science/tel-04523762v1>
58. Parikh AS, Li Y, Mazul A, Yu VX, Thorstad W, Rich J, et al. Immune Cell Deconvolution Reveals Possible Association of  $\gamma\delta$  T Cells with Poor Survival in Head and Neck Squamous Cell Carcinoma. *Cancers (Basel)*. 5 de octubre de 2023;15(19):4855.
59. Shao Y, Chen C, Yu X, Yan J, Guo J, Ye G. Comprehensive analysis of scRNA-seq and bulk RNA-seq data via machine learning and bioinformatics reveals the role of lysine metabolism-related genes in gastric carcinogenesis. *BMC Cancer*. 9 de abril de 2025;25(1):644.
60. Bhinder B, Friedl V, Sethuraman S, Risso D, Chiotti KE, Mashl RJ, et al. Pan-cancer immune and stromal deconvolution predicts clinical outcomes and mutation profiles. *Sci Rep*. 4 de julio de 2025;15(1):23921.
61. Zaitsev A, Chelushkin M, Dyikanov D, Cheremushkin I, Shpak B, Nomie K, et al. Precise reconstruction of the TME using bulk RNA-seq and a machine learning algorithm trained on artificial transcriptomes. *Cancer Cell*. 8 de agosto de 2022;40(8):879-894.e16.
62. Computational de novo discovery of distinguishing genes for biological processes and cell types in complex tissues | *PLOS One* [Internet]. [citado 13 de septiembre de 2025]. Disponible en: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0193067>
63. CAM3.0: determining cell type composition and expression from bulk tissues with fully unsupervised deconvolution | *Bioinformatics* | Oxford Academic [Internet]. [citado 13 de septiembre de 2025]. Disponible en: <https://academic.oup.com/bioinformatics/article/40/3/btae107/7614270>
64. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 5 de mayo de 2016;44(8):e71.
65. Chen Y, Chen L, Lun ATL, Baldoni PL, Smyth GK. edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Research*. 11 de enero de 2025;53(2):gkaf018.

66. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart – biological queries made easy. *BMC Genomics*. 14 de enero de 2009;10(1):22.
67. Wang Y, Duval AJ, Adli M, Matei D. Biology-driven therapy advances in high-grade serous ovarian cancer. *Journal of Clinical Investigation*. 2 de enero de 2024;134(1):e174013.
68. Li S, Jiang B, Zhou H, Yang S, Yang L, Hong Y. Development of a prognostic immune cell-based model for ovarian cancer using multiplex immunofluorescence. *J Transl Med*. 19 de junio de 2025;23(1):688.