

# ***digitalDLSorteR: paquete de R para la deconvolución de muestras bulk RNA-seq basado en Redes Neuronales.***

Trabajo Fin de Máster

Máster en Bioinformática y Biología Computacional

---

Diego Mañanes Cayero

Universidad Autónoma de Madrid  
Curso 2019-2020



Universidad Autónoma  
de Madrid



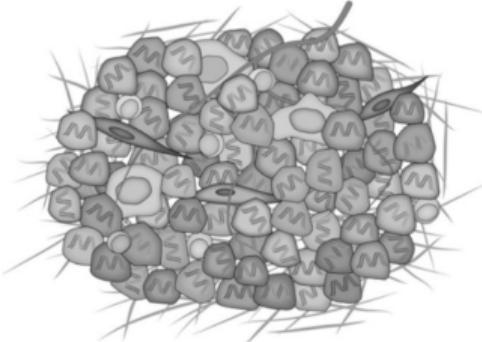
# Contenidos

---

1. Introducción y contexto
2. Objetivos
3. digitalDLSorteR: transformación de la *pipeline* en paquete de R
4. Análisis de datos *scRNA-seq* de cáncer de mama
5. Puesta en práctica de digitalDLSorteR: deconvolución muestras de cáncer de mama
6. Conclusiones y trabajo futuro

## **Introducción y contexto**

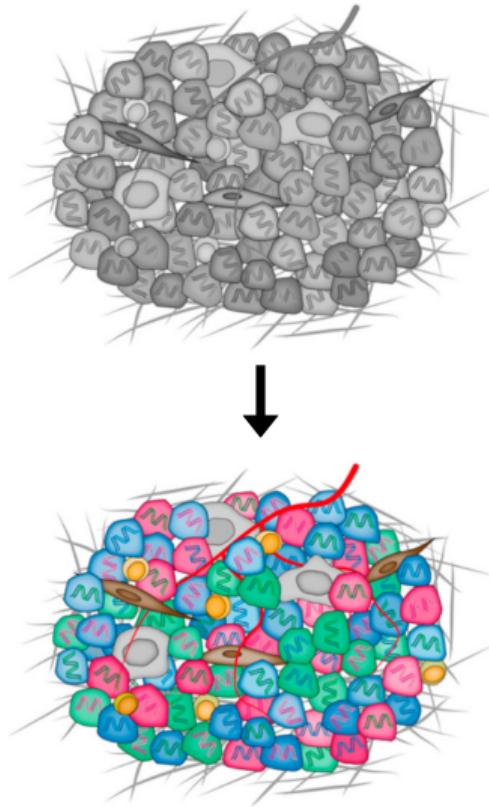
---



## Por qué

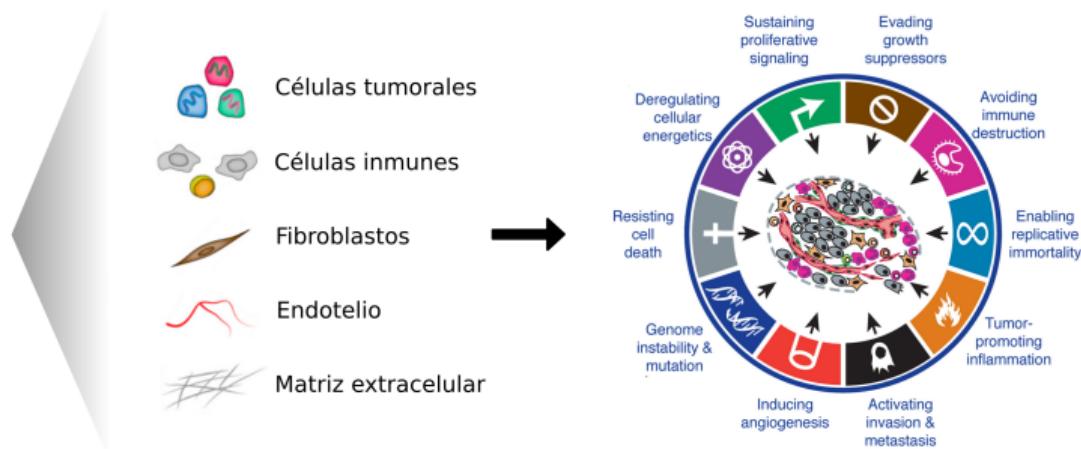
- Diferentes poblaciones tumorales.
- Micro-entorno tumoral: diferentes tipos celulares con complejas interacciones.
- Sello de Identidad del cáncer.

# Cáncer y micro-entorno tumoral

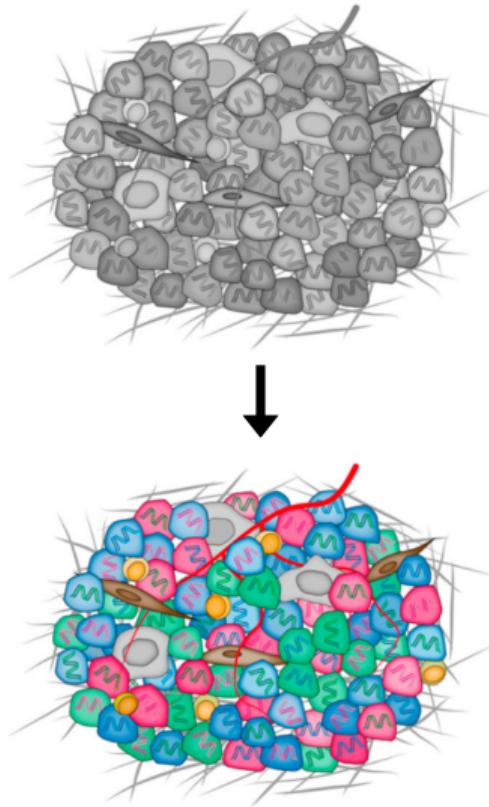


## Por qué

- Diferentes poblaciones tumorales.
- Micro-entorno tumoral: diferentes tipos celulares con complejas interacciones.
- Sello de Identidad del cáncer.

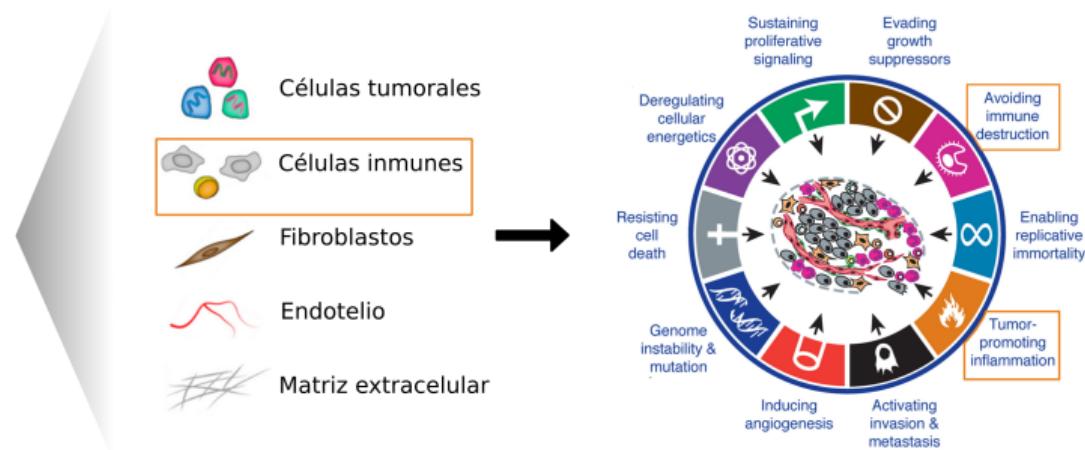


# Cáncer y micro-entorno tumoral



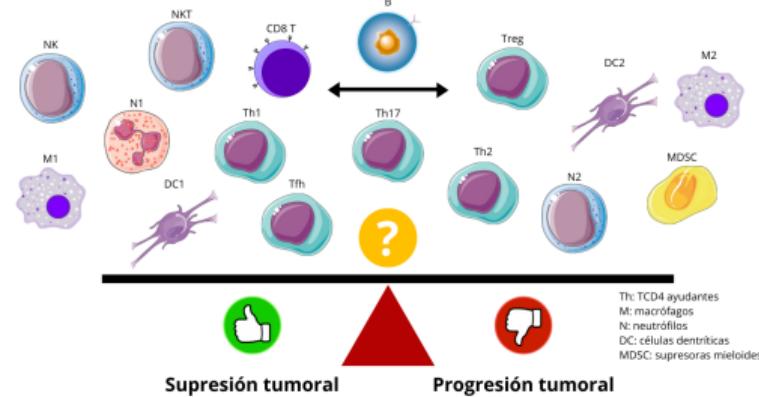
## Por qué

- Diferentes poblaciones tumorales.
- Micro-entorno tumoral: diferentes tipos celulares con complejas interacciones.
- Sello de Identidad del cáncer.



# Papel del sistema inmune en la enfermedad

- En tumores sólidos establece lo que se conoce como contexto inmune: localización, densidad y organización funcional.
- Sistema inmune presenta un papel dicotómico:
  - Supresión tumoral: linfocitos T CD8+, células B de memoria, etc.
  - Progresión tumoral: linfocitos T reguladores, algunos macrófagos, etc.
- **Terapias:** Quimioterapias inmunogénicas como oxaliplatino, anticuerpos inhibidores de puntos de control inmunitario: PD-1, CTLA-4, receptores coestimuladores de la respuesta inmumne, etc.



# Caso de estudio: cáncer de mama

Enfermedad altamente **heterogénea desde el punto de vista molecular**.

## Subtipos intrínsecos del cáncer

- Luminal A (ER+).
- Luminal B (ER+/HER2+).
- HER2 enriquecido (HER2+).
- Triple negativo (TNBC).

## Importancia del sistema inmune

- ↑ linfocitos T CD8+ → ↑ supervivencia.
- ↑ linfocitos T CD4+ reguladores → ↑ efecto inmunosupresor.
- ↑ TILs → beneficioso en pacientes tratados con trastuzumab.
- Quimioterapias neoadyuvantes.

Subtipo cáncer	Luminal A	Luminal B	HER2 enriquecido	Basal o triple negativo (TNBC)
% de cánceres de mama	50%	25%	15%	10%
Fenotipo	ER+ PR+ HER2-	ER+ PR+ HER2+	ER- PR- HER2+	ER- PR- HER2-
Prognosis	Buena		Mala	
Valor de TILs en la prognosis				
Tratamiento	Terapia endocrina Anti-HER2 mAb Quimioterapia			

# Estudio de la heterogeneidad celular

Necesidad de métodos para el estudio de la heterogeneidad celular.

## Tradicionalmente

A nivel de proteína mediante técnicas inmunohistoquímicas, inmunofluorescencia y citometría de flujo.

Pequeña combinación de marcadores génicos

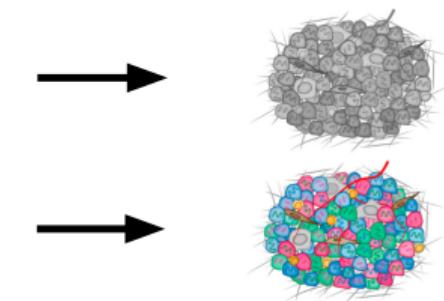
## Tecnologías de alto rendimiento

A nivel transcriptómico mediante NGS: *RNA-seq*.

Estatus funcional completo

Dos variantes:

- *Bulk RNA-seq* (nivel tisular): los niveles de expresión corresponden al sumatorio de tipos celulares presentes en las muestras.
- *Single-cell RNA-seq* (nivel celular): los niveles de expresión corresponden a cada célula individual que compone la muestra.



# *scRNA-seq*: Ventajas e inconvenientes

## Ventajas

- Estatus funcional de cada célula.
- Caracterización de las poblaciones celulares. Potencial identificación de tipos celulares no esperados.
- Potencial aplicación traslacional: mejora del diagnóstico, terapias dirigidas, etc.

## Inconvenientes

- Altos costes económicos y protocolos complejos.
- No es práctica su aplicación en grandes cohortes de muestras.
- Baja eficiencia de captura de ARN.
- Mayor ruido técnico.

# *scRNA-seq*: Ventajas e inconvenientes

## Ventajas

- Estatus funcional de cada célula.
- Caracterización de las poblaciones celulares. Potencial identificación de tipos celulares no esperados.
- Potencial aplicación translacional: mejora del diagnóstico, terapias dirigidas, etc.

## Inconvenientes

- Altos costes económicos y protocolos complejos.
- No es práctica su aplicación en grandes cohortes de muestras.
- Baja eficiencia de captura de ARN.
- Mayor ruido técnico.

**Resultado:**

*Bulk RNA-seq* sigue siendo el estándar.

# *scRNA-seq*: Ventajas e inconvenientes

## Ventajas

- Estatus funcional de cada célula.
- Caracterización de las poblaciones celulares. Potencial identificación de tipos celulares no esperados.
- Potencial aplicación translacional: mejora del diagnóstico, terapias dirigidas, etc.

## Inconvenientes

- Altos costes económicos y protocolos complejos.
- No es práctica su aplicación en grandes cohortes de muestras.
- Baja eficiencia de captura de ARN.
- Mayor ruido técnico.

### Resultado:

*Bulk RNA-seq* sigue siendo el estándar.

**Problema:** No tiene en cuenta en qué proporción contribuye cada tipo celular a los niveles de expresión medidos.

# scRNA-seq: Ventajas e inconvenientes

## Ventajas

- Estatus funcional de cada célula.
- Caracterización de las poblaciones celulares. Potencial identificación de tipos celulares no esperados.
- Potencial aplicación translacional: mejora del diagnóstico, terapias dirigidas, etc.

## Inconvenientes

- Altos costes económicos y protocolos complejos.
- No es práctica su aplicación en grandes cohortes de muestras.
- Baja eficiencia de captura de ARN.
- Mayor ruido técnico.

**Resultado:**

*Bulk RNA-seq* sigue siendo el estándar.

