

METODOLOGÍAS DE
IDENTIFICACIÓN DE
TIPOS CELULARES A
PARTIR DE DATOS DE
SECUENCIACIÓN DE
RNA DE CÉLULA
SENCILLA

ESTUDIANTE: MELISSA ROBLES

ASESOR: JORGE DUITAMA

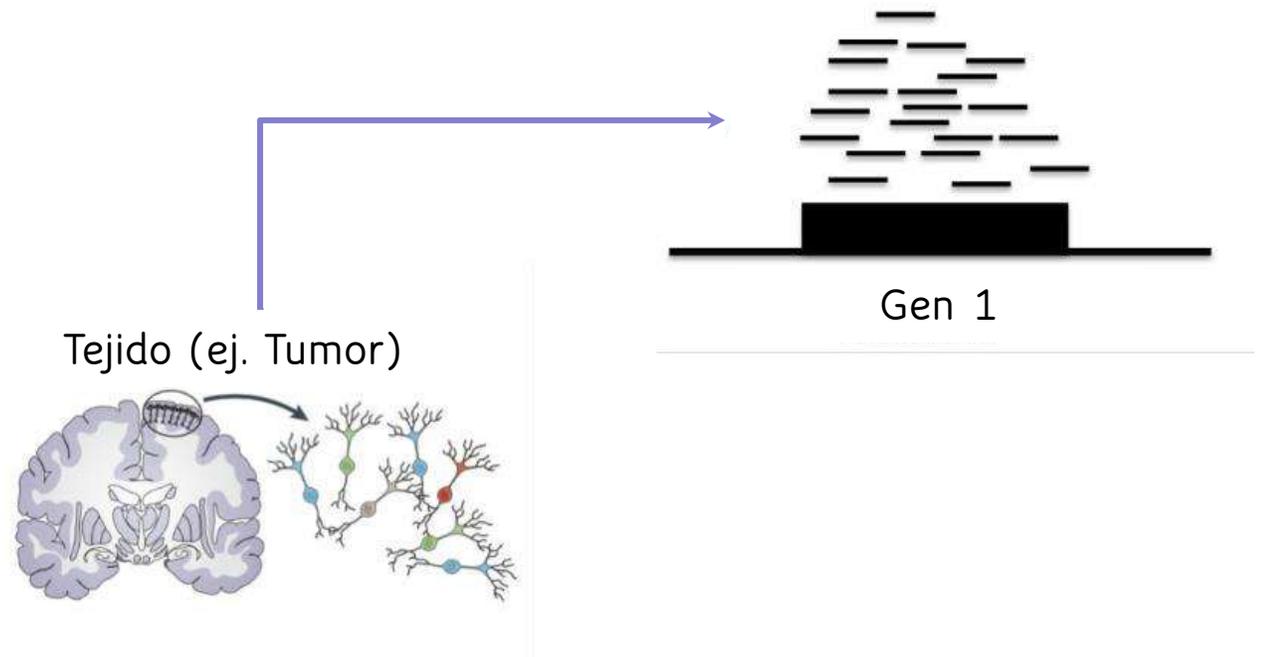
Introducción

Secuenciación de RNA

Objetivo: Identificar qué genes están activos y cuantificar su nivel de expresión en una muestra.

¿Por qué es importante?

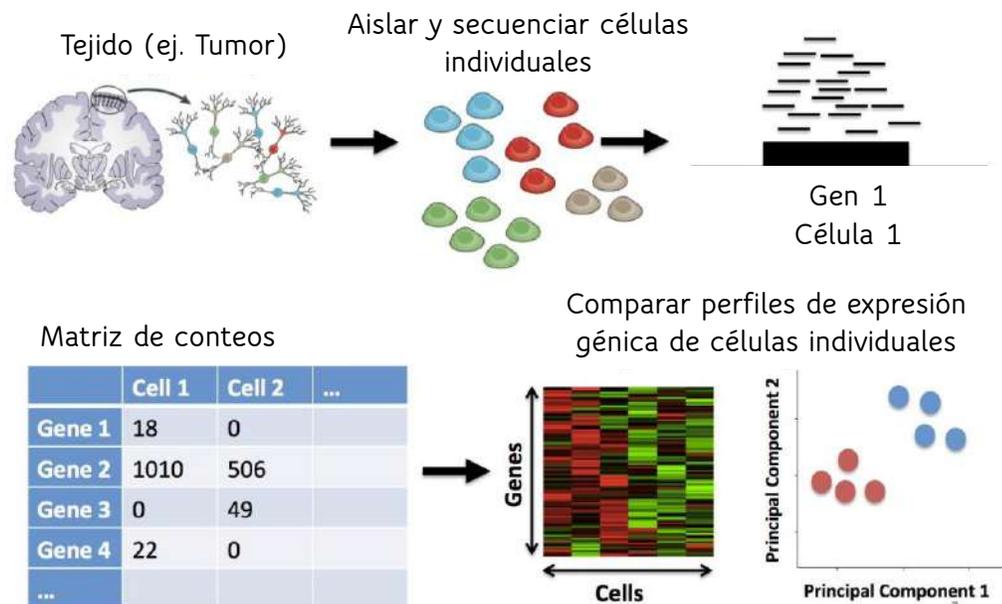
- La expresión de los genes cambia dependiendo del tejido analizado, el tipo celular, la edad, enfermedades, entre otros.
- Cuantificar el nivel de expresión de los genes permite:
 - Detectar diferencias a nivel de gen entre pacientes enfermos y sanos.
 - Estudiar el efecto de un nuevo medicamento en pacientes.



Introducción

Secuenciación de RNA de célula sencilla

Secuenciación de RNA de célula única (scRNA-seq)



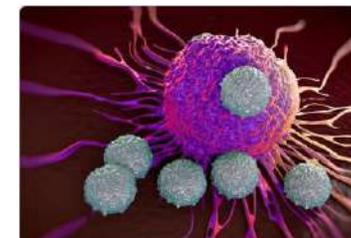
¿Por qué es importante?

- Mediante este tipo de secuenciación se conoce la expresión génica a nivel de célula y no de muestra.
- Permite un análisis más detallado. En particular, puedo hacer análisis a nivel de **tipo celular**.

Tipo celular: Clasificación de las células por sus propiedades moleculares, morfológicas, fisiológicas y funcionales.



Eritrocitos



Linfocito

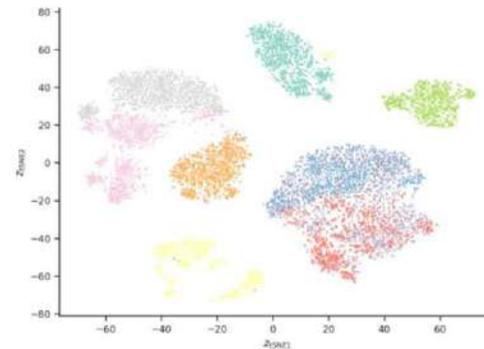
Introducción

Problema: Identificación de tipos celulares

Objetivo: Agrupar las células de un experimento de scRNA-seq por su perfil de expresión.

Matriz de conteos

	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			



$$f : \text{Cells} \rightarrow \{1, \dots, K\}$$

Desafíos

- Alta dimensionalidad:
 - Matrices de conteos en humanos presentan entre 20.000 y 35.000 genes.
- Matriz dispersa:
 - >90% de entradas son 0.
- Ruido y errores técnicos
 - Inflación de ceros (dropouts).

Objetivos

Objetivo principal

Proponer y desarrollar una solución de software que integre técnicas novedosas para agrupamiento de células a partir de datos de secuenciación de RNA de célula sencilla.

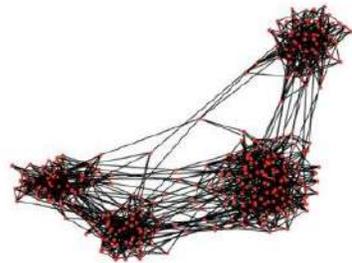
Objetivos específicos

01. Proponer nuevas metodologías para agrupamiento de células a partir de datos de RNA basados en la estructura de los datos.
02. Comparar las metodologías propuestas con herramientas comúnmente utilizadas en la actualidad para el agrupamiento de datos de RNA y con metodologías tradicionales para agrupamiento de datos de alta dimensionalidad.
03. Desarrollar un demo de una aplicación que implemente los mejores métodos desarrollados durante el proyecto.

Metodología

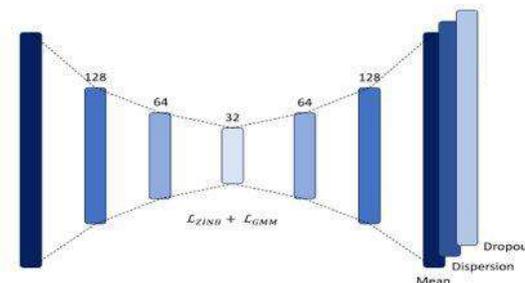
Metodologías basadas en grafos

- Utilizar grafos para representar los datos en una estructura más sencilla.
- Utilizar algoritmos de detección de comunidades sobre el grafo construido.



Metodologías basadas en redes neuronales

- Utilizar arquitecturas de tipo Autoencoder para
 - Reducción de dimensionalidad
 - Limpieza del ruido
 - Aprender parámetros de las distribuciones de los clusters.



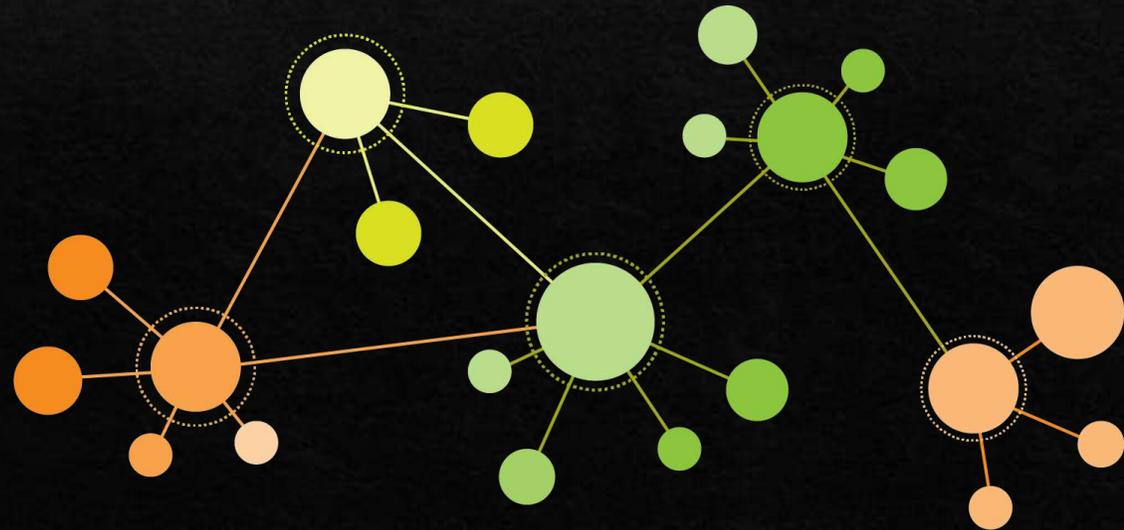
Metodologías basadas en estadística bayesiana

- Utilizar la base de *The Human Protein Atlas* para asignar tipos celulares.