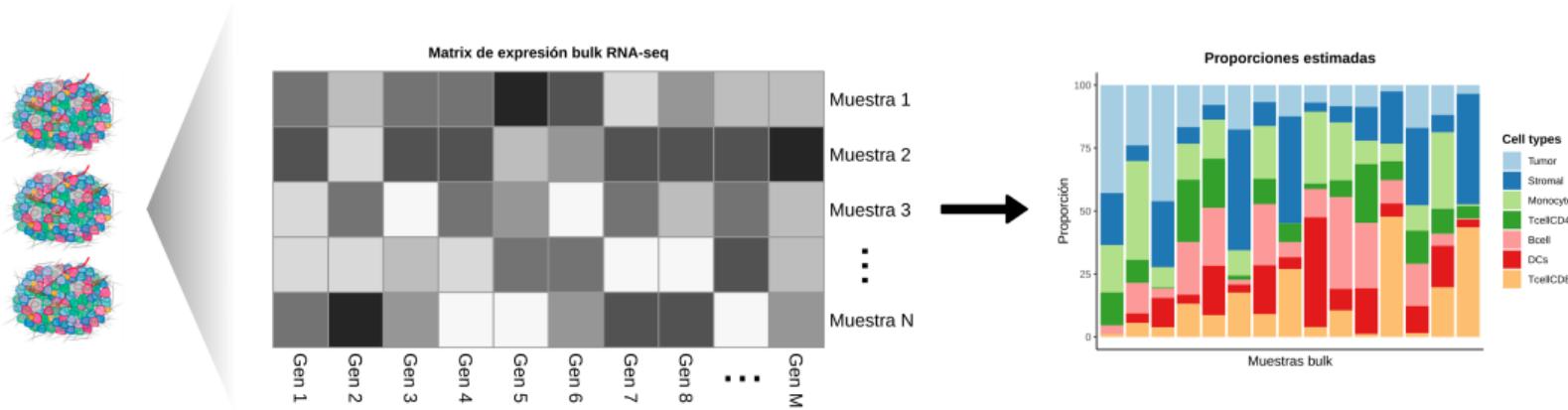
**Problema:** No tiene en cuenta en qué proporción contribuye cada tipo celular a los niveles de expresión medidos.

**Necesidad:** Métodos computacionales que permitan estimar las proporciones de cada tipo celular medidas en muestras *bulk RNA-seq*.

# Deconvolución de muestras *bulk RNA-seq*

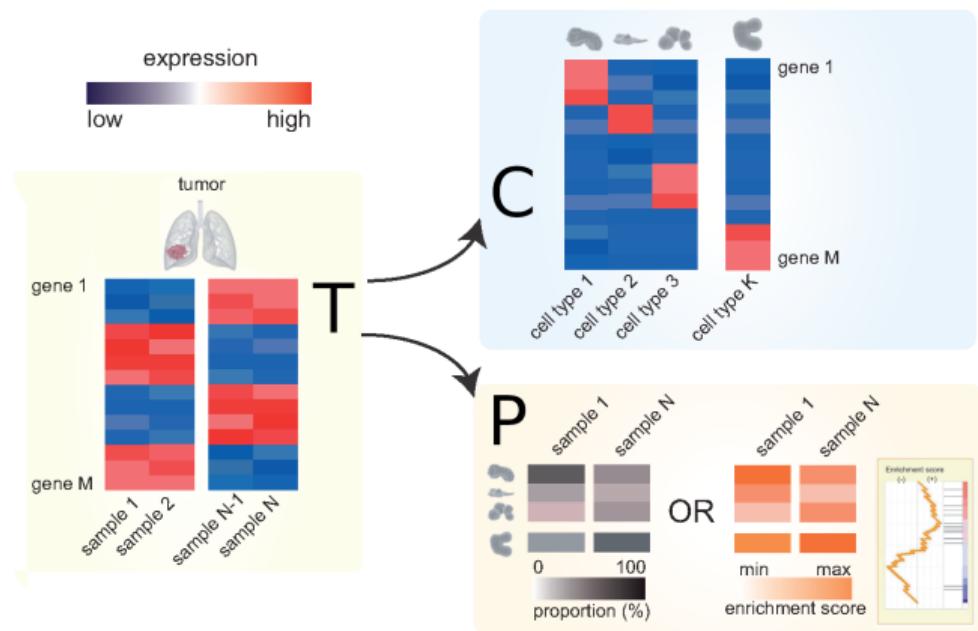
**Deconvolución:** estimación de la señal individual de cada uno de los componentes (tipos celulares) a partir de una mezcla de los mismos (muestra tisular).

- Sustituto de experimentos *scRNA-seq* por sus altos costes económicos o la imposibilidad de su aplicación.
- Control de la contribución de cada tipo celular a las muestras tisulares → evitar factores de confusión en análisis de expresión diferencial.



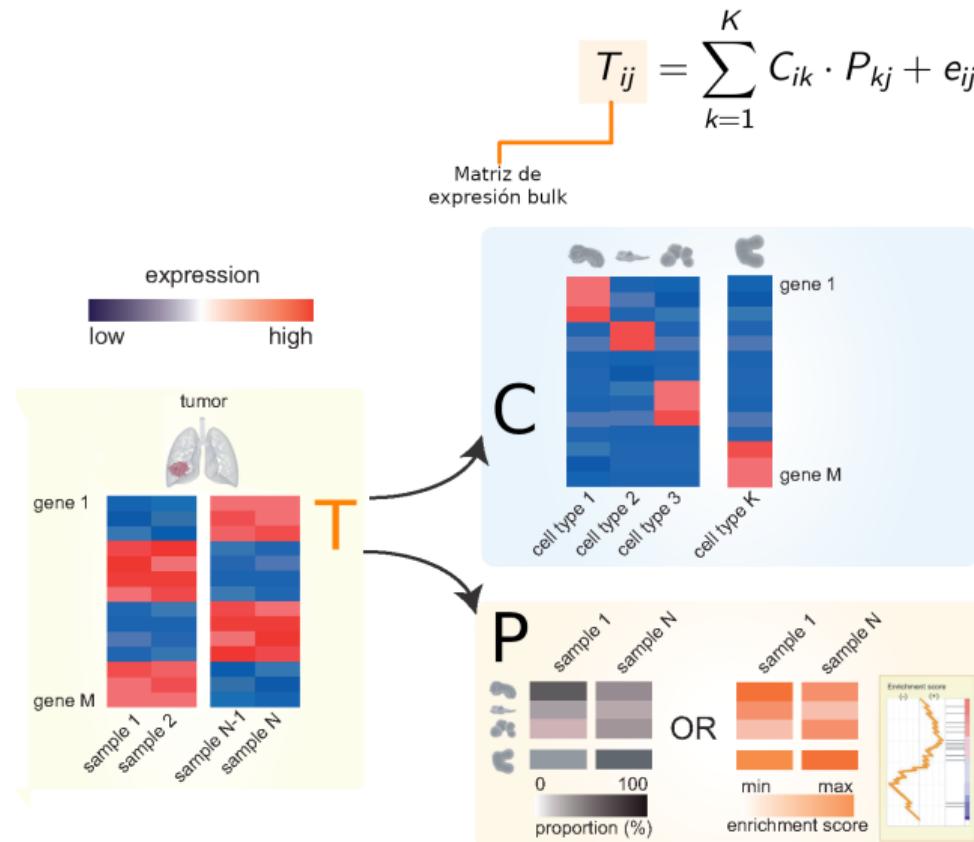
# Deconvolución: marco de trabajo

$$T_{ij} = \sum_{k=1}^K C_{ik} \cdot P_{kj} + e_{ij}$$



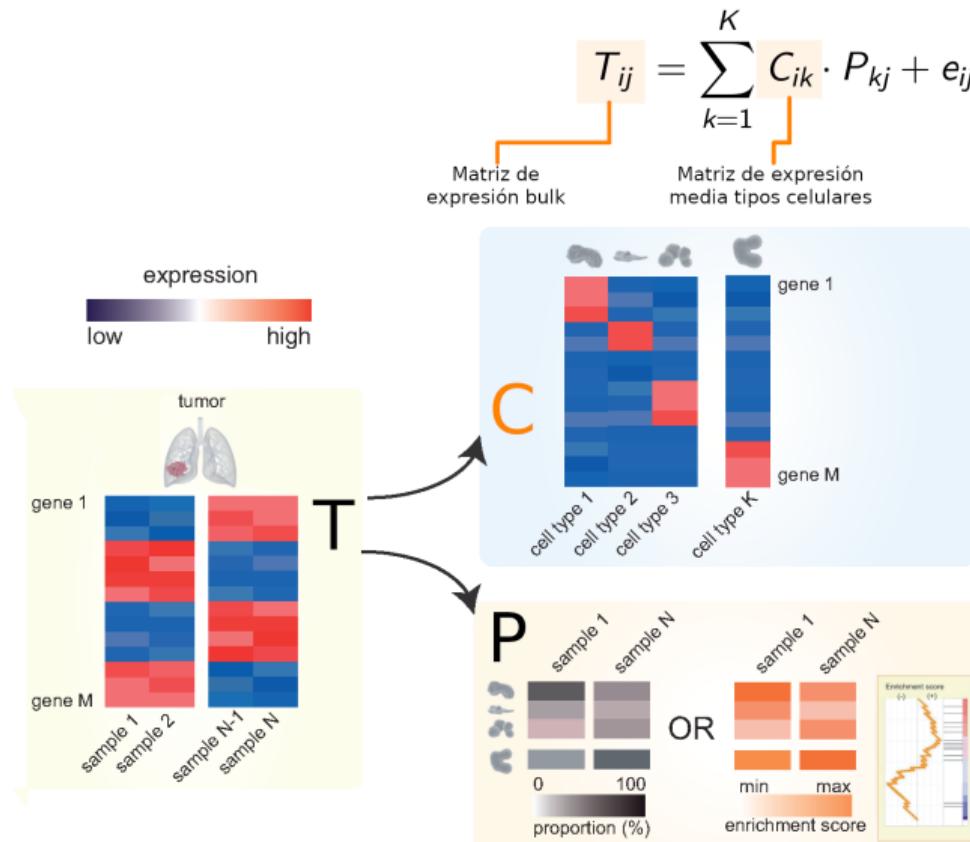
- $i$ : genes ( $i = 1 \dots M$ ).
- $j$ : muestras *bulk* ( $j = 1 \dots N$ ).
- $k$ : tipos celulares ( $k = 1 \dots K$ ).

# Deconvolución: marco de trabajo

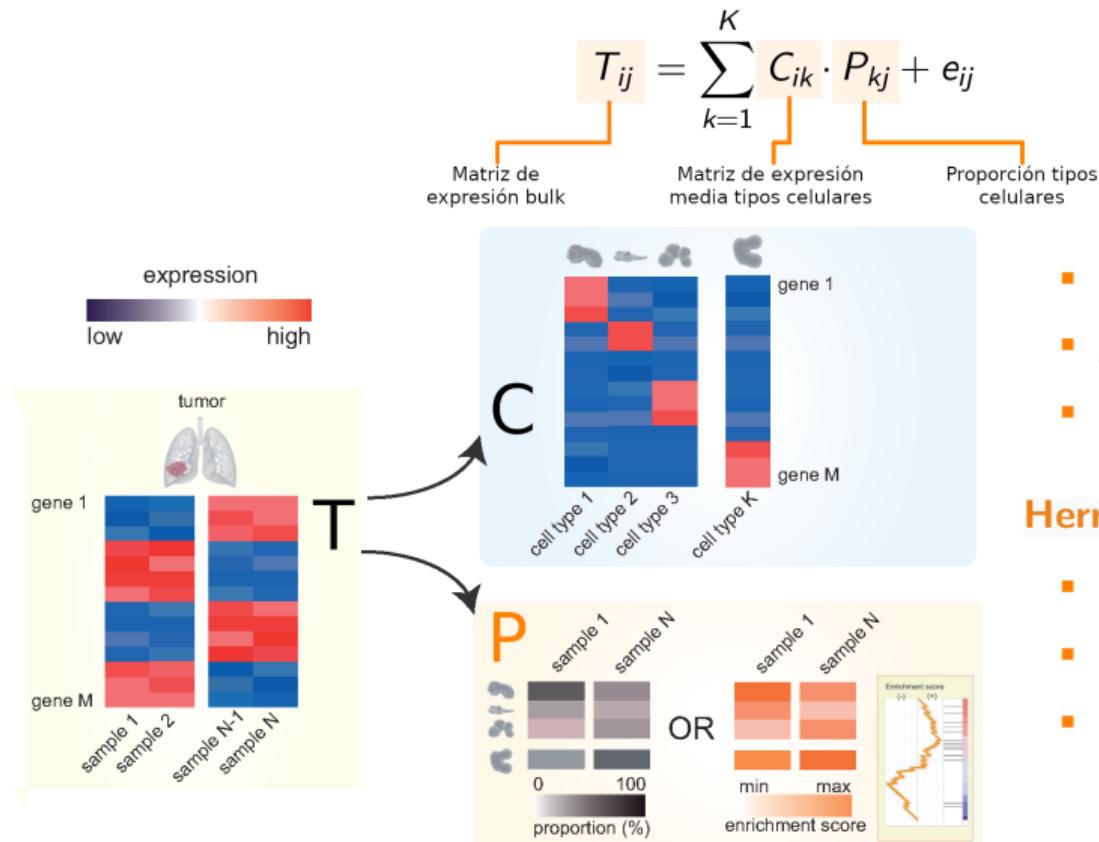


- $i$ : genes ( $i = 1 \dots M$ ).
- $j$ : muestras *bulk* ( $j = 1 \dots N$ ).
- $k$ : tipos celulares ( $k = 1 \dots K$ ).

# Deconvolución: marco de trabajo



# Deconvolución: marco de trabajo



- $i$ : genes ( $i = 1 \dots M$ ).
- $j$ : muestras *bulk* ( $j = 1 \dots N$ ).
- $k$ : tipos celulares ( $k = 1 \dots K$ ).

## Herramientas publicadas

- Enriquecimiento de sets de genes.
- Mínimos cuadrados ordinarios.
- $v$ -SVR

# Método de deconvolución digitalDLSorter

## Características

- Método basado en **Aprendizaje Profundo** → revolución en el campo del Aprendizaje Automático durante los últimos años por su desempeño.
- Uso de datos **scRNA-seq** → perfiles de expresión procedentes del propio entorno de estudio (micro-entorno tumoral cáncer de mama, cáncer de colon, entorno neuronal, etc.).

## Implementación

- *Pipeline* escrita en varios lenguajes de programación.
- Cada paso escribe los datos intermedios en disco como ficheros tabulados.
- No ofrece la opción de utilizar modelos preentrenados.
- Uso complicado por otros usuarios.

## **Objetivos**

---

1. Transformación de la *pipeline* original en un paquete de R.
  - Unificación de los lenguajes de programación → R.
  - Evitar la lectura/escritura de ficheros tabulados en disco.
  - Funcionalidades nuevas → uso de ficheros HDF5 como *back-end*, implementación de parámetros para construir modelos más personalizados.
  - Generación de documentación y viñeta → facilitar su uso.
2. Análisis de datos *scRNA-seq* procedentes de cáncer de mama.
  - Separación tipos tumorales y no tumorales.
  - Identificación de tipos celulares no tumorales.
3. Aplicación de la herramienta implementada.
  - Construcción de varios modelos y comparativa.
  - Incorporación de los mejores como modelos preentrenados en el paquete.

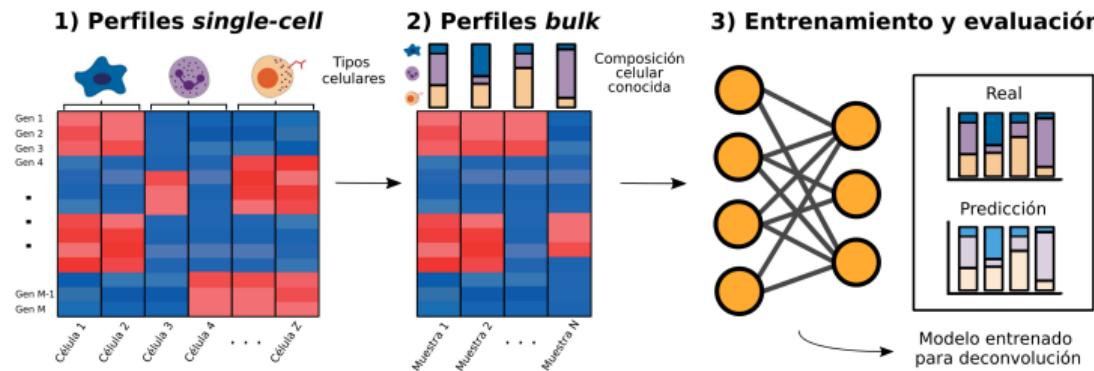
1. Transformación de la *pipeline* original en un paquete de R.
  - Unificación de los lenguajes de programación → R.
  - Evitar la lectura/escritura de ficheros tabulados en disco.
  - Funcionalidades nuevas → uso de ficheros HDF5 como *back-end*, implementación de parámetros para construir modelos más personalizados.
  - Generación de documentación y viñeta → facilitar su uso.
2. Análisis de datos *scRNA-seq* procedentes de cáncer de mama.
  - Separación tipos tumorales y no tumorales.
  - Identificación de tipos celulares no tumorales.
3. Aplicación de la herramienta implementada.
  - Construcción de varios modelos y comparativa.
  - Incorporación de los mejores como modelos preentrenados en el paquete.

1. Transformación de la *pipeline* original en un paquete de R.
  - Unificación de los lenguajes de programación → R.
  - Evitar la lectura/escritura de ficheros tabulados en disco.
  - Funcionalidades nuevas → uso de ficheros HDF5 como *back-end*, implementación de parámetros para construir modelos más personalizados.
  - Generación de documentación y viñeta → facilitar su uso.
2. Análisis de datos *scRNA-seq* procedentes de cáncer de mama.
  - Separación tipos tumorales y no tumorales.
  - Identificación de tipos celulares no tumorales.
3. Aplicación de la herramienta implementada.
  - Construcción de varios modelos y comparativa.
  - Incorporación de los mejores como modelos preentrenados en el paquete.

## **digitalDLSorteR: transformación de la *pipeline* en paquete de R**

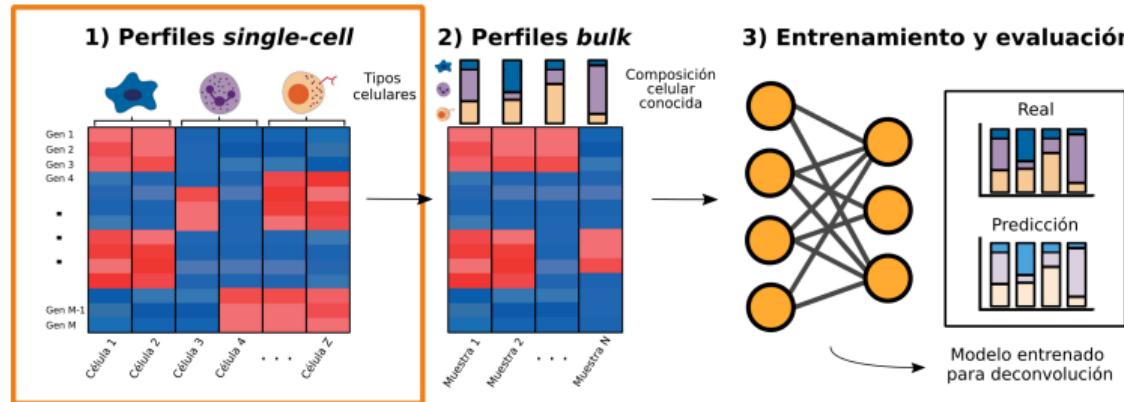
---

# Fundamento del método



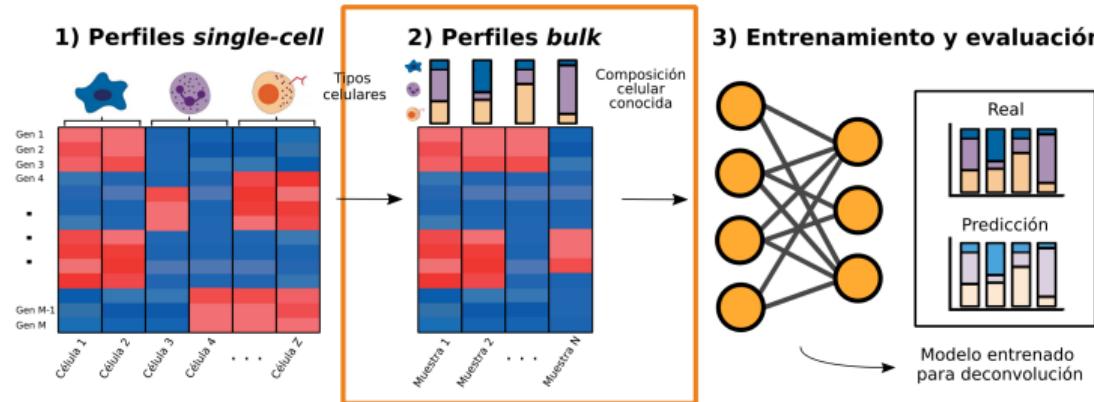
1. Entrada de perfiles *single-cell* caracterizados. Simulación de nuevos perfiles con el modelo ZINB-WaVE en caso de tipos celulares poco representados o pocas células de partida.
2. Simulación de perfiles *bulk RNA-seq* de composición celular conocida: agregación de perfiles *single-cell* en función del tipo celular al que pertenecen.
3. Entrenamiento y evaluación de la Red Neuronal Profunda: deconvolución de nuevas muestras *bulk*.

# Fundamento del método



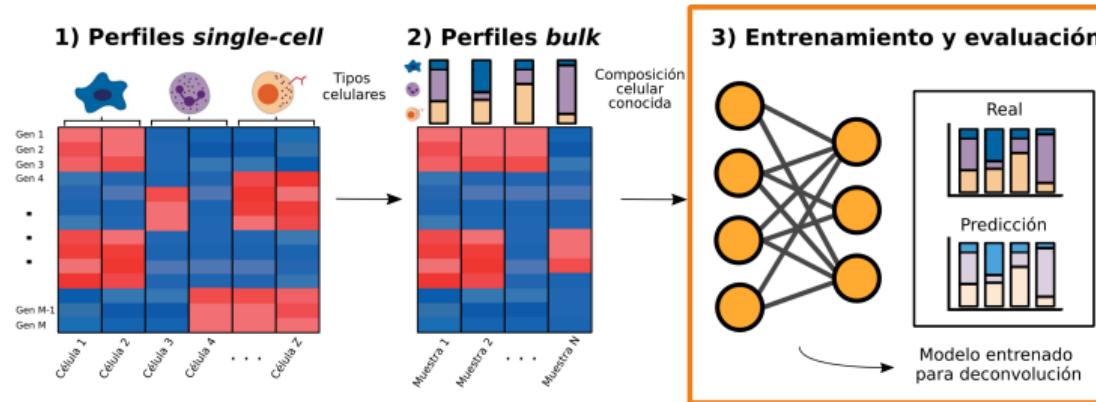
1. Entrada de perfiles *single-cell* caracterizados. Simulación de nuevos perfiles con el modelo ZINB-WaVE en caso de tipos celulares poco representados o pocas células de partida.
2. Simulación de perfiles *bulk RNA-seq* de composición celular conocida: agregación de perfiles *single-cell* en función del tipo celular al que pertenecen.
3. Entrenamiento y evaluación de la Red Neuronal Profunda: deconvolución de nuevas muestras *bulk*.

# Fundamento del método



1. Entrada de perfiles *single-cell* caracterizados. Simulación de nuevos perfiles con el modelo ZINB-WaVE en caso de tipos celulares poco representados o pocas células de partida.
2. Simulación de perfiles *bulk RNA-seq* de composición celular conocida: agregación de perfiles *single-cell* en función del tipo celular al que pertenecen.
3. Entrenamiento y evaluación de la Red Neuronal Profunda: deconvolución de nuevas muestras *bulk*.

# Fundamento del método



1. Entrada de perfiles *single-cell* caracterizados. Simulación de nuevos perfiles con el modelo ZINB-WaVE en caso de tipos celulares poco representados o pocas células de partida.
2. Simulación de perfiles *bulk RNA-seq* de composición celular conocida: agregación de perfiles *single-cell* en función del tipo celular al que pertenecen.
3. **Entrenamiento y evaluación de la Red Neuronal Profunda:** deconvolución de nuevas muestras *bulk*.

## Fundamento del método

$$T_{ij} = \sum_{k=1}^K C_{ik} \cdot P_{kj} + e_{ij}$$

Matriz de expresión bulk      Matriz de expresión media tipos celulares      Proporción tipos celulares

1. Entrada de perfiles *single-cell* caracterizados. Simulación de nuevos perfiles con el modelo ZINB-WaVE en caso de tipos celulares poco representados o pocas células de partida.
2. Simulación de perfiles *bulk RNA-seq* de composición celular conocida: agregación de perfiles *single-cell* en función del tipo celular al que pertenecen.
3. Entrenamiento y evaluación de la Red Neuronal Profunda: deconvolución de nuevas muestras *bulk*.

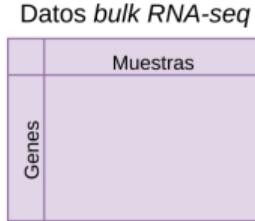
## Objetivo del método como paquete

1. Permitir la deconvolución directa de muestras *bulk RNA-seq*.
2. Permitir la construcción de nuevos modelos a partir de datos *scRNA-seq*.

digitalDLSorter ofrece  
dos flujos de trabajo

- 
1. Uso de modelos preentrenados integrados en el paquete
  2. Construcción de nuevos modelos a partir de *scRNA-seq*

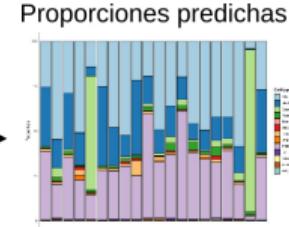
# Uso de modelos preentrenados



-deconvDigitalDLSorter→



-barPlotCellTypes→



```
1 deconvResults <- deconvDigitalDLSorter(  
2   data = TCGA.breast.small,  
3   model = "breast.chung.generic",  
4   normalize = TRUE  
5 )  
6 ## barplot showing results  
7 barPlotCellTypes(deconvResults)
```

- Modelos para la cuantificación de células inmunes en cáncer de mama.
  1. Modelo genérico: 7 tipos celulares.
  2. Modelo específico: 13 tipos celulares.
- Datos procedentes de Chung et al., 2017 ([GSE75688](#)).
- Agregación de tipos celulares con `simplify.set` y `simplify.majority`.

# Construcción de nuevos modelos

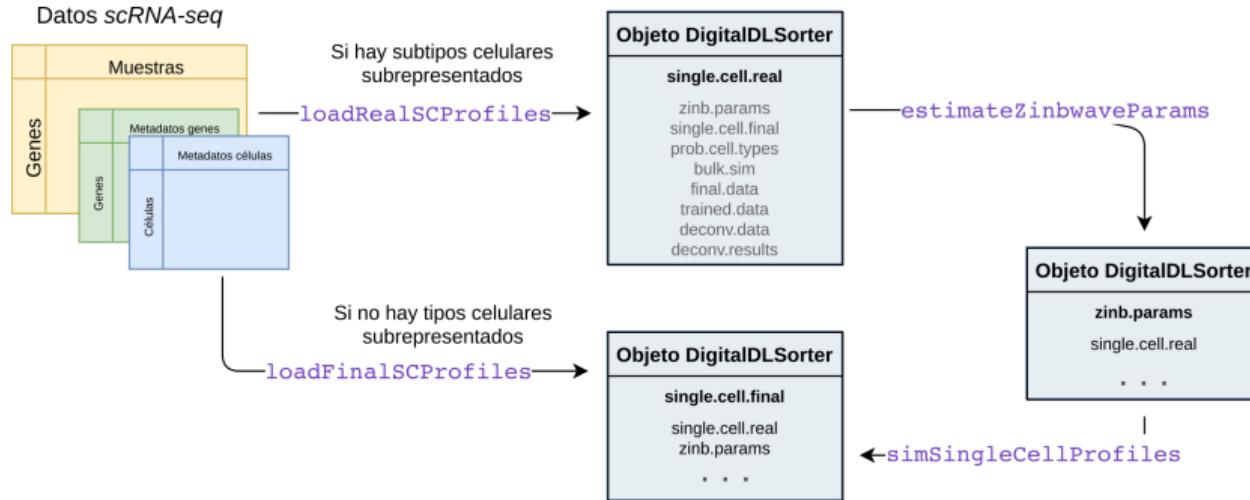
## Clases

- *DigitalDLSorter*: núcleo del paquete.
- *ProbMatrixCellTypes*: información relativa a las matrices de composición celular.
- *DigitalDLSorterDNN*: información relativa a la Red Neuronal Profunda.
- Otras clases: *SingleCellExperiment*, *SummarizedExperiment*, etc.

## Flujo de trabajo

1. Carga de datos y simulación de nuevos perfiles *single-cell* (si es necesario).
2. Generación de la matriz de composición celular.
3. Simulación de perfiles *bulk RNA-seq* de acuerdo a las proporciones establecidas en el paso anterior. Preparación de los datos para el entrenamiento.
4. Entrenamiento de la red neuronal y evaluación.
5. Carga de nuevos datos *bulk* y su deconvolución.