31 tejidos

85 tipos celulares

METODOLOGÍAS BASADAS EN GRAFOS



Metodología basada en grafos

Definición de similaridad entre células:

- Distancia euclídeana.
- Distancia coseno.
- Distancia coseno sobre la matriz tf-idf.
- Correlación de Pearson ($1 - c_{ij}$)
- Correlación de Pearson considerando la dimensión intrínseca local de los datos..

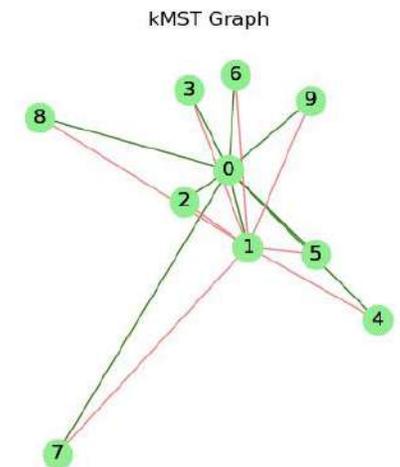
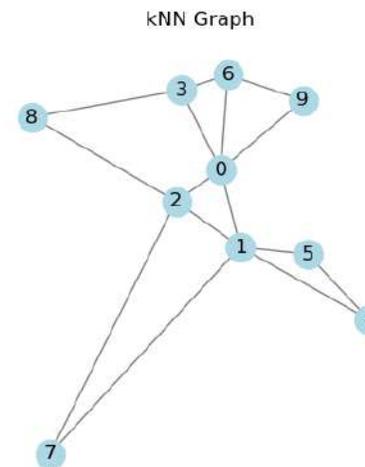
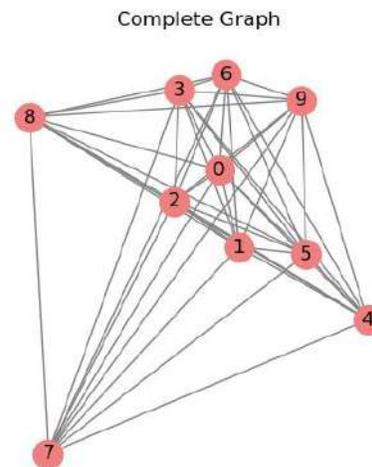
Tipos de grafos:

- Completo
- k-NN
- k-MST

Detección de comunidades:

- Louvain

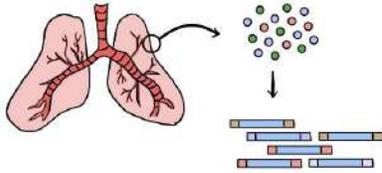
Tipos de grafos considerados



Metodología basada en grafos

Experimento scRNA-seq

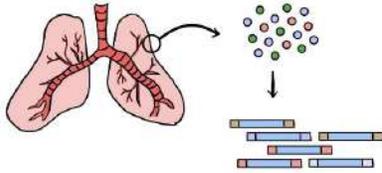
a)



Metodología basada en grafos

Experimento scRNA-seq

a)



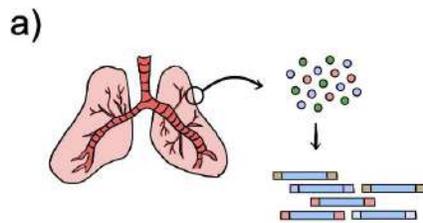
b)

Conteos

genes	1	0	2	...	0	0
	0	0	0	...	1	0
	0	3	0	...	9	2
	⋮	⋮	⋮	⋮	⋮	⋮
	0	1	3	...	1	0
	0	0	0	...	0	0
	0	0	0	...	0	2
	cells					

Metodología basada en grafos

Experimento scRNA-seq



Conteos

b)

	1	0	2	...	0	0
	0	0	0	...	1	0
	0	3	0	...	9	2
	⋮	⋮	⋮	⋱	⋮	⋮
genes	0	1	3	...	1	0
	0	0	0	...	0	0
	0	0	0	...	0	0
	0	0	0	...	0	2
						cells

Filtro de genes

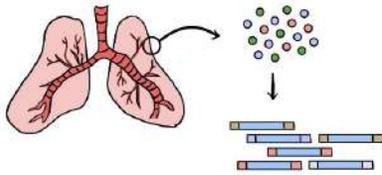
c)

	1	0	2	...	0	0
	0	3	0	...	9	2
	⋮	⋮	⋮	⋱	⋮	⋮
genes	0	1	3	...	1	0
	0	0	0	...	0	2
						cells

Metodología basada en grafos

Experimento scRNA-seq

a)



Conteos

b)

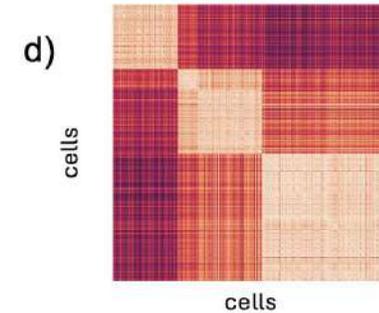
genes	1	0	2	...	0	0
	0	0	0	...	1	0
	0	3	0	...	9	2
	⋮	⋮	⋮	⋱	⋮	⋮
	0	1	3	...	1	0
	0	0	0	...	0	0
	0	0	0	...	0	0
	0	0	0	...	0	2
					cells	

Filtro de genes

c)

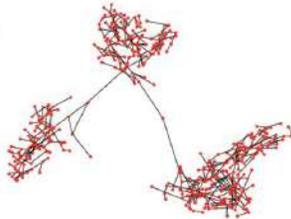
genes	1	0	2	...	0	0
	0	3	0	...	9	2
	⋮	⋮	⋮	⋱	⋮	⋮
	0	1	3	...	1	0
	0	0	0	...	0	2
					cells	

Correlación de Pearson



Construcción del grafo

e)



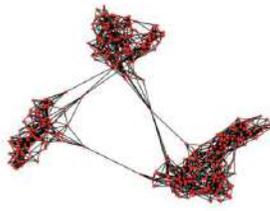
1-MST



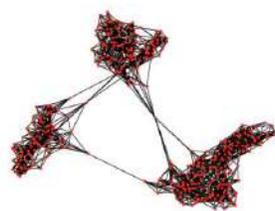
2-MST



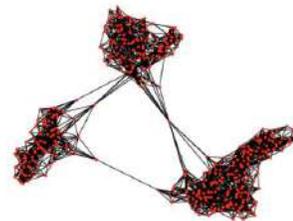
3-MST



4-MST



5-MST

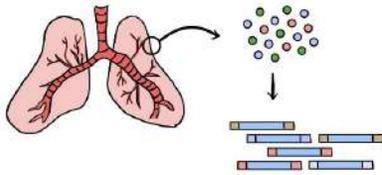


6-MST

Metodología basada en grafos

Experimento scRNA-seq

a)



Conteos

b)

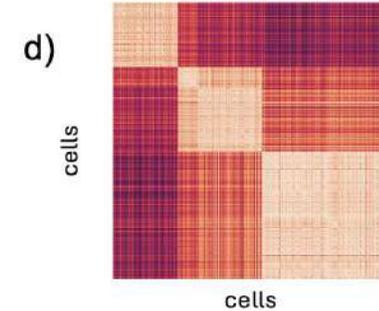
	1	0	2	...	0	0
genes	0	0	0	...	1	0
	0	3	0	...	9	2
	⋮	⋮	⋮	⋱	⋮	⋮
	0	1	3	...	1	0
	0	0	0	...	0	0
	0	0	0	...	0	2
					cells	

Filtro de genes

c)

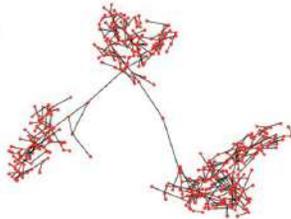
	1	0	2	...	0	0
genes	0	3	0	...	9	2
	⋮	⋮	⋮	⋱	⋮	⋮
	0	1	3	...	1	0
	0	0	0	...	0	2
					cells	

Correlación de Pearson



Construcción del grafo

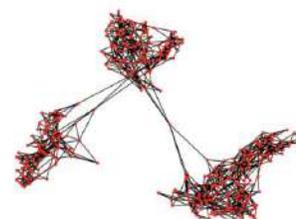
e)



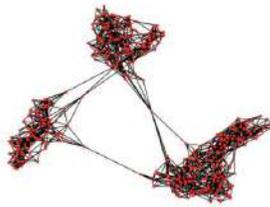
1-MST



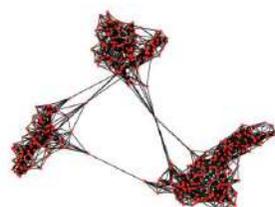
2-MST



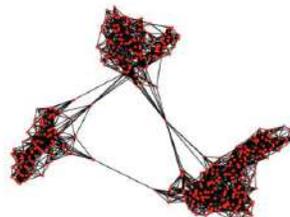
3-MST



4-MST

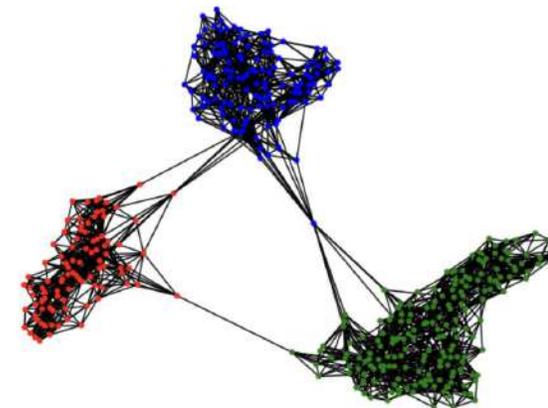


5-MST



6-MST

f)