# 1. Carga de datos y simulación de perfiles *single-cell*



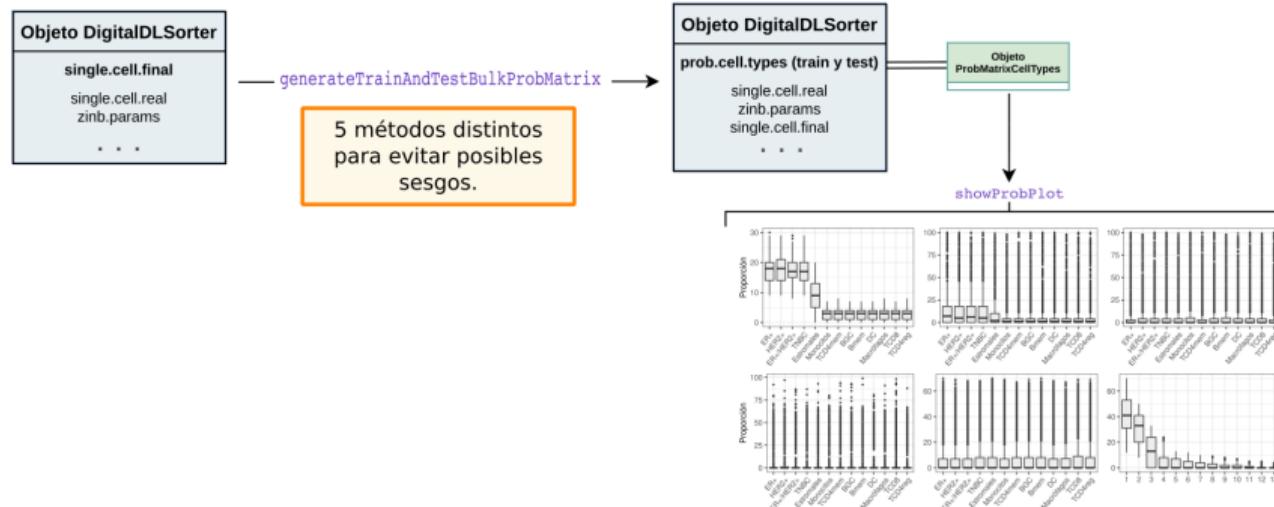
- Carga de los datos en un objeto de la clase *DigitalDLSorter* en el slot `single.cell.real` o `single.cell.final` → matriz de expresión, metadatos de células y metadatos de genes.
- Si es necesario: estimación de parámetros mediante el modelo ZINB-WaVE y simulación de nuevos perfiles *single-cell*: distribución binomial negativa cero inflada.

# 1. Carga de datos y simulación de perfiles *single-cell*

```
1 DDLSChungSmall <- loadRealSCPProfiles(  
2   ## SingleCellExperiment object  
3   single.cell.real = sc.chung.breast,  
4   cell.ID.column = "Cell_ID",  
5   gene.ID.column = "external_gene_name",  
6   min.cells = 1,  
7   min.counts = 1,  
8   project = "Chung_example"  
9 )
```

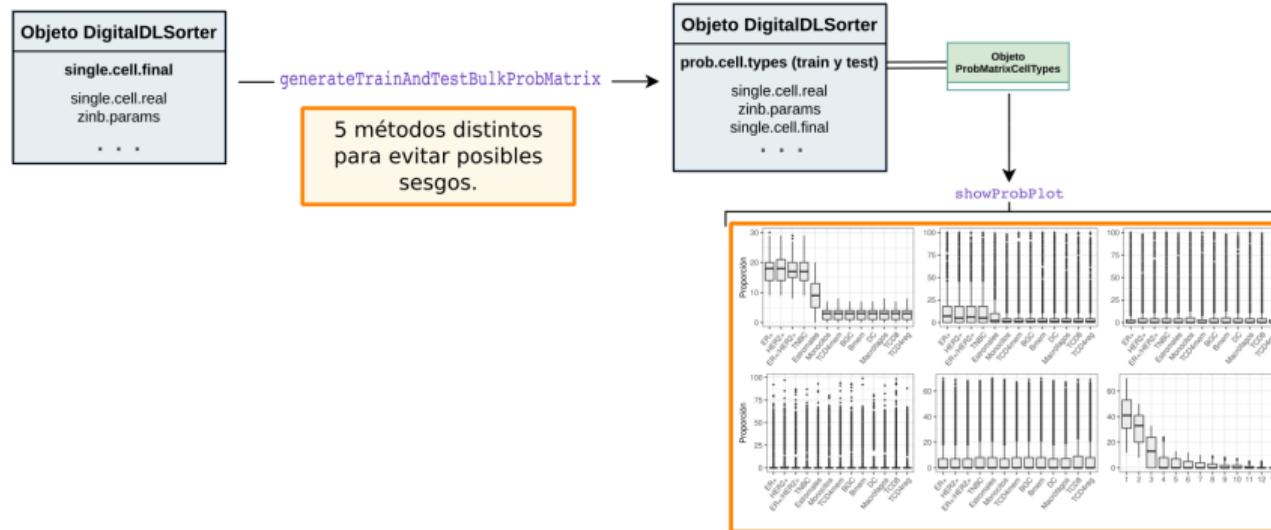
```
1 ## estimation of ZINB-Wave parameters  
2 DDLSChungSmall <- estimateZinbwaveParams(  
3   object = DDLSChungSmall,  
4   cell.ID.column = "Cell_ID",  
5   gene.ID.column = "external_gene_name",  
6   cell.type.column = "Cell_type",  
7   cell.cov.columns = "Patient",  
8   gene.cov.columns = "gene_length"  
9 )  
10 ## simulation of new profiles  
11 DDLSChungSmall <- simSingleCellProfiles(  
12   object = DDLSChungSmall,  
13   cell.ID.column = "Cell_ID",  
14   cell.type.column = "Cell_type",  
15   n.cells = 10 # 1000 in real situations  
16 )
```

## 2. Generación de la matriz de composición celular



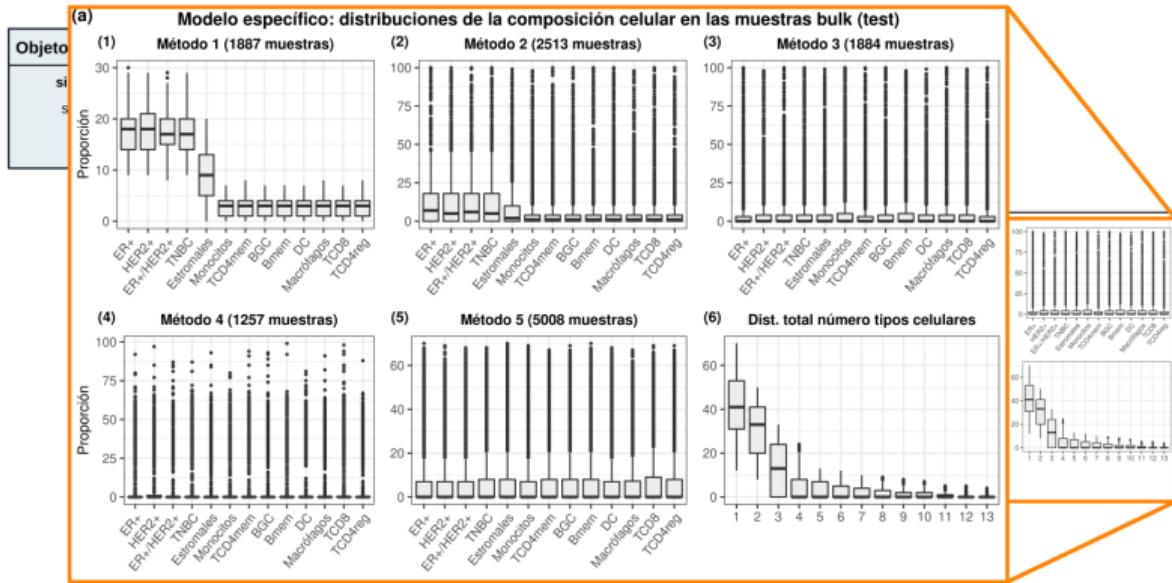
- Generación de un objeto `ProbMatrixCellTypes` en el slot `prob.cell.types`.
- 5 métodos distintos para evitar posibles sesgos en la composición de las muestras *bulk*.
- Perfiles *single-cell* son separados en entrenamiento y test.

## 2. Generación de la matriz de composición celular



- Generación de un objeto *ProbMatrixCellTypes* en el slot `prob.cell.types`.
- 5 métodos distintos para evitar posibles sesgos en la composición de las muestras *bulk*.
- Perfiles *single-cell* son separados en entrenamiento y test.

## 2. Generación de la matriz de composición celular

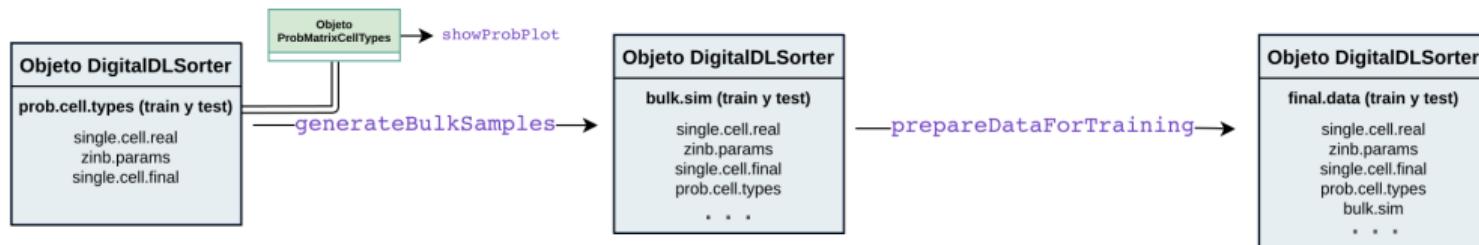


- Generación de un objeto *ProbMatrixCellTypes* en el slot *prob.cell.types*.
- 5 métodos distintos para evitar posibles sesgos en la composición de las muestras *bulk*.
- Perfiles *single-cell* son separados en entrenamiento y test.

## 2. Generación de la matriz de composición celular

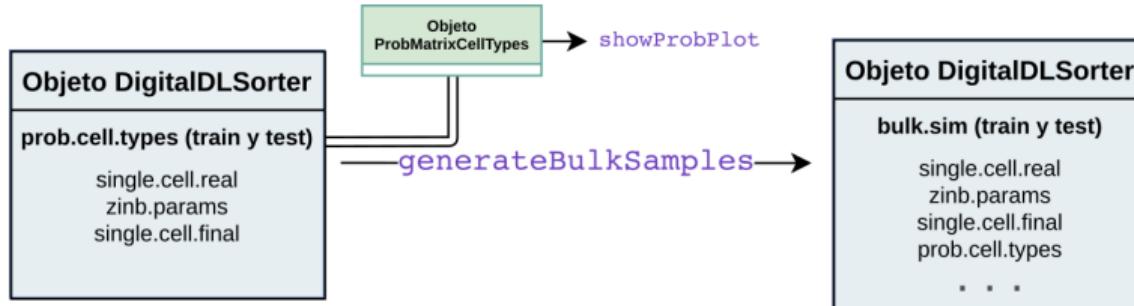
```
1 ## data.frame with prior information about cell types
2 probMatrix <- data.frame(
3   Cell_types = c("ER+", "HER2+", "ER+ and HER2+", "TNBC",
4                 "Stromal", "Monocyte", "TCD4me", "BGC",
5                 "Bmem", "DC", "Macrophage", "TCD8", "TCD4reg"),
6   from = c(rep(30, 4), 1, rep(0, 8)),
7   to = c(rep(70, 4), 50, rep(15, 8))
8 )
9 DDLSChungSmall <- generateTrainAndTestBulkProbMatrix(
10   object = DDLSChungSmall,
11   cell.type.column = "Cell_type",
12   prob.design = probMatrix,
13   proportions.train = c(10, 5, 20, 15, 10, 40),
14   proportions.test = c(10, 5, 20, 15, 10, 40),
15   n.cells = 100,
16   n.bulk.samples = 31000
17 )
```

### 3. Simulación de perfiles *bulk* y preparación de los datos



1. Simulación de los perfiles *bulk* de acuerdo a las proporciones y al número de células establecidas en el paso anterior.
2. Preparación de los datos para el entrenamiento de la red neuronal y su evaluación: combinación de perfiles, normalización...

### 3. Simulación de perfiles *bulk*



$$T_{ij} = \sum_{k=1}^K \sum_{z=1}^Z C_{izk}$$

de forma que  $\begin{cases} i = 1 \dots M \\ j = 1 \dots N \\ Z = 1 \dots 100 \cdot P_{jk} \\ \sum_{k=1}^K Z \cdot P_{jk} = 100 \end{cases}$

- Sumatorio de los niveles de expresión del gen  $i$  de `n.cells` (100 por defecto) células ( $z$ ) en función del tipo celular ( $k$ ) al que pertenezcan para cada muestra *bulk* ( $j$ ).
- Implementación: posibilidad de utilizar ficheros HDF5 como *back-end*: paquetes `DelayedArray` y `HDF5Array`.

2 matrices de 23.555 filas y 18.777 columnas:  
6,5GB → 4,4kB

### 3. Preparación de los datos para entrenamiento y evaluación



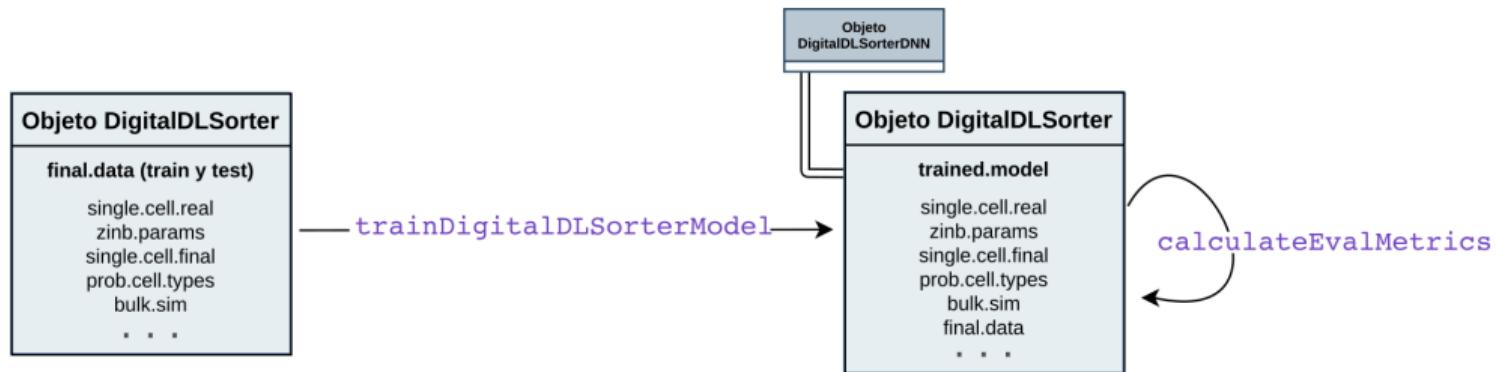
- Combinación de perfiles mediante el argumento `combine`:
  - Solo perfiles *single-cell*.
  - Solo perfiles *bulk*.
  - Combinación de ambos.
- Normalización: cálculo CPM y establecimiento de  $\mu = 0$  y  $\sigma = 1$ .
- Aleatorización de las muestras.
- Transposición de la matriz resultante.

Posibilidad de usar ficheros HDF5 → rápida lectura de las muestras durante entrenamiento y predicción.

### 3. Simulación de perfiles *bulk* y preparación de datos

```
1 ## simulation of bulk samples
2 DDLSChungSmall <- generateBulkSamples(
3   DDLSChungSmall,
4   threads = 2,
5   type.data = "both",
6   file.backend = "DDLS_bulk.h5" # if NULL, works in-memory
7 )
8
9 ## preparing samples for training (train) and prediction (test)
10 DDLSChungSmall <- prepareDataForTraining(
11   object = DDLSChungSmall,
12   type.data = "both",
13   combine = "bulk", # or both or single-cell
14   file.backend = "DDLS_final.h5" # if NULL, works in-memory
15 )
```

## 4. Entrenamiento y evaluación del modelo



1. Entrenamiento de la Red Neuronal Profunda con los datos seleccionados en el slot `final.data`.
2. Cálculo de métricas de error para determinar el desempeño del modelo durante la resolución del problema de deconvolución sobre las muestras de test.

## 4. Entrenamiento de la Red Neuronal Profunda

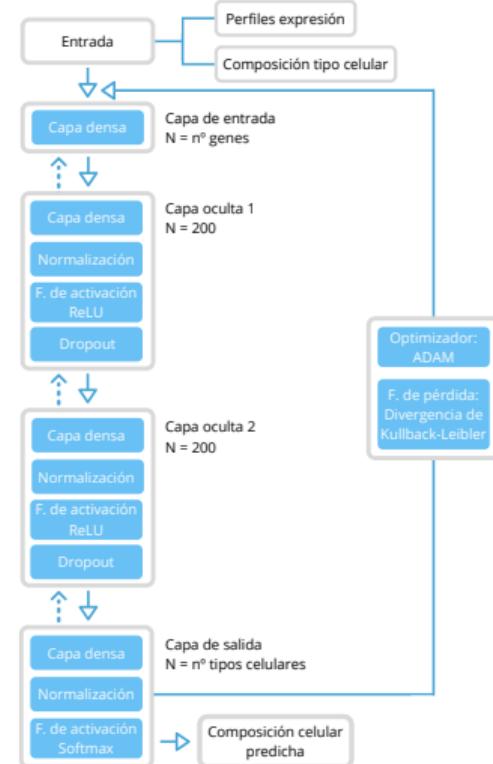
Uso de los paquetes **Keras** y **Tensorflow**: API de R para el módulo Keras de Python.

```
1 DDLSChungSmall <- trainDigitalDLSorterModel(  
2   object = DDLSChungSmall,  
3   batch.size = 128,  
4   num.epochs = 20,  
5   val = FALSE  
6 )
```

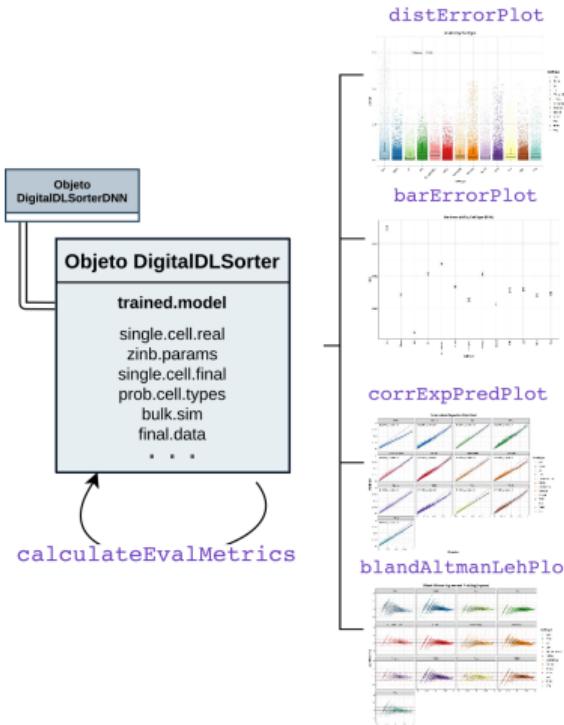
**Instalación:** README del repositorio:  
[diegommcc/digitalDLSorteR](https://github.com/diegommcc/digitalDLSorteR).

### Parámetros

- Número de épocas con `num.epochs`.
- Tamaño de *batch* con `batch.size`.
- Subset de validación con `val` y `freq.val`.
- Función de pérdida con `loss` (Divergencia de Kullback-Leibler por defecto).



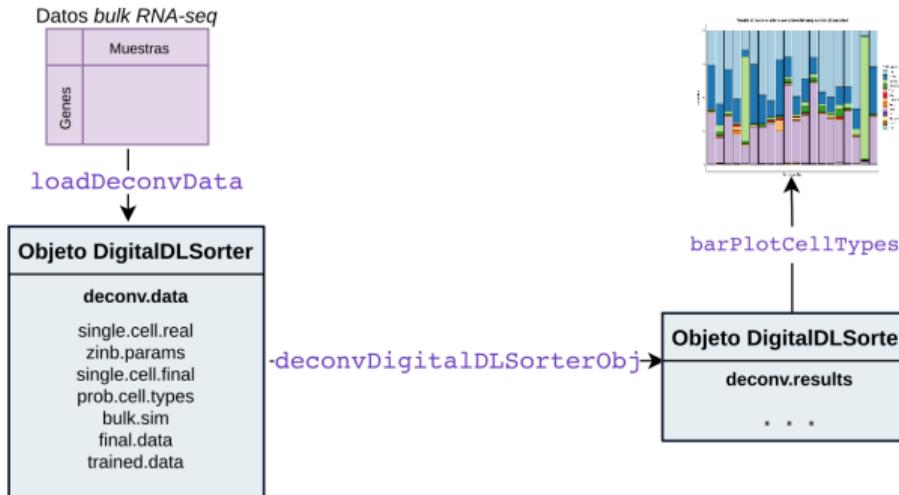
## 4. Evaluación del modelo sobre los datos de test



- `calculateEvalMetrics`: cálculo de error absoluto y error cuadrático.
- `distErrorPlot` y `barErrorPlot`: distribución de los errores en función de tipo celular, número de tipos celulares.
- `corrExpPredPlot`: gráficos de correlación → coeficiente de correlación de Pearson ( $R$ ) y Coeficiente de correlación de concordancia (CCC).
- `blandAltmanLehPlot`: gráfico de concordancia entre predicción y real.

Determinación del desempeño del modelo

## 5. Carga de nuevos datos *bulk* y deconvolución



- Carga de nuevas muestras *bulk* en el objeto *DigitalDLSorter* mediante `loadDeconvDataFromSummarizedExperiment` o `loadDeconvDataFromFile`.
- Deconvolución y representación.

## 5. Carga de nuevos datos *bulk* y deconvolución

```
1 ## new SummarizedExperiment object
2 TCGA.breast <- SummarizedExperiment(
3   assay = list(counts = TCGA.breast.small)
4 )
5 ## load SE object into DigitalDLSorter object
6 DDLSChungSmall <- loadDeconvDataFromSummarizedExperiment(
7   object = DDLSChungSmall,
8   se.object = TCGA.breast,
9   name.data = "TCGA.breast"
10 )
11 ## deconvolution using this data
12 DDLSChungSmall <- deconvDigitalDLSorterObj(
13   object = DDLSChungSmall,
14   name.data = "TCGA.breast",
15   normalize = TRUE
16 )
17 ## see results
18 barPlotCellTypes(DDLSChungSmall, name.data = "TCGA.breast")
```

# Análisis de datos *scRNA-seq* de cáncer de mama

---

## Datos utilizados

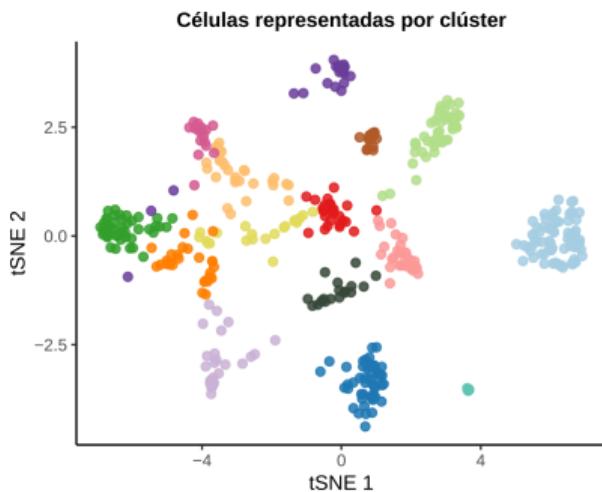
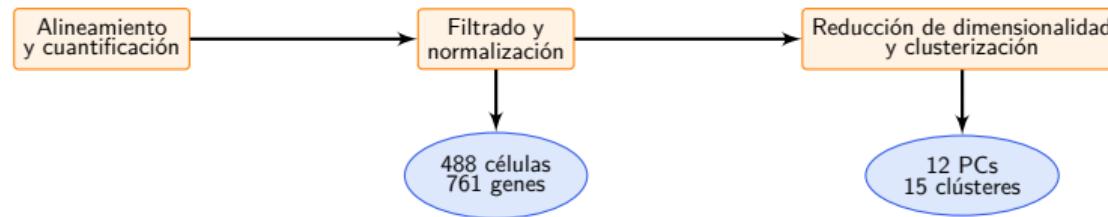
Datos *scRNA-seq* de **cáncer de mama** procedentes del artículo Chung at al., 2017.

- 11 pacientes que representan los 4 subtipos de cáncer de mama.
- Total células: 549.
- Se utilizaron las lecturas obtenidas de la base de datos SRA ([SRP066982](#)).

Paciente	Subtipo	Tejido	# células
BC01	Luminal A	Tumor primario	26
BC02	Luminal A	Tumor primario	56
BC03	Luminal B	Tumor primario	37
BC04	HER2	Tumor primario	59
BC05	HER2	Tumor primario	77
BC06	HER2	Tumor primario	25
BC07	TNBC	Tumor primario	51
BC08	TNBC	Tumor primario	23
BC09	TNBC	Tumor primario	60
BC10	TNBC	Tumor primario	16
BC11	TNBC	Tumor primario	11
BC03LN	Luminal B	Tejido linfático	55
BC07LN	TNBC	Tejido linfático	53

# Análisis: separación de células tumorales y no tumorales

- Alineamiento y cuantificación con **RSEM** y **STAR**.
- Análisis de datos *scRNA-seq* mediante el paquete de R **Seurat**.



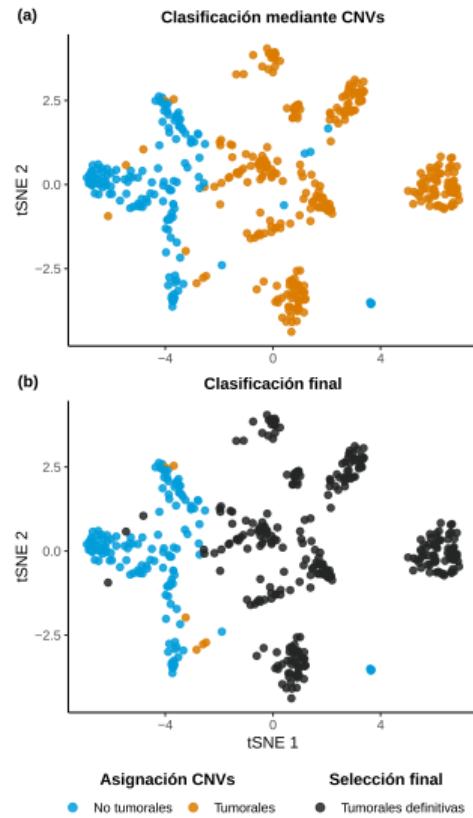
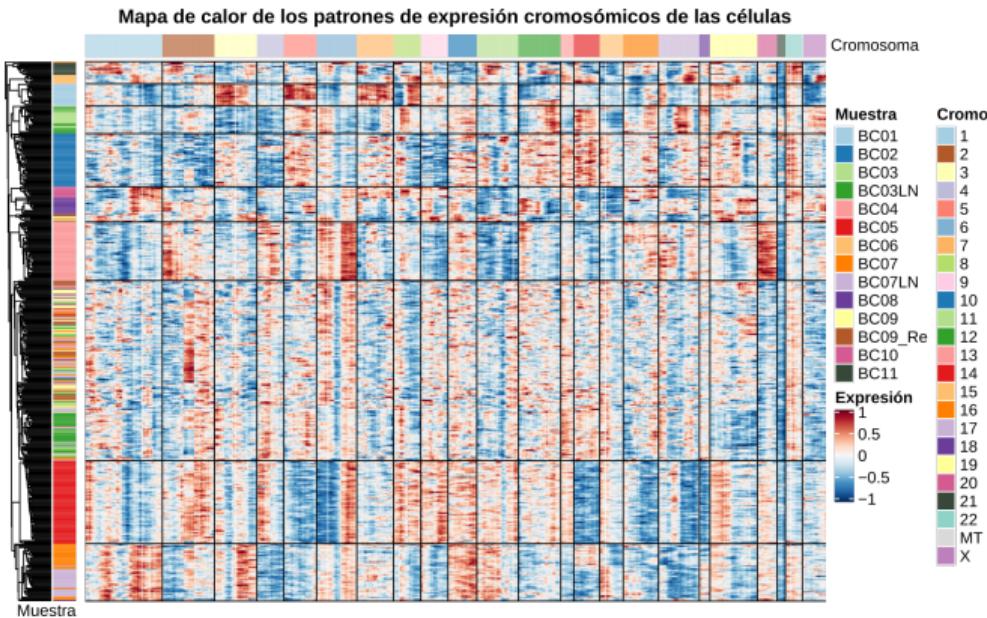
Clúster

- C0
- C1
- C2
- C3
- C4
- C5
- C6
- C7
- C8
- C9
- C10
- C11
- C12
- C13
- C14

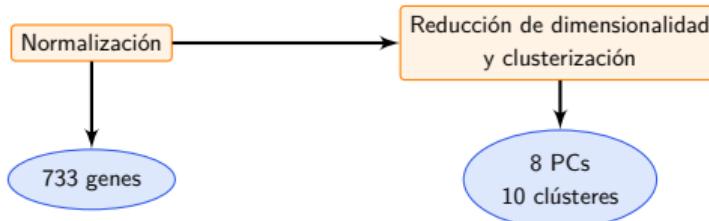
- Filtrado: 488 células.
- Normalización: 761 genes más variables.
- Reducción de la dimensionalidad: 12 componentes principales.
- Clusterización: 15 clústeres.