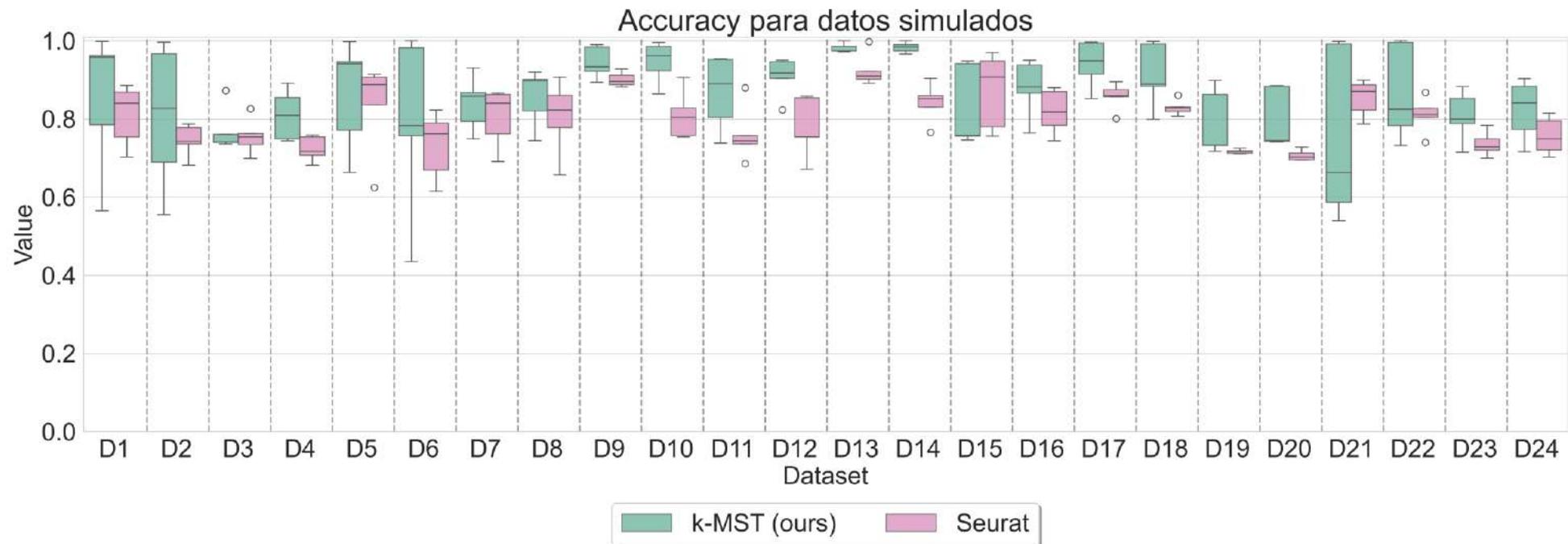


Datos simulados

- Se generaron datos simulados con la herramienta **Symsim** (simulador de datos de scRNA-seq).
- Los **parámetros** variados fueron:
 - Número de genes, número de células, número de clusters y profundidad promedio.
- Por cada combinación se crearon 5 réplicas para probar la estabilidad de los algoritmos
- Porcentaje de ceros promedio: **83.39%**.

Parámetro	Valores
Número de células	500, 750, 1000
Número de genes	1500, 2000
Número de clusters	3, 5
Profundidad promedio	1000, 2000

Resultados metodología basada en grafos



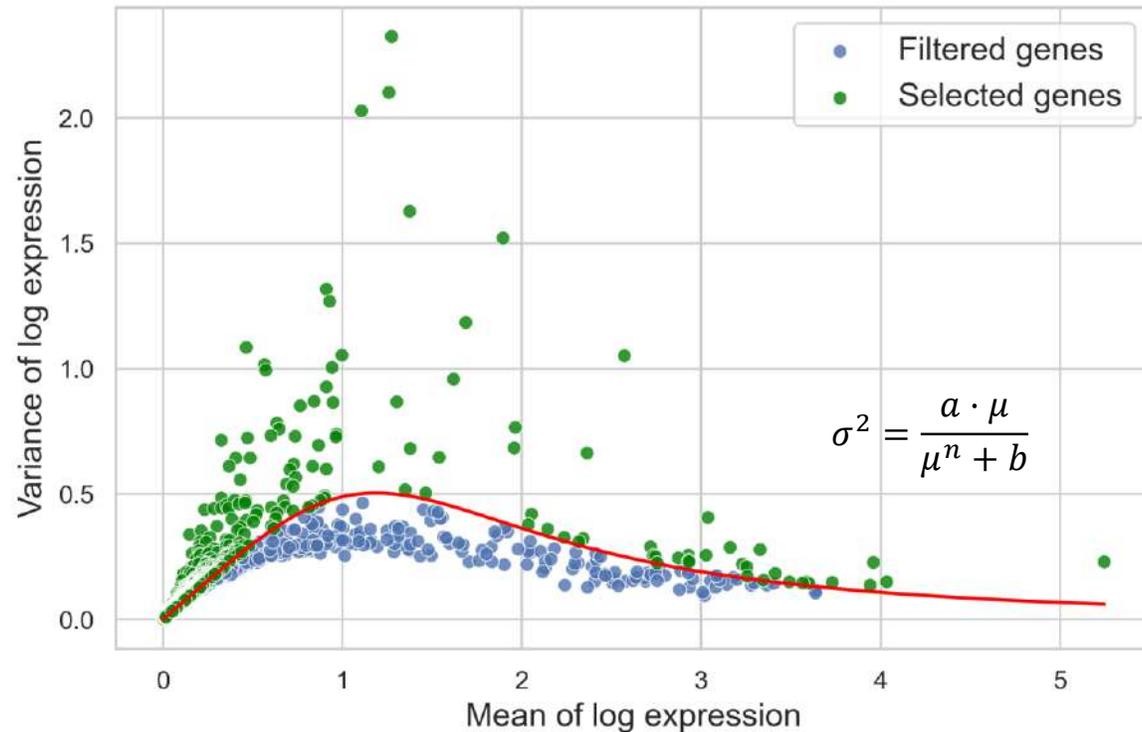
Promedio de accuracy

k-MST	85.9%
Seurat	79.9%

Promedio de std sobre réplicas

k-MST	9.35%
Seurat	5.44%

Resultados metodología basada en grafos



Selección de genes:

- Sin selección
- 5000 genes con mayor varianza.
- 5000 genes con mayor varianza observada con respecto a la esperada.

Procesamiento de datos

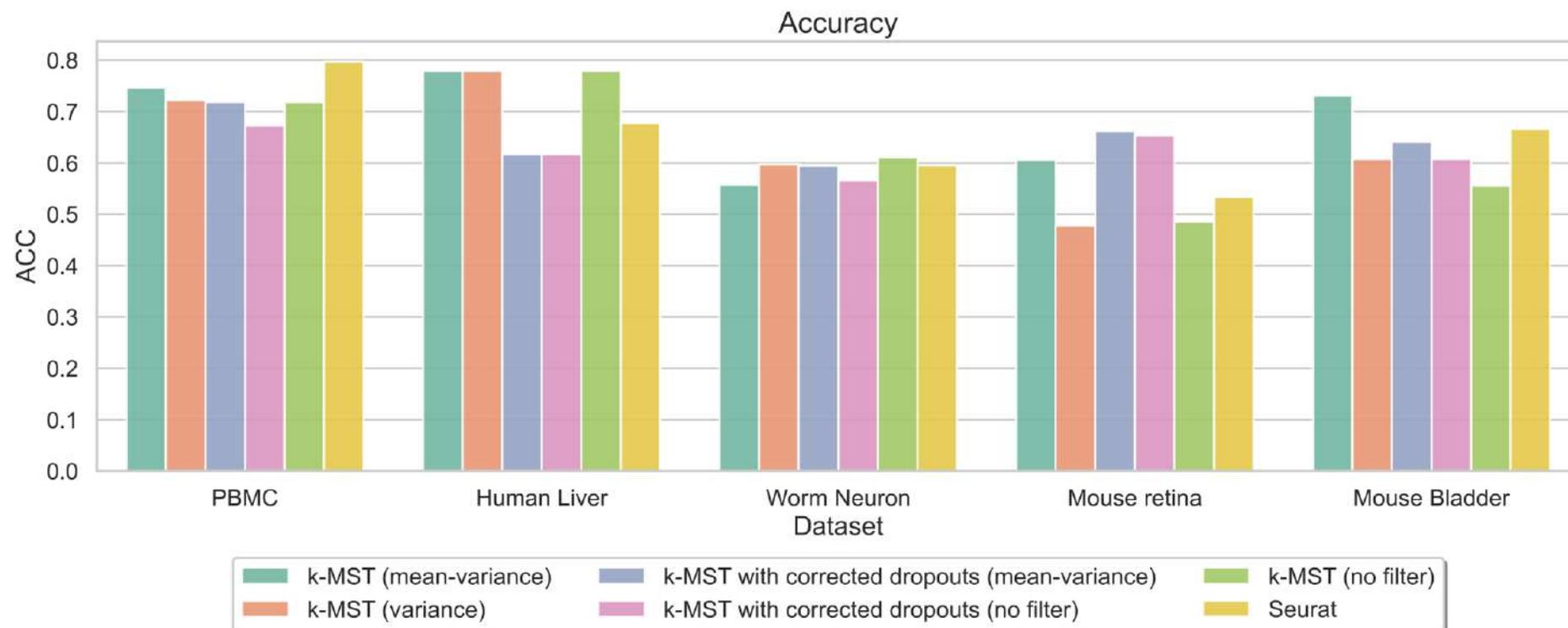
- Con técnicas de corrección de errores basadas en redes neuronales.
- Sin la técnica.

Datos reales

- Se utilizaron cinco datos reales de diferentes especies, tejidos, números de células y número de clusters.
- Los clusters fueron etiquetados por expertos.

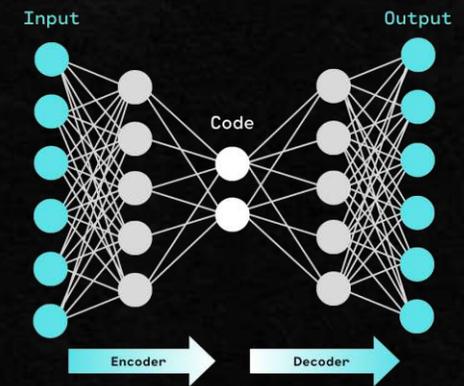
Dataset	Especie	Tejido	Células	Clusters	% Ceros	Secuenciación
PBMC	Humano	Sangre	4271	8	96.08%	10X
Human Liver	Humano	Hígado	8444	11	90.76%	10X
Mouse bladder	Ratón	Vejiga	2746	16	94.86%	Microwell-seq
Mouse retina	Ratón	Retina	14653	39	88.34%	Drop-seq
Worm neuron	Gusano	Cerebro	4186	10	98.61%	Sci-RNA-seq

Resultados metodología basada en grafos



- Se compara la metodología con diferentes procesamientos contra Seurat para cinco datasets reales.
- Las aproximaciones tienen mejores resultados en todos los datasets menos en el de PBMC (sangre)
- El preprocesamiento que muestra mayor estabilidad es la selección de genes con mayor varianza sobre la esperada.