# Análisis: separación de células tumorales y no tumorales

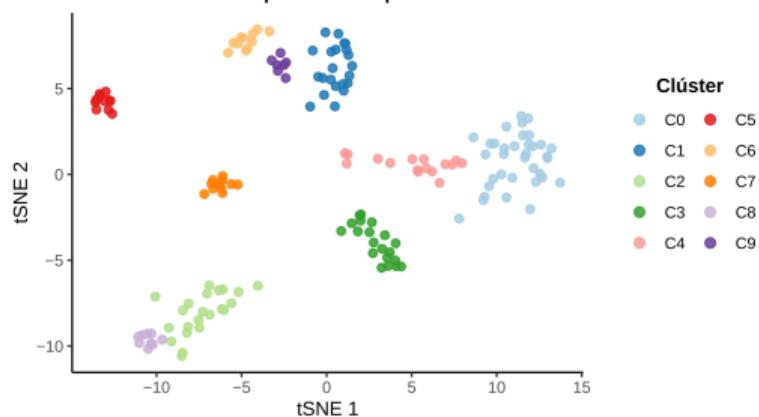
- Aproximación seguida por los autores del artículo: comparativa de los **patrones de expresión** entre células y tejido mamario normal.
- Cálculo de Z-scores y clusterización jerárquica (UPGMA).



# Análisis e identificación de células no tumorales



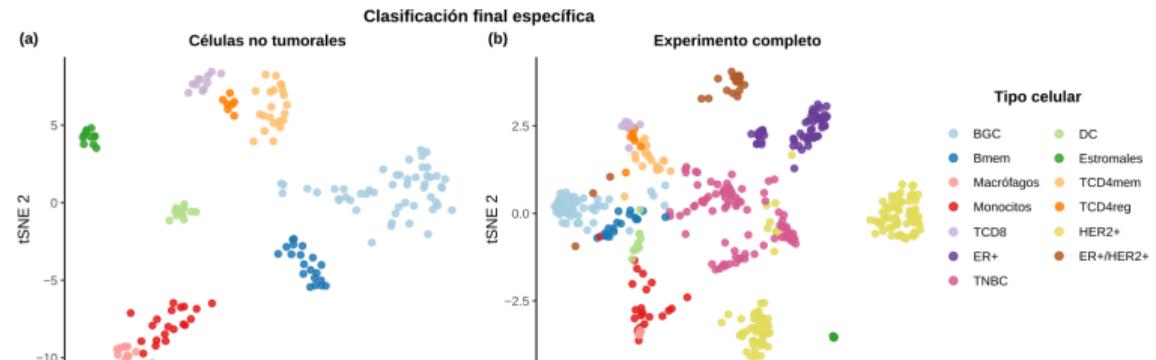
Células representadas por clúster



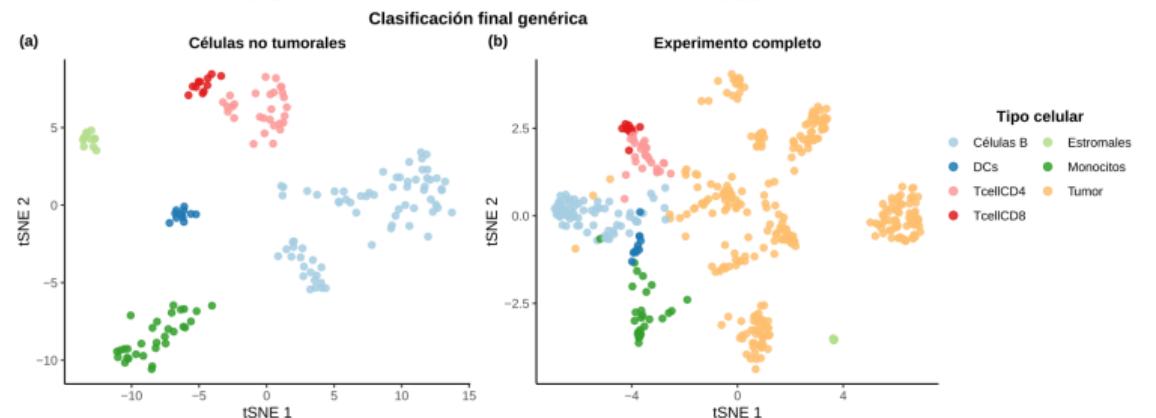
- Normalización: 733 genes más variables.
- Reducción de la dimensionalidad: 8 componentes principales.
- Clusterización: 10 clústeres.

Caracterización mediante el paquete de R SingleR  
y el análisis manual de marcadores específicos de tipo celular.

# Análisis e identificación de células no tumorales



Mayor resolución:  
13 tipos celulares

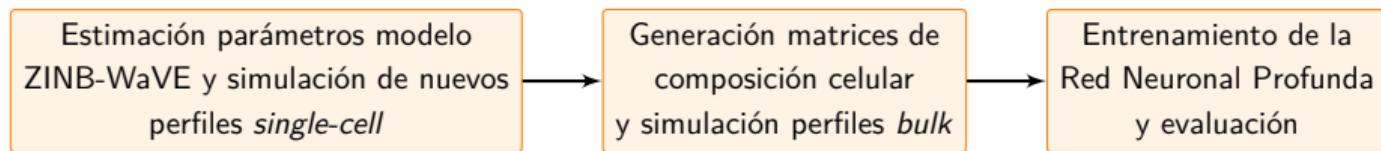


Menor resolución:  
7 tipos celulares

# **Puesta en práctica de digitalDLSorteR: deconvolución muestras de cáncer de mama**

---

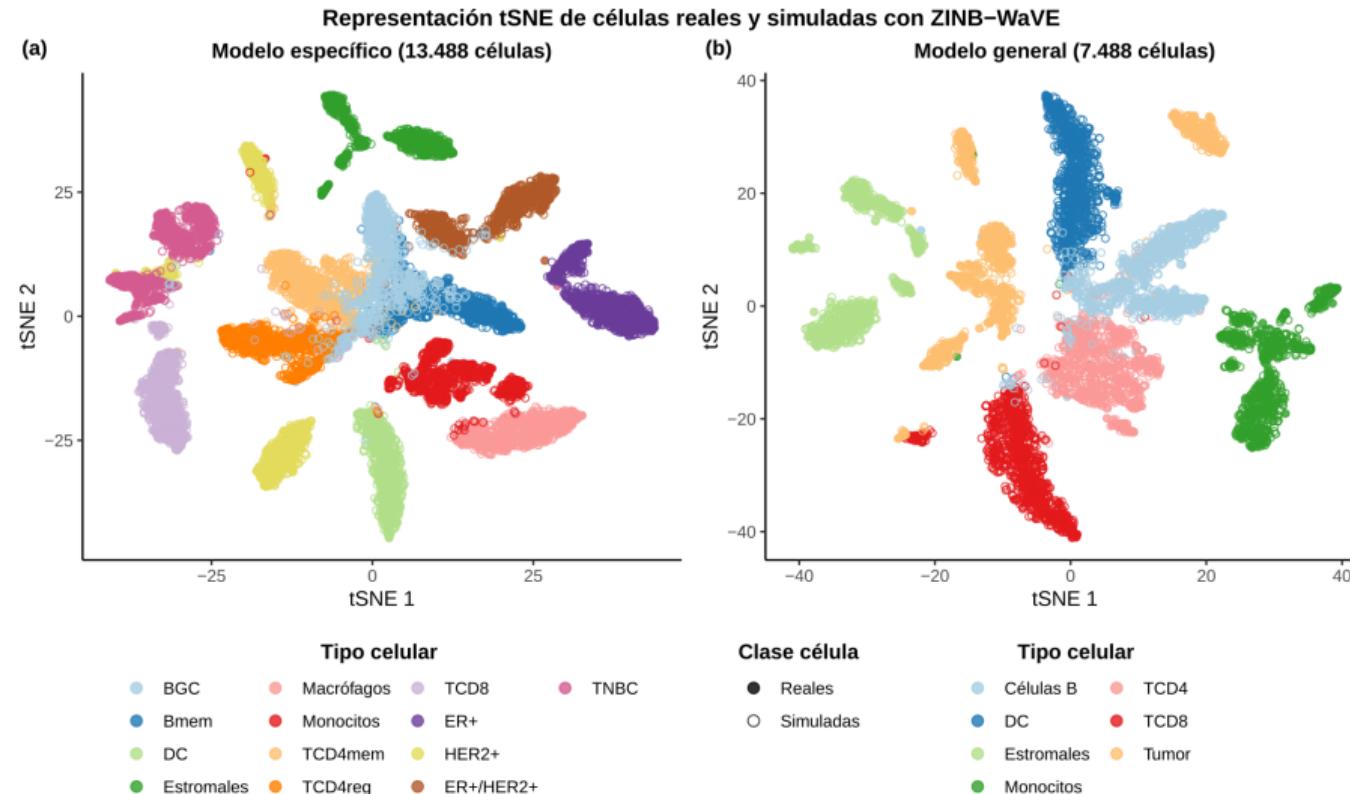
Flujo de trabajo mostrado anteriormente:



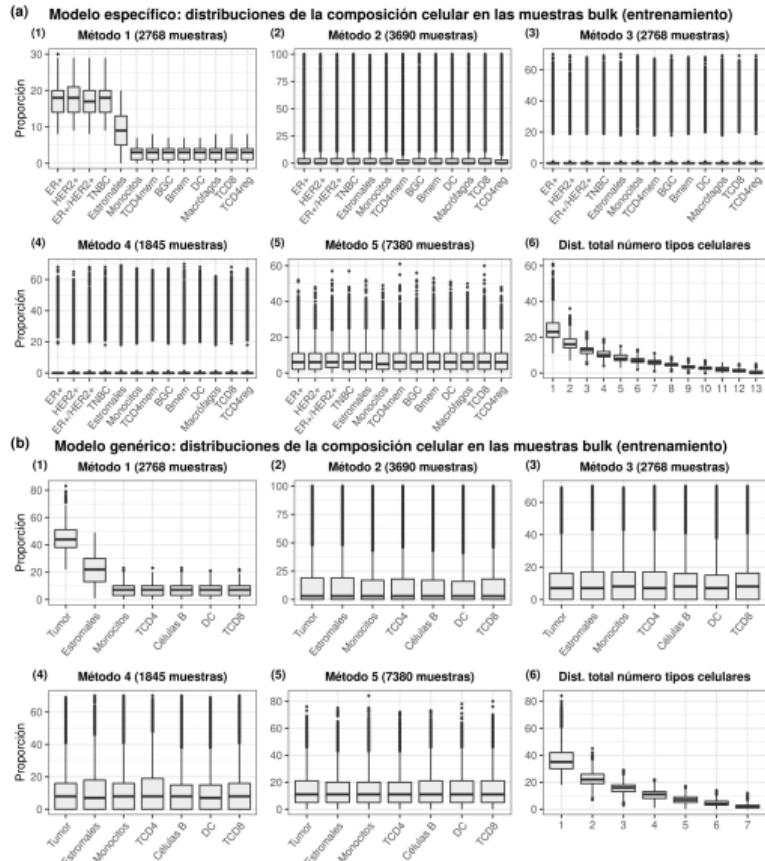
Dos modelos de deconvolución a diferente nivel de resolución. Aplicación sobre muestras reales:

- Modelo específico: 13 tipos celulares → mayor resolución aunque mayor probabilidad de error.
- Modelo genérico: 7 tipos celulares → menor resolución aunque menor probabilidad de error.

# Estimación parámetros ZINB-WaVE y simulación de nuevas células



# Matrices de composición celular y modelos



## Parámetros utilizados

- 31.000 muestras *bulk* compuestas por 100 células cada una.
- Proporciones generadas por cada modelo: por defecto → 85% con mayor aleatoriedad.

## Modelos construidos (datos entrenamiento)

1. Modelo específico 1 → solo *bulk*: 18.451 muestras.
2. Modelo específico 2 → solo *single-cell*: 8.992 células.
3. Modelo específico 3 → solo *bulk* + *single-cell*: 18.451 muestras y 8.992 células.
4. Modelo genérico: solo *bulk* → 18.451 muestras.

# Entrenamiento de la Red Neuronal Profunda

25 épocas en modelos específicos y 20 épocas en genérico. El resto de parámetros son los establecidos en `trainDigitalDLSorterModel` por defecto.

Conjunto de datos	KLD		Precisión		MAE	
	Entr.	Test	Entr.	Test	Entr.	Test
<b>Modelo específico 1 (<i>bulk</i>)</b>	0,0515	0,0751	0,7442	0,885	0,015	0,014
<b>Modelo específico 2 (<i>sc</i>)</b>	0,0289	0,5718	1	0,6131	7,6197	0,0465
<b>Modelo específico 3 (<i>bulk + sc</i>)</b>	0,0445	0,0385	0,7953	0,8101	0,0123	0,0112
<b>Modelo genérico (<i>bulk</i>)</b>	0,0222	0,0154	0,9051	0,9629	0,0186	0,0112

- Modelo específico 2 con los peores resultados, gran sobreajuste → necesidad de incluir muestras *bulk*.
- Modelo específico 3 superior al Modelo específico 1 → necesaria una comparativa más profunda.
- Modelo genérico con las mejores puntuaciones → problema más sencillo.

# Entrenamiento de la Red Neuronal Profunda

25 épocas en modelos específicos y 20 épocas en genérico. El resto de parámetros son los establecidos en `trainDigitalDLSorterModel` por defecto.

Conjunto de datos	KLD		Precisión		MAE	
	Entr.	Test	Entr.	Test	Entr.	Test
<b>Modelo específico 1 (<i>bulk</i>)</b>	0,0515	0,0751	0,7442	0,885	0,015	0,014
<b>Modelo específico 2 (<i>sc</i>)</b>	0,0289	0,5718	1	0,6131	7,6197	0,0465
<b>Modelo específico 3 (<i>bulk + sc</i>)</b>	0,0445	0,0385	0,7953	0,8101	0,0123	0,0112
<b>Modelo genérico (<i>bulk</i>)</b>	0,0222	0,0154	0,9051	0,9629	0,0186	0,0112

- Modelo específico 2 con los peores resultados, gran sobreajuste → necesidad de incluir muestras *bulk*.
- Modelo específico 3 superior al Modelo específico 1 → necesaria una comparativa más profunda.
- Modelo genérico con las mejores puntuaciones → problema más sencillo.