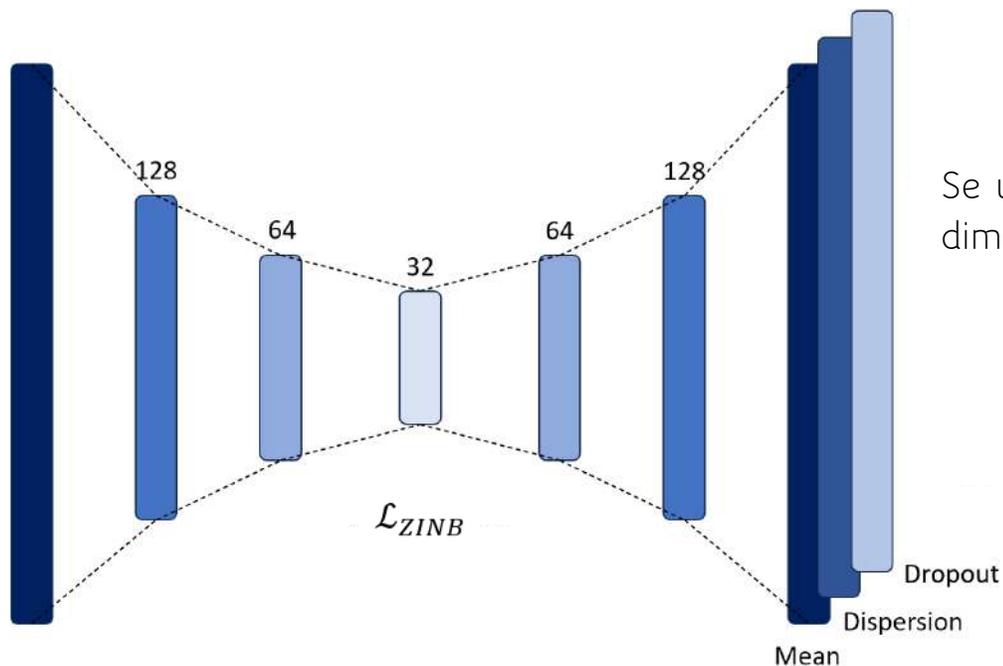
METODOLOGÍAS BASADAS EN REDES NEURONALES



Metodología basada en redes neuronales

Suposición: Conteos de expresión de datos de célula única se modelan como una distribución binomial negativa con inflación de ceros.

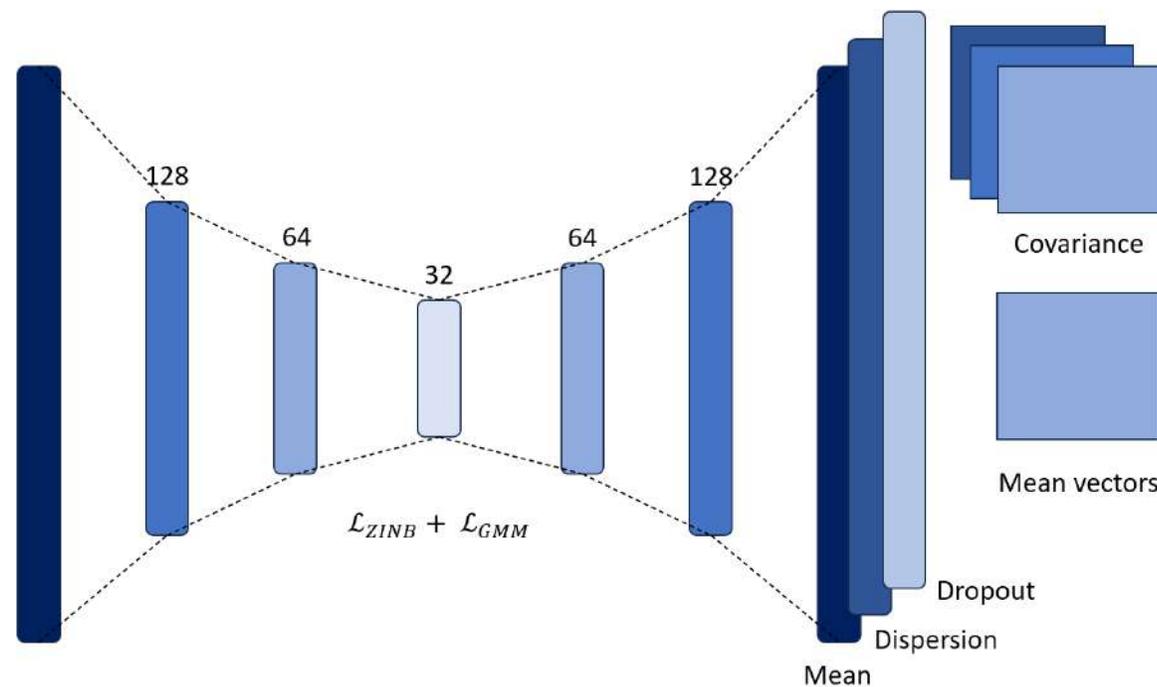
$$ZINB(X = k) = \begin{cases} \pi + (1 - \pi) \cdot NB(k|\mu, \theta) & \text{si } k = 0 \\ (1 - \pi) \cdot NB(k|\mu, \theta) & \text{si } k > 0 \end{cases} .$$



Se utilizan arquitecturas de tipo Autoencoder para reducción de dimensionalidad y corrección de errores en datos de scRNA-seq.

$$\mathcal{L}_{ZINB} = \sum_{ij} -\log(ZINB(X_{ij}|\pi_{ij}, \mu_{ij}, \theta_{ij})) .$$

Metodología basada en redes neuronales



Basándonos en el método scDCC, es posible aprender los parámetros de los clusters durante el entrenamiento de la red.

$$C_k \sim \mathcal{N}(\mu, \Sigma)$$

Se incluye una nueva función de pérdida basada en los Modelos de Mezcla Gaussiana (GMM)

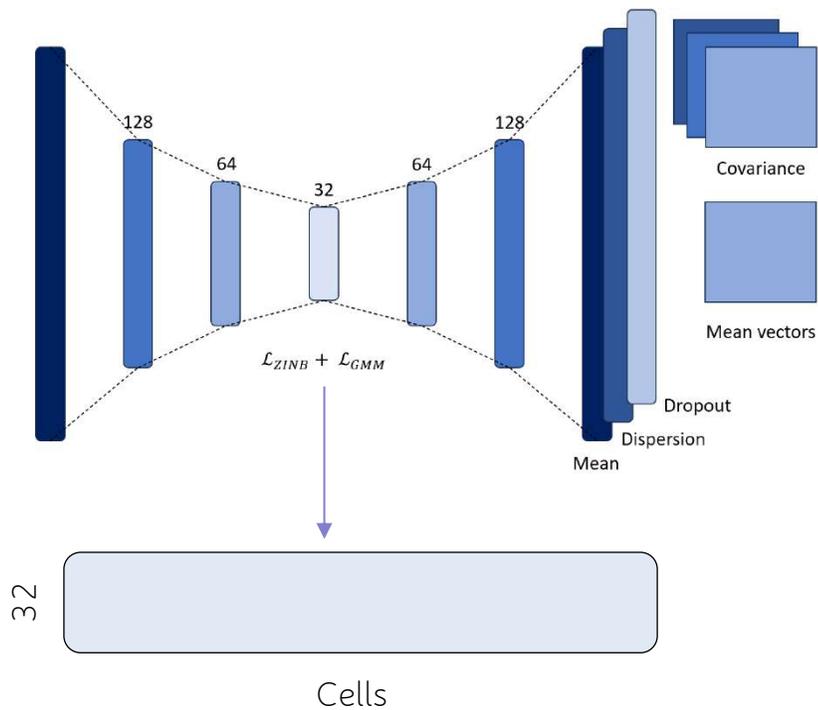
$$\begin{aligned} \mathcal{L}_{GMM} &= \sum_{i=1}^n -\ln p(x_i | \pi, \mu, \Sigma) \\ &= \sum_{i=1}^n -\ln \sum_c \pi_c \cdot \mathcal{N}(x_i, \mu_c, \Sigma_c). \end{aligned}$$

Entrenamiento:

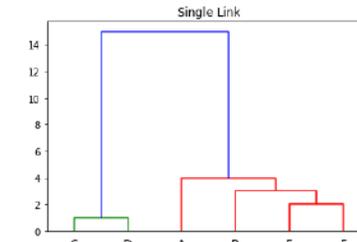
1. Únicamente con \mathcal{L}_{ZINB} para que los datos se ajusten a la distribución.
2. Ajuste con las dos funciones de pérdida para determinar los parámetros de los clusters.

Metodología basada en redes neuronales

Otras aproximaciones



Single Linkage
Método jerárquico



Birch
Método jerárquico para datos en dimensión alta

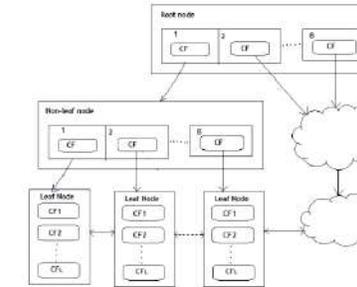
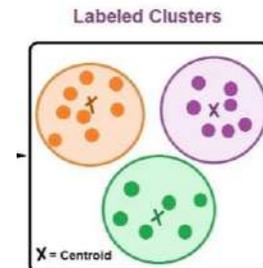
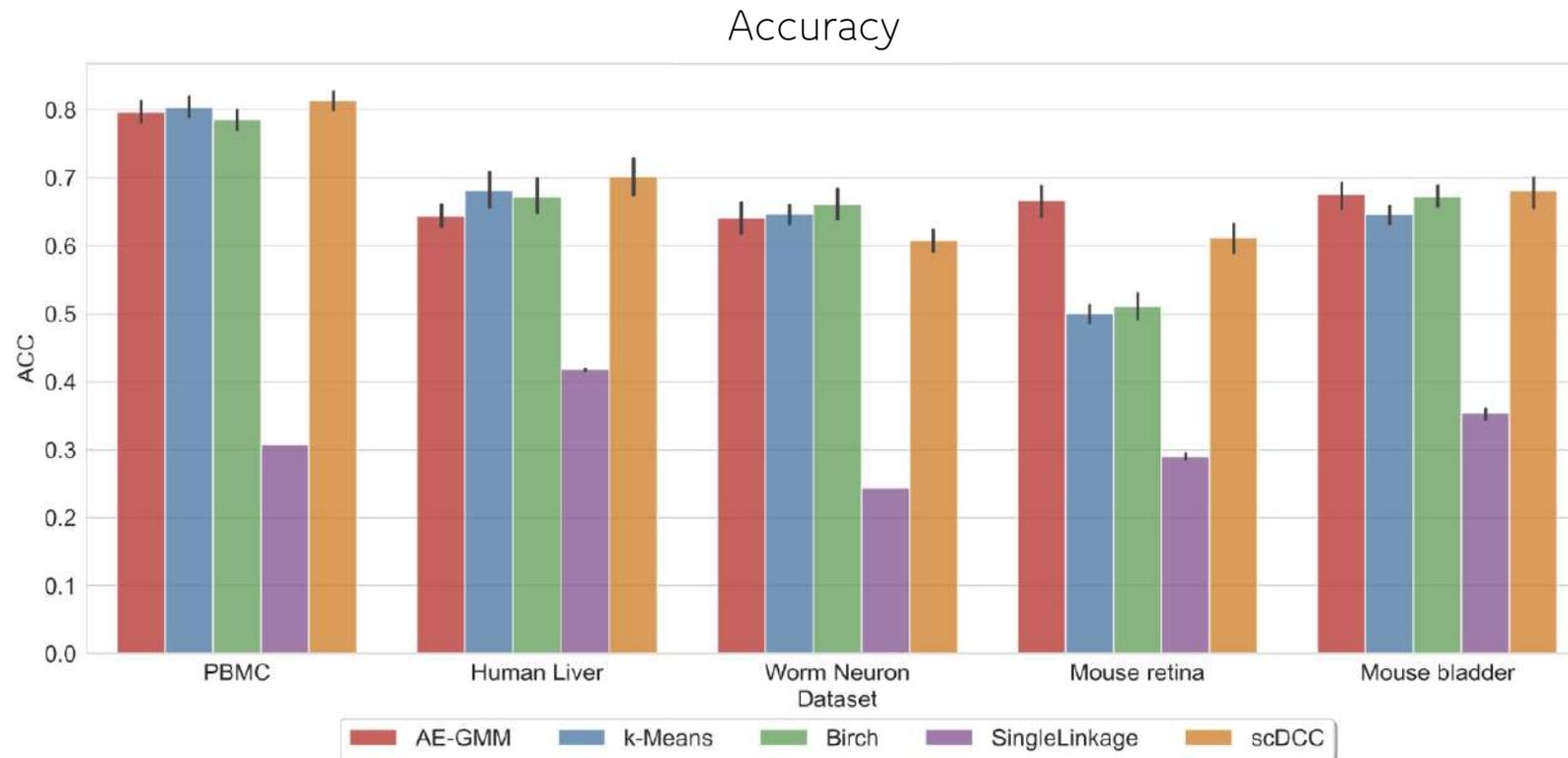