# Entrenamiento de la Red Neuronal Profunda

25 épocas en modelos específicos y 20 épocas en genérico. El resto de parámetros son los establecidos en `trainDigitalDLSorterModel` por defecto.

Conjunto de datos	KLD		Precisión		MAE	
	Entr.	Test	Entr.	Test	Entr.	Test
<b>Modelo específico 1 (<i>bulk</i>)</b>	0,0515	0,0751	0,7442	0,885	0,015	0,014
<b>Modelo específico 2 (<i>sc</i>)</b>	0,0289	0,5718	1	0,6131	7,6197	0,0465
<b>Modelo específico 3 (<i>bulk + sc</i>)</b>	0,0445	0,0385	0,7953	0,8101	0,0123	0,0112
<b>Modelo genérico (<i>bulk</i>)</b>	0,0222	0,0154	0,9051	0,9629	0,0186	0,0112

- Modelo específico 2 con los peores resultados, gran sobreajuste → necesidad de incluir muestras *bulk*.
- **Modelo específico 3 superior al Modelo específico 1 → necesaria una comparativa más profunda.**
- Modelo genérico con las mejores puntuaciones → problema más sencillo.

# Entrenamiento de la Red Neuronal Profunda

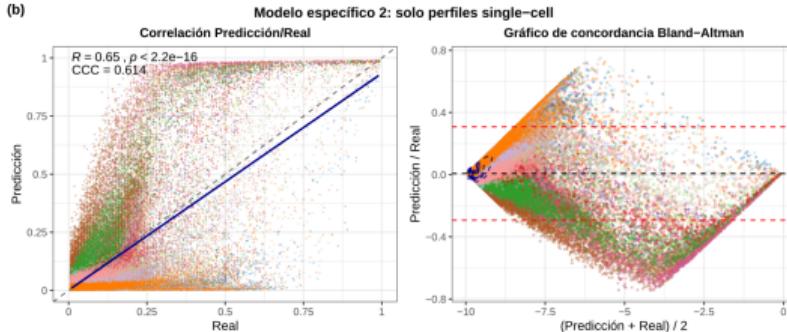
25 épocas en modelos específicos y 20 épocas en genérico. El resto de parámetros son los establecidos en `trainDigitalDLSorterModel` por defecto.

Conjunto de datos	KLD		Precisión		MAE	
	Entr.	Test	Entr.	Test	Entr.	Test
<b>Modelo específico 1 (<i>bulk</i>)</b>	0,0515	0,0751	0,7442	0,885	0,015	0,014
<b>Modelo específico 2 (<i>sc</i>)</b>	0,0289	0,5718	1	0,6131	7,6197	0,0465
<b>Modelo específico 3 (<i>bulk + sc</i>)</b>	0,0445	0,0385	0,7953	0,8101	0,0123	0,0112
<b>Modelo genérico (<i>bulk</i>)</b>	0,0222	0,0154	0,9051	0,9629	0,0186	0,0112

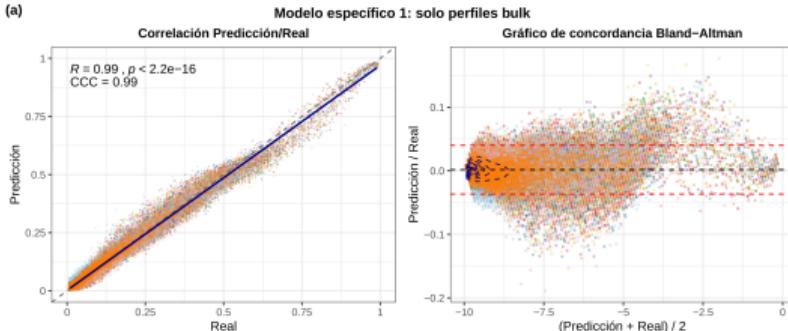
- Modelo específico 2 con los peores resultados, gran sobreajuste → necesidad de incluir muestras *bulk*.
- Modelo específico 3 superior al Modelo específico 1 → necesaria una comparativa más profunda.
- **Modelo genérico con las mejores puntuaciones → problema más sencillo.**

# Evaluación del desempeño de la deconvolución: modelos específicos

(b)



(a)

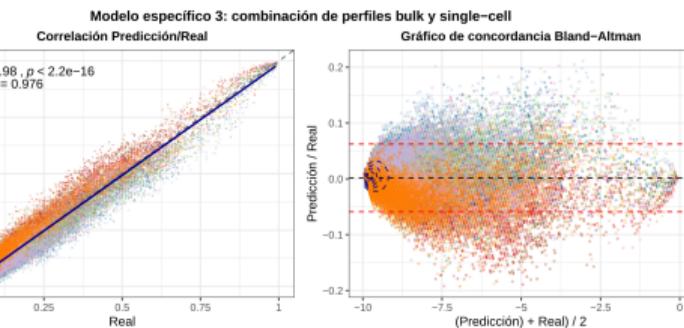


## Resultados

- Modelo específico 2: peores resultados  $\rightarrow R = 0,65$  y  $CCC = 0,614$
- Modelos específicos 1 y 3: similares, ligeramente superior el 1.

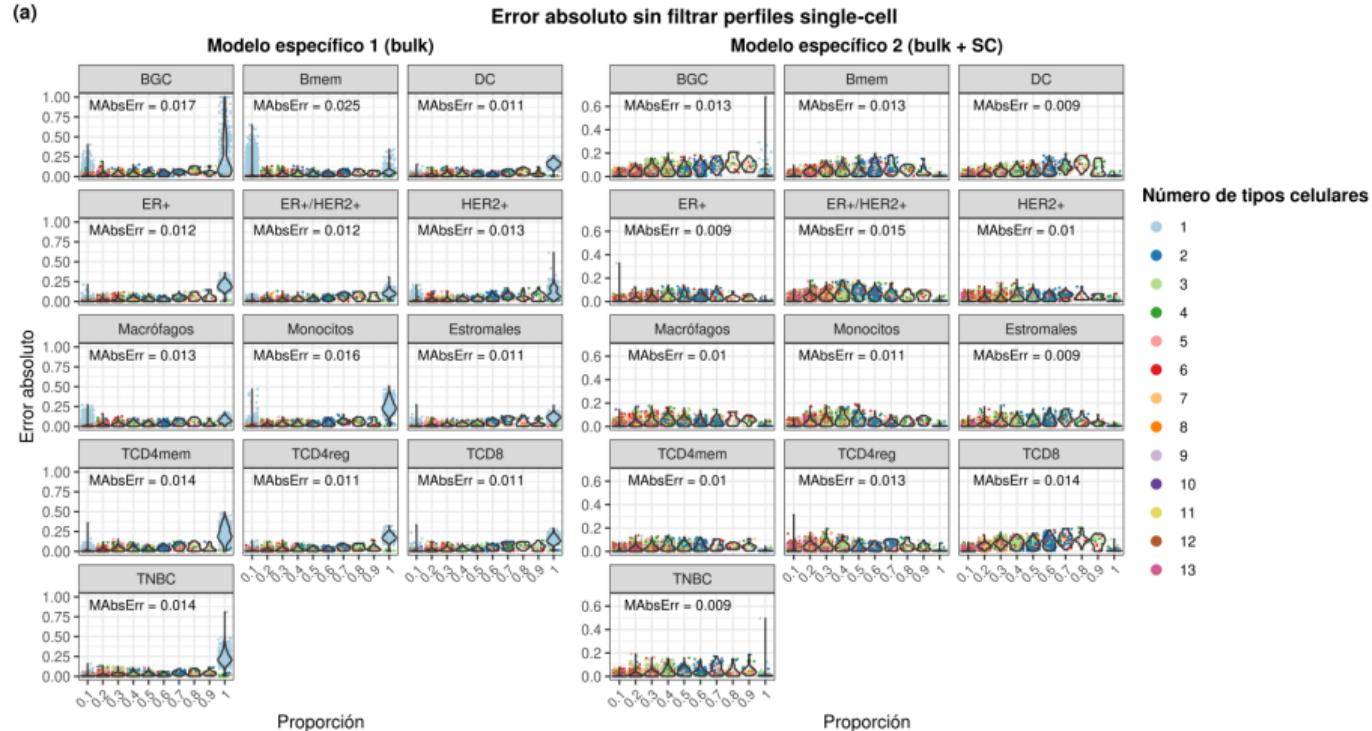
II

(c)



# Modelo específico 1 vs Modelo específico 3

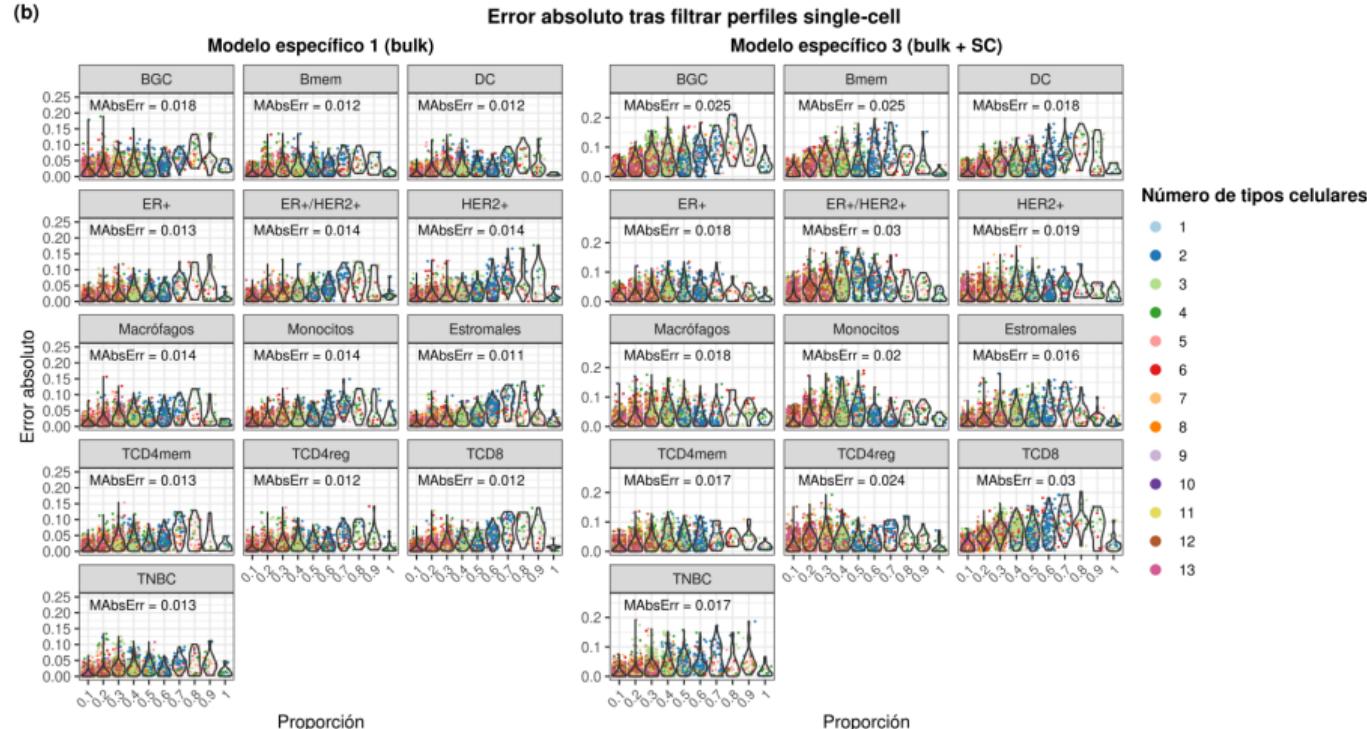
(a)



- Modelo entrenado con *single-cell* + *bulk* → superior sin filtrar perfiles *single-cell*.
- Modelo entrenado solo con *bulk* → superior retirando los perfiles *single-cell*.

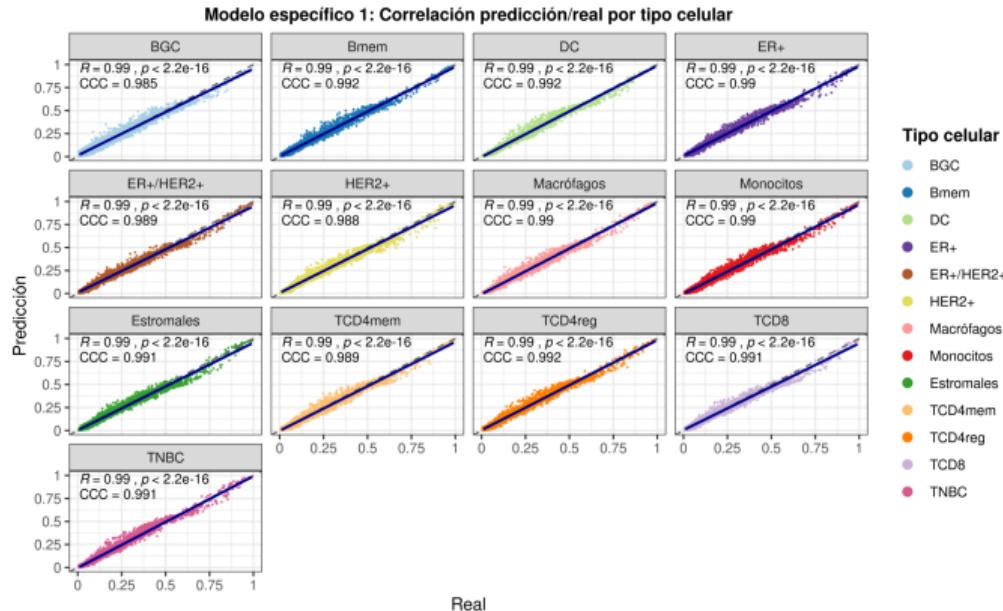
# Modelo específico 1 vs Modelo específico 3

(b)



- Modelo entrenado con *single-cell* + *bulk* → superior sin filtrar perfiles *single-cell*.
- Modelo entrenado solo con *bulk* → superior retirando los perfiles *single-cell*.