Fig: A CF tree structure

k-Means
Método basado en centroides

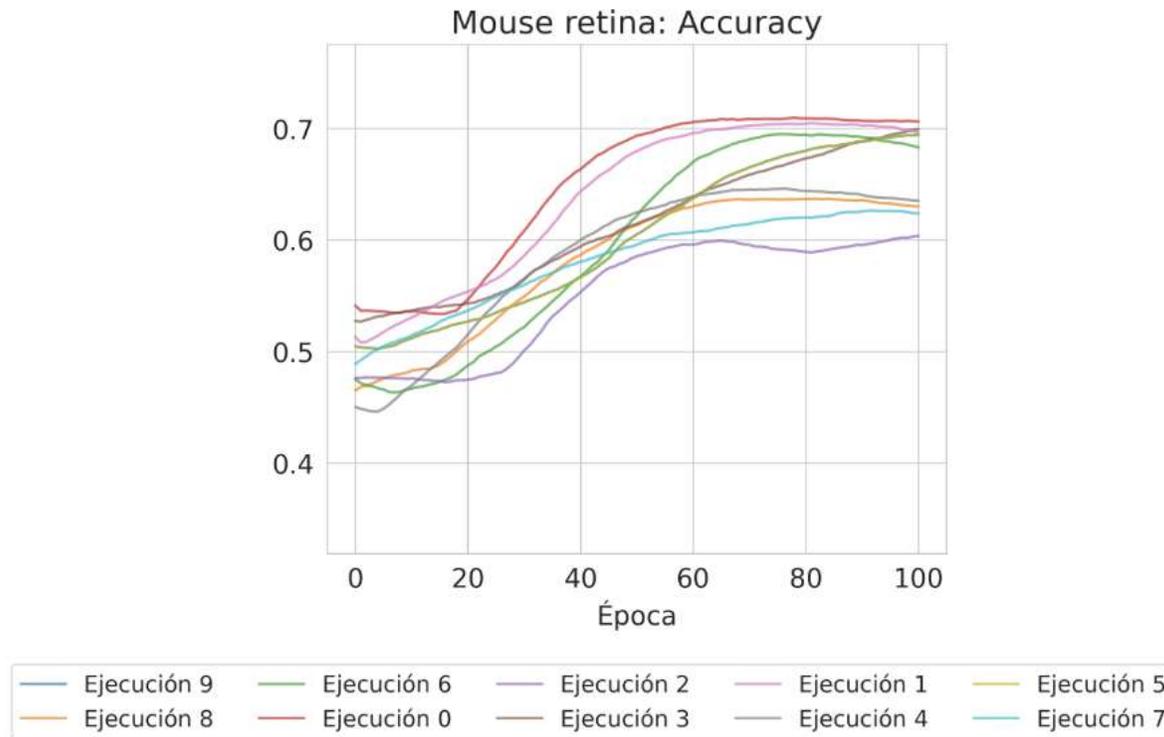


Resultados redes neuronales

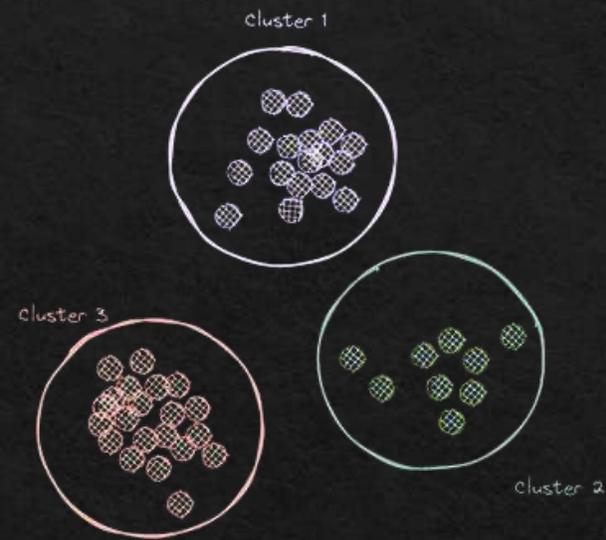


- En experimentos reales se comparan los métodos basados en redes neuronales con scDCC con 10 repeticiones cada ejecución
- Se observa que los resultados de scDCC y AE-GMM son comparables a las aproximaciones que utilizan la red neuronal únicamente como un proceso de reducción de dimensionalidad, seguido de métodos tradicionales.
- El dataset de retina de ratón evidenció una mejora en las métricas a favor del modelo AE-GMM.

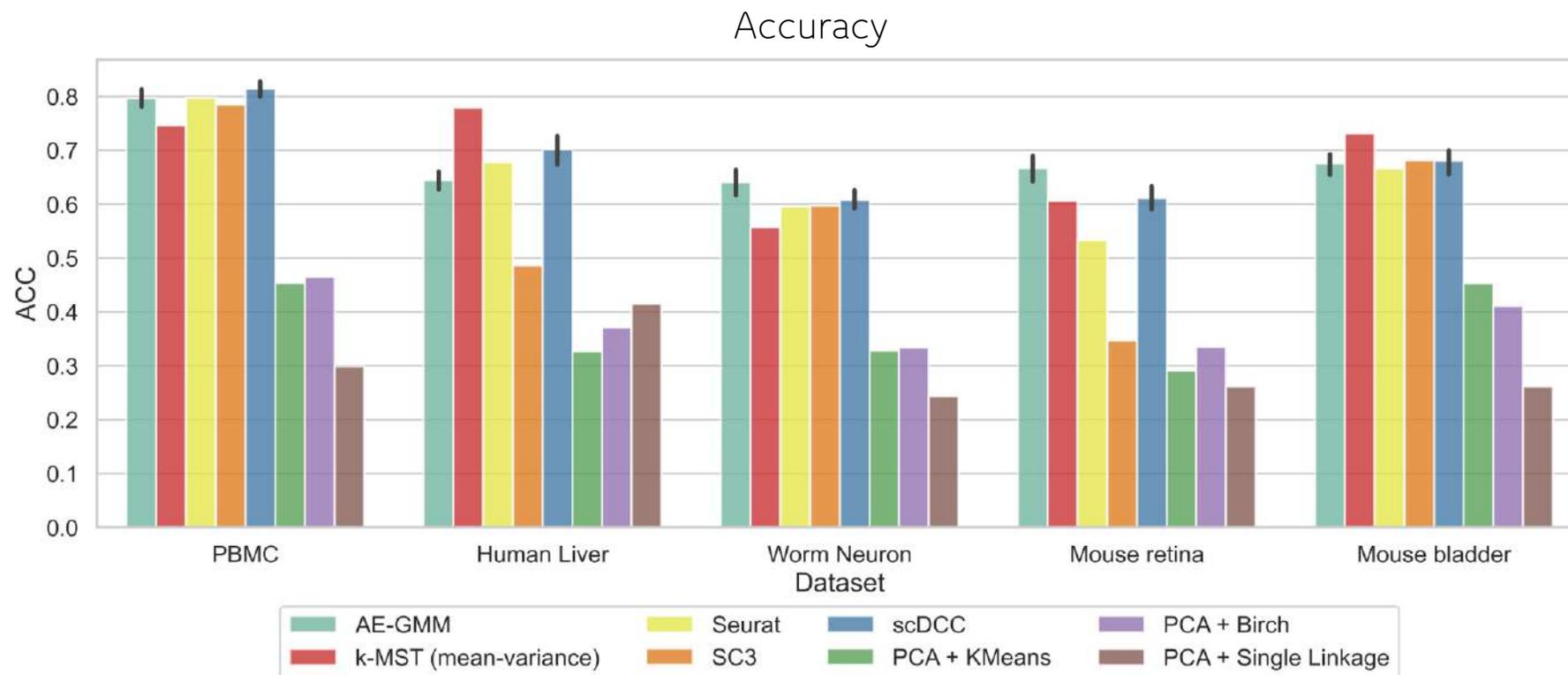
Resultados redes neuronales



COMPARACIÓN MÉTODOS NO SUPERVISADOS

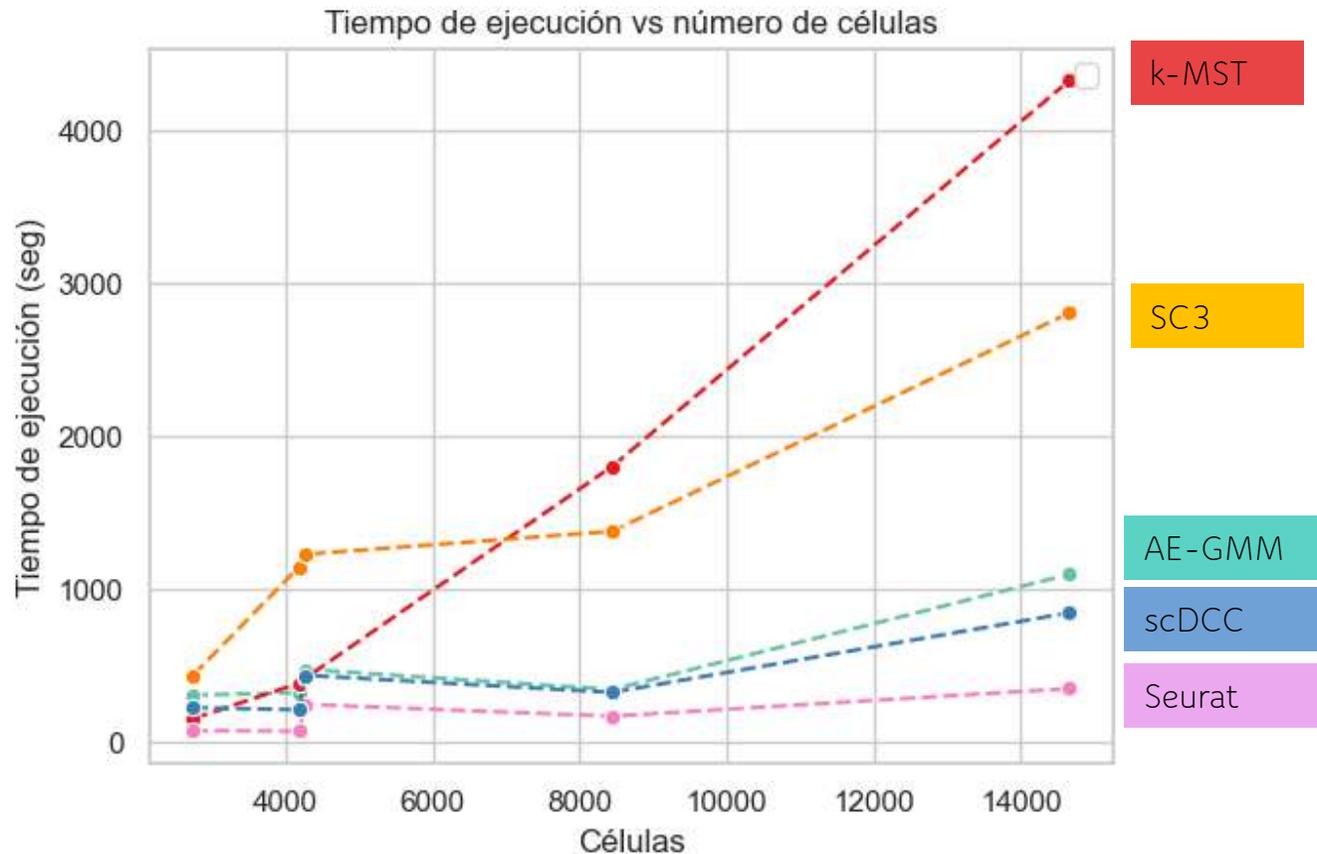


Comparación de modelos no supervisados



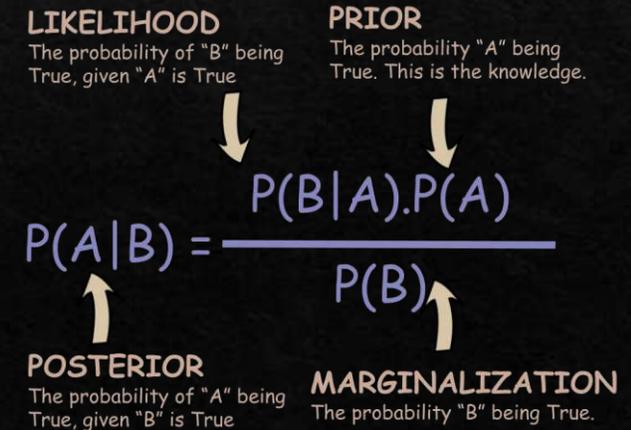
- Comparación de los métodos no supervisados desarrollados con métodos para clustering de scRNA-seq y métodos tradicionales.
- AE-GMM presenta los mejores resultados en Mouse Retina y Worm Neuron y resultados comparables con scDCC en PBMC.
- k-MST presenta los mejores resultados en Human Liver y Mouse Bladder.
- Los métodos tradicionales presentan los peores rendimientos.

Comparación de modelos no supervisados



- Se compara el tiempo de ejecución contra el número de células de los datasets.
- Los métodos basados en redes neuronales se ejecutaron con GPU.
- Para los datasets con **más de 7000 células** (Human Liver y Mouse retina), k-MST es el método que **toma más tiempo** de ejecución.
- Para los datasets **con menos de 7000 células** el que tiene peor rendimiento es SC3.
- Los métodos basados en redes neuronales son similares en cuanto a tiempo de ejecución.
- Seurat es el método que requiere de menos tiempo de ejecución en todos los conjuntos de datos.

METODOLOGÍAS BASADAS EN ESTADÍSTICA BAYESIANA



Metodología basada en estadística bayesiana

Objetivo: Utilizar la base de datos de Human Protein Atlas como datos de entrenamiento para un modelo supervisado.

Human Protein Atlas

genes	Cell type 1 Tissue 1		Cell type M Tissue N			
	representations	representations	representations	representations		
2	...	3	...	0	...	0
0	...	0	...	1	...	1
0	...	0	...	9	...	7
⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	...	3	...	1	...	2
0	...	0	...	0	...	0
0	...	0	...	1	...	1

Entrenar un modelo supervisado de clasificación

Problema: Algunas combinaciones de tejido y tipo celular (>50%) tienen una única representación en la base

Teorema de Bayes:

Se estima la probabilidad de observar el tipo celular H_k dado el perfil de expresión de la célula C_j .

El tipo celular a asignar es aquel que maximice la probabilidad $P(H_k|C_j)$.

$$P(H_k|C_j) = \frac{P(C_j|H_k) \cdot P(H_k)}{P(C_j)}$$

Likelihood: Probabilidad de observar el perfil de expresión en el tipo celular

Prior: utilizar información conocida sobre el tejido secuenciado

Total o marginal: Es constante para el mismo perfil de expresión C_j .

Metodología basada en estadística bayesiana

Prior:

Si N es el conjunto de combinaciones Tipo celular-Tejido de HPA y N_t es el conjunto de combinaciones que corresponden al tejido del experimento, entonces:

$$P(H_k|C_j) = \frac{P(C_j|H_k) \cdot P(H_k)}{P(C_j)}$$

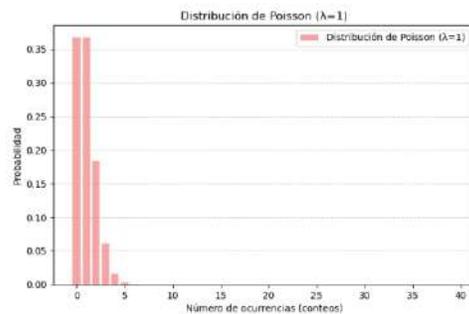
$$P(H_k) = \begin{cases} \frac{0.9}{|N_t|} & \text{si } H_k \in N_t \\ \frac{0.1}{|N|-|N_t|} & \text{si } H_k \notin N_t \end{cases}$$

Likelihood:

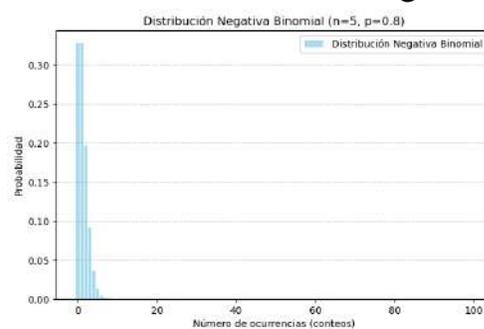
Se supone independencia de los genes. Los parámetros de las distribuciones se calculan a partir del HPA

$$P(C_j|H_k) = \prod_i P(C_{ij}|H_k)$$

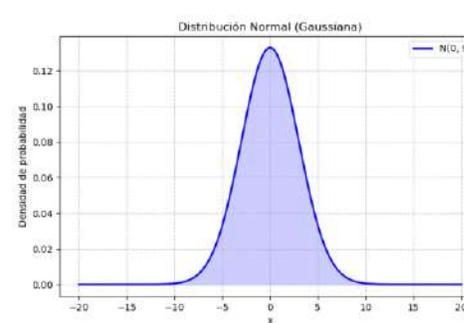
Distribución Poisson



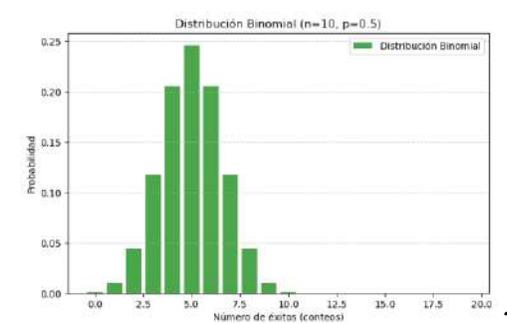
Distribución Binomial Negativa



Distribución Gaussiana



Distribución Multinomial



Metodología basada en estadística bayesiana

Consideraciones:

- Coincidencia de genes en HPA y en la matriz de conteos.
- Problemas numéricos por la magnitud de los valores en HPA.

$$\hat{x}_{ij} = x_{ij} \cdot \frac{L}{\sum_k x_{kj}}$$

Las matrices de conteo del experimento y HPA se normalizan para que tengan el mismo conteo total L. Este valor corresponde a la mediana de los conteos por célula del experimento.

- Imposibilidad de estimar varianzas y dispersiones para tipos celulares con una única representación.

Varianza $\sigma^2 \sim 0.08$

Dispersión $\theta \sim 1.8$

- Uso de coeficiente de Laplace $\alpha = 1$ para evitar problemas de estimación de parámetros por los valores en cero del HPA.