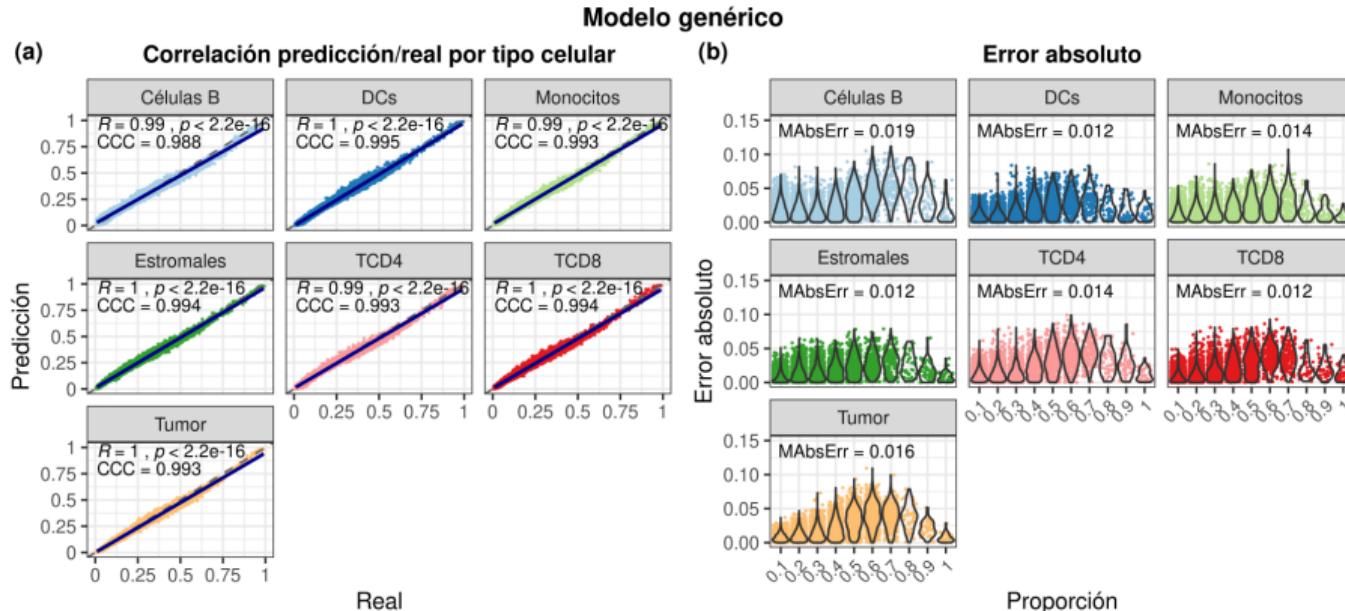
# Modelo específico 1



- Niveles de correlación cercanos a 0,99 tanto en  $R$  como en CCC.
- Errores mayores en tipos celulares tumorales y BGC.

Buen desempeño general en todos los tipos celulares.

# Modelo genérico



- Mismas tendencias que las observadas en el modelo específico.
- Errores menores en general.

Buen desempeño general en todos los tipos celulares.

# Modelos de deconvolución sobre datos *bulk* reales

Datos *bulk RNA-seq* de cáncer de mama procedentes del proyecto TCGA (*The Cancer Genome Atlas*): 1222 muestras.

## Problema

- No se tienen las proporciones celulares de las muestras.
- Análisis de correlación entre las proporciones predichas.

## Se espera

Tipos pro-tumorales correlacionen positivamente con proporciones tumorales:

- Linfocitos T reguladores
- Macrófagos.

Tipos anti-tumorales correlacionen negativamente con proporciones tumorales:

- Células B de memoria.
- Linfocitos T CD8+.

# Modelos de deconvolución sobre datos *bulk* reales

Datos *bulk RNA-seq* de cáncer de mama procedentes del proyecto TCGA (*The Cancer Genome Atlas*): 1222 muestras.

## Problema

- No se tienen las proporciones celulares de las muestras.
- Análisis de correlación entre las proporciones predichas.

## Se espera

Tipos pro-tumorales correlacionen positivamente con proporciones tumorales:

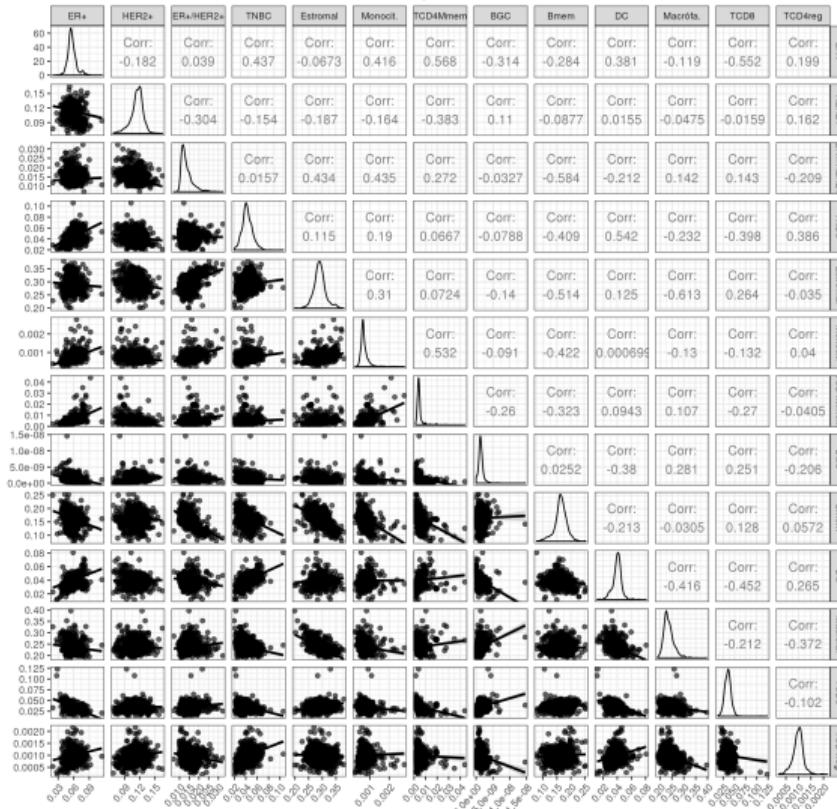
- Linfocitos T reguladores
- Macrófagos.

Tipos anti-tumorales correlacionen negativamente con proporciones tumorales:

- Células B de memoria.
- Linfocitos T CD8+.

# Modelo específico sobre datos TCGA

Evaluación del modelo específico sobre los datos TCGA



- Lincofitos T CD8+ y linfocitos B de memoria → anticorrelacionan con 3 de los 4 subtipos de cáncer.
- Linfocitos T CD4+ reguladores: correlacionan positivamente con 3 de los 4 subtipos de cáncer.

## Problema

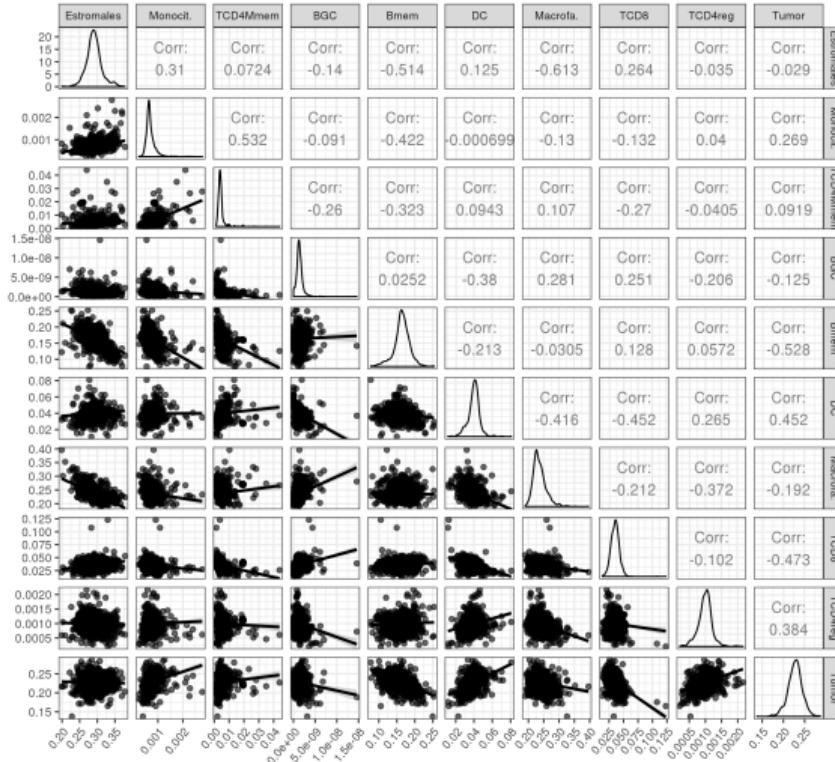
ER+/HER2+ no es predicho según lo esperado. → proporciones medias de 1,14%.

## Solución

Uso del argumento `simplify.set` agrupando los 4 subtipos intrínsecos de cáncer bajo la etiqueta 'Tumor'.

# Modelo específico sobre datos TCGA

Modelo específico: tipos tumorales colapsados



- Lincofitos T CD8+ y linfocitos B de memoria → anticorrelacionan con 3 de los 4 subtipos de cáncer.
- Linfocitos T CD4+ reguladores: correlacionan positivamente con 3 de los 4 subtipos de cáncer.

## Problema

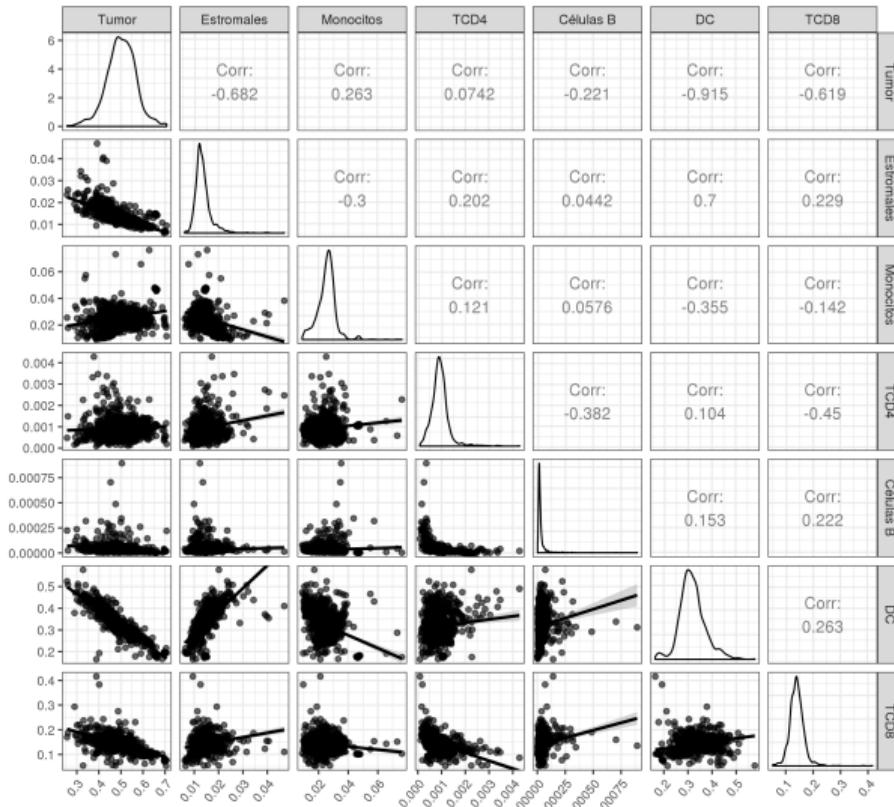
ER+/HER2+ no es predicho según lo esperado. → proporciones medias de 1,14%.

## Solución

Uso del argumento `simplify.set` agrupando los 4 subtipos intrínsecos de cáncer bajo la etiqueta 'Tumor'.

# Modelo genérico sobre datos TCGA

Evaluación del modelo genérico sobre los datos TCGA



## Mismas tendencias

- Lincofitos T CD8+ y linfocitos B de memoria → anticorrelacionan con la proporción tumoral.
- Linfocitos T CD4+ reguladores: correlacionan positivamente con la proporción tumoral.
- Consenso entre ambos modelos excepto en las células dendríticas: posible problema en los datos de entrenamiento.

## **Conclusiones y trabajo futuro**

---

## Resultados obtenidos

- Modelos entrenados con perfiles *scRNA-seq* procedentes del entorno de estudio.
- Mejora constante de conjuntos de datos *scRNA-seq* → mejores modelos de deconvolución.
- ZINB-WaVE ofrece buenos perfiles *single-cell* simulados.
- Uso de Redes Neuronales: buenas representaciones de los tipos celulares.
- Simulación de perfiles *bulk*: reduce el ruido intrínseco de los perfiles *single-cell* y fuerza a encontrar patrones específicos.

## Limitaciones y trabajo futuro

- Altamente dependiente de la calidad de los datos de partida.
- Interpretabilidad pobre: naturaleza de 'caja negra' de las redes neuronales.

## Aportaciones

- Implementación formal de cada uno de los pasos.
- Funciones de evaluación del desempeño de los modelos.
- Parámetros que permiten modelos más personalizables.
- Uso de ficheros HDF5 como *back-end*.

## Limitaciones y trabajo futuro

- Ausencia del directorio test en el paquete: han sido incluidos los test correspondientes a la carga de datos (fichero loadData.R) y a la generación de las matrices de composición celular (fichero generateTranAndTestProbMatrix.R).
- Carga de las muestras *bulk* simuladas en memoria antes de su escritura en disco como ficheros HDF5.
- Imposibilidad de parallelizar entrenamiento red neuronal.
- Permitir el uso de redes neuronales arquitecturas diferentes.
- Implementación del método como servicio Web (aplicación Shiny).

## Referencias

---

- W. Chung, H. H. Eum, H. O. Lee, K. M. Lee, H. B. Lee, K. T. Kim, H. S. Ryu, S. Kim, J. E. Lee, Y. H. Park, Z. Kan, W. Han, y W. Y. Park. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat Commun*, 8:15081, may. 2017. doi: 10.1038/ncomms15081.
- C. Torroja y F. Sanchez-Cabo. DigitalDlsorter: Deep-Learning on scRNA-Seq to deconvolute gene expression data. *Front Genet*, 10:978, oct. 2019. doi: 10.3389/fgene.2019.00978.

**Gracias por vuestra atención.**

## Extra

### F-measure

$$F(c_i, k_j) = \frac{2 \times Pr_{ij} \times Re_{ij}}{Pr_{ij} + Re_{ij}}$$

$$F_{mean}(C, K) = \sum_{ci} \frac{|c_i|}{N} \max_{kj} F(c_i, k_j)$$

### Normalized Mutual Information

$$NMI = \frac{I(X; Y)}{\sqrt{H(X)H(Y)}}$$

### Non-redundancy score (NRS)

$$NRS(A_p) = \sum_{k=1:c} |coeff(k)| \times eigenvalue(k)$$