Metodología basada en estadística bayesiana

Input 1

Human Protein Atlas

	Cell type 1 Tissue 1				Cell type M Tissue N			
genes	2	...	3	...	0	...	0	...
	0	...	0	...	1	...	1	...
	0	...	0	...	9	...	7	...
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	4	...	3	...	1	...	2	...
	0	...	0	...	0	...	0	...
	0	...	0	...	1	...	1	...

representations

Input 2

scRNA-seq Experiment

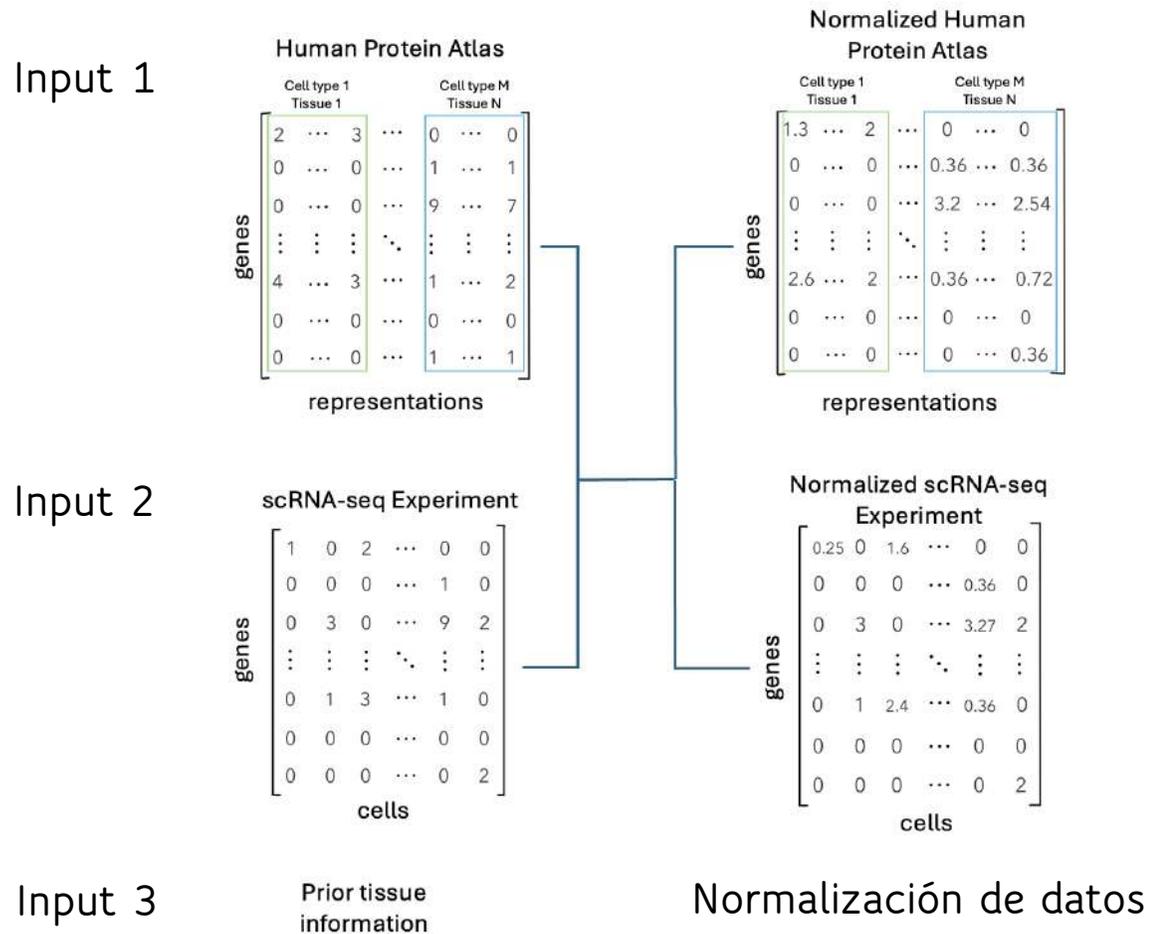
genes	1	0	2	...	0	0
	0	0	0	...	1	0
	0	3	0	...	9	2
	⋮	⋮	⋮	⋮	⋮	⋮
	0	1	3	...	1	0
	0	0	0	...	0	0
	0	0	0	...	0	2

cells

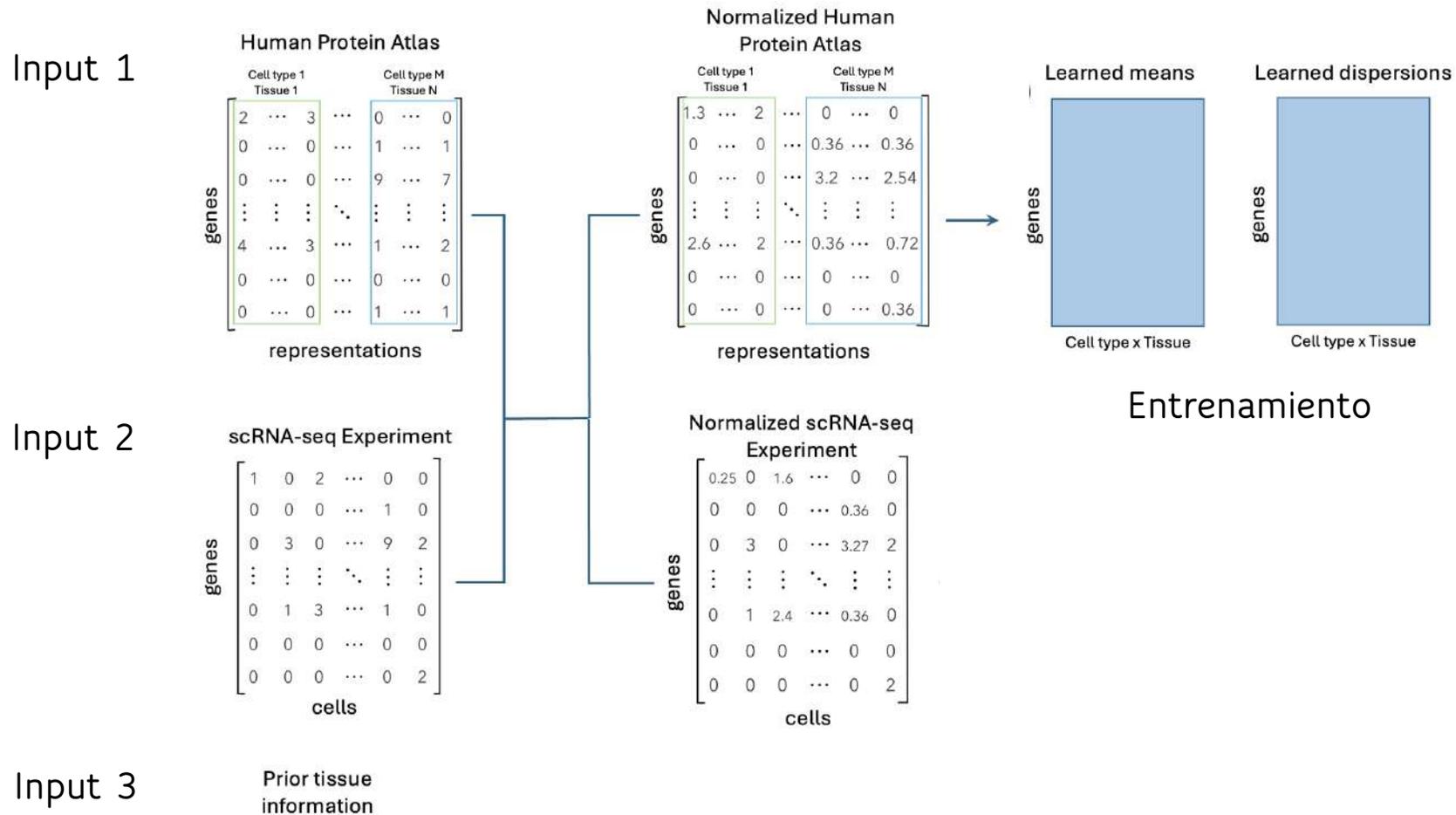
Input 3

Prior tissue
information

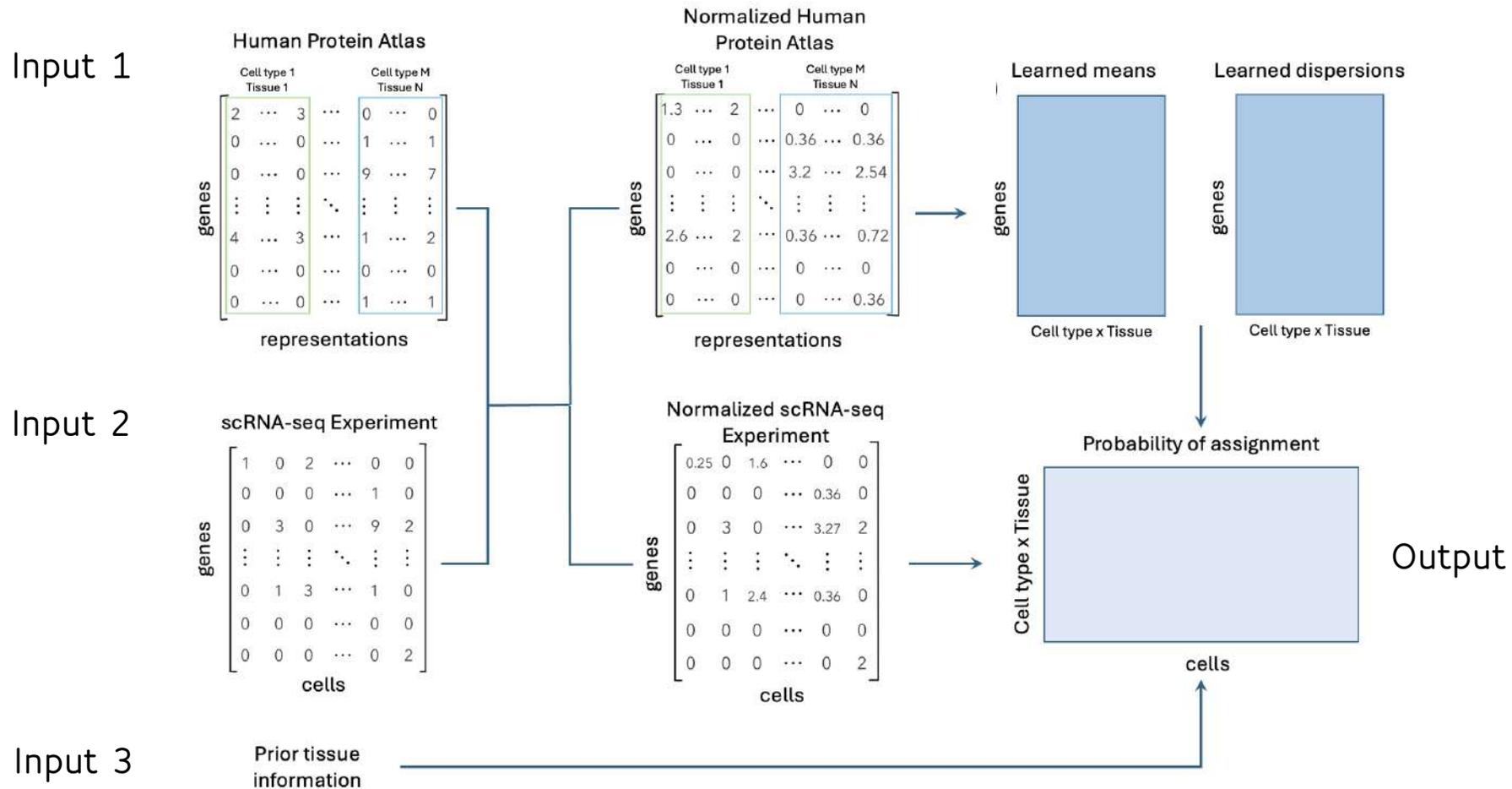
Metodología basada en estadística bayesiana



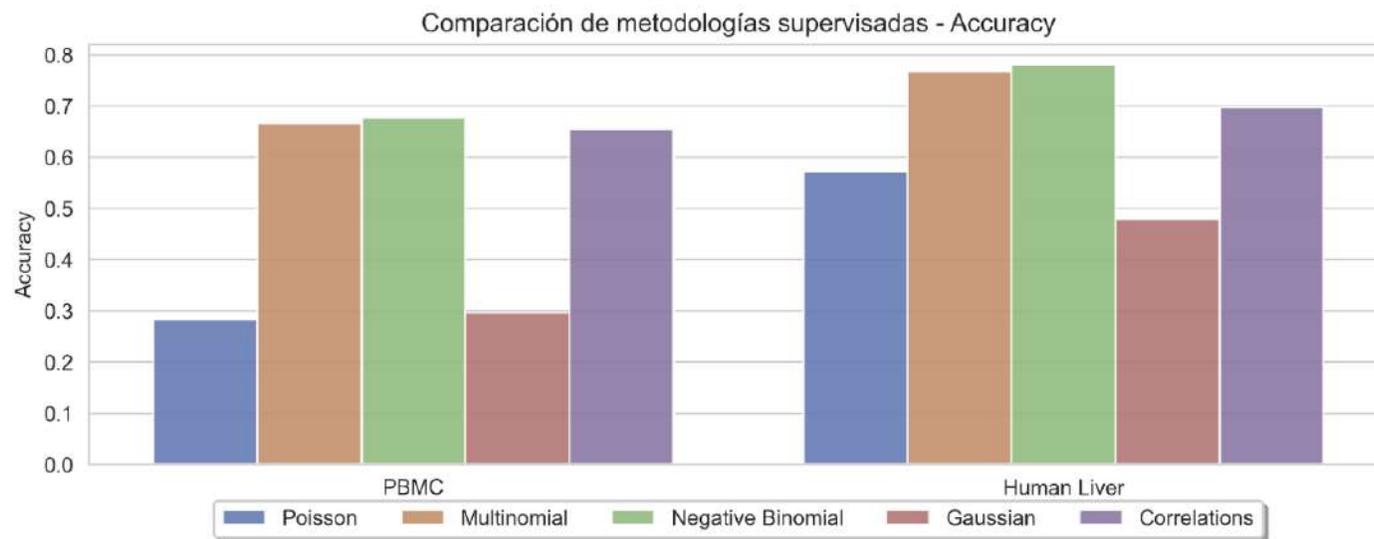
Metodología basada en estadística bayesiana



Metodología basada en estadística bayesiana

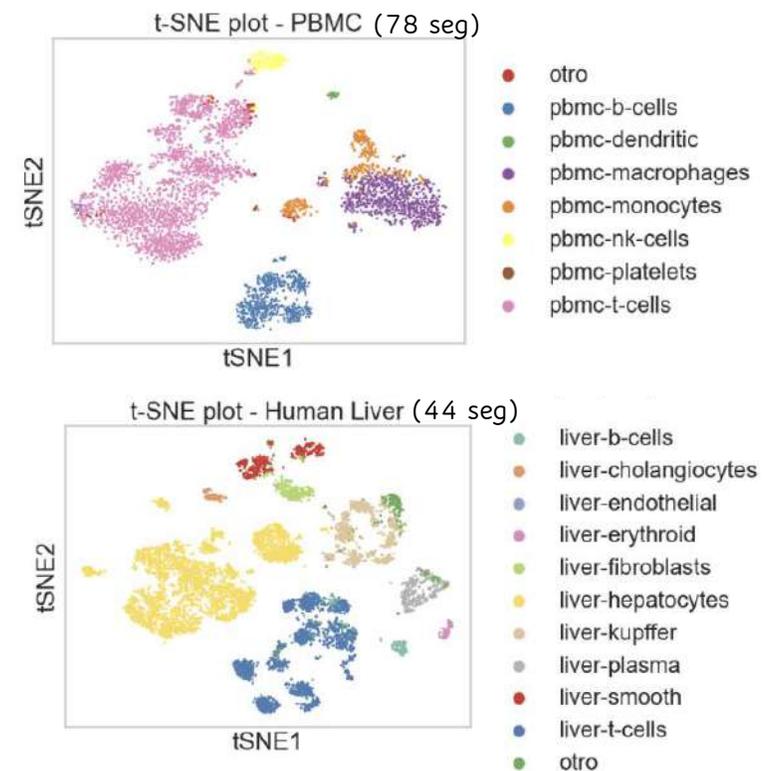


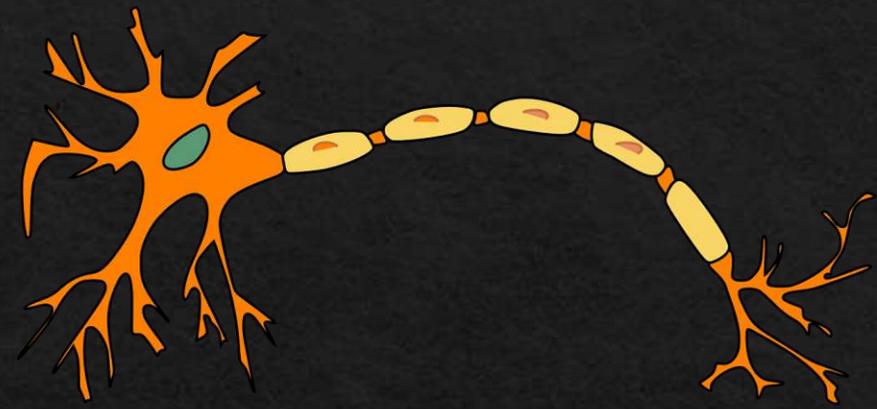
Resultados estadística bayesiana



- Los datasets PBMC y Human Liver son los únicos datos en humanos entre nuestras bases de prueba.
- Se comparan los métodos NB con la asignación directamente por correlación.
- Gaussian y Poisson obtuvieron los peores resultados.
- El que obtuvo mejor resultado en ambas bases fue Negative Binomial.

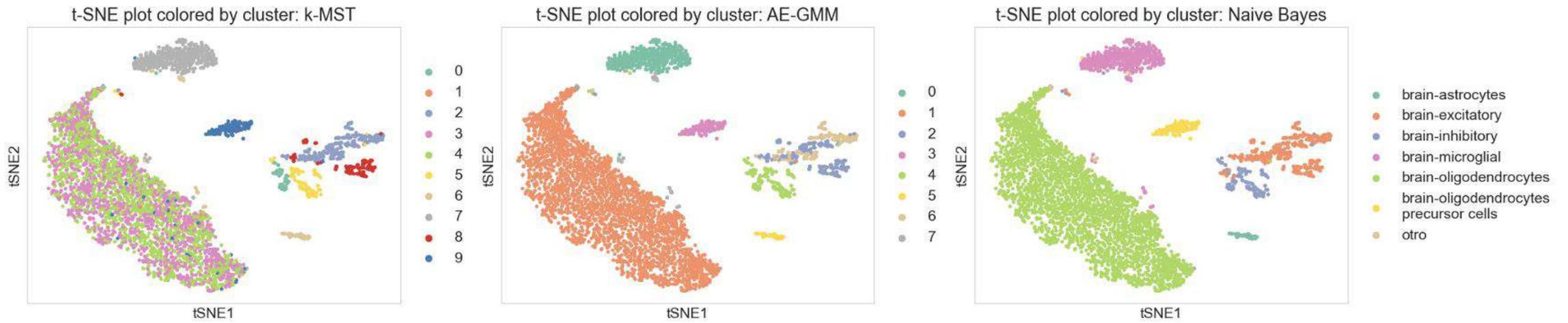
Asignación - Binomial Negativa





COMPARACIÓN DATASET DE EPILEPSIA

Resultados base de epilepsia

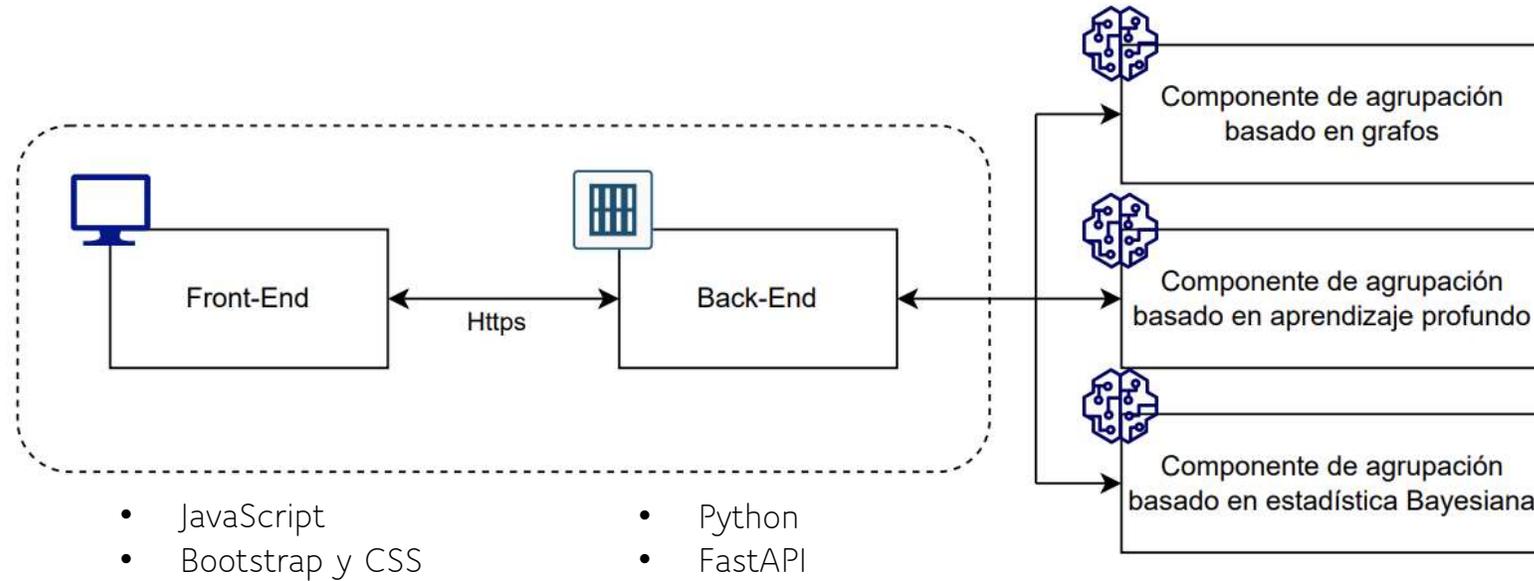


- No es posible obtener métricas supervisadas para estas asignaciones debido a la falta de una referencia previa.
- k-MST presentó los resultados más divergentes, evidenciando un fenómeno de overclustering, particularmente en el cluster más grande, el cual fue subdividido en dos grupos principales.
- Las metodologías AE-GMM y Naive Bayes identificaron correctamente este cluster como el grupo de oligodendrocitos.
- En los clusters más pequeños los métodos con mejores resultados fueron k-MST y Naive Bayes.

DEMO DE LA HERRAMIENTA



Demo de la herramienta



Demo de la herramienta

Agrupamiento de datos de Single Cell

Seleccione los archivos

Cargue los archivos necesarios: barcodes.tsv.gz, features.tsv.gz y matrix.mtx.gz

Choose File	barcodes.tsv	Barcodes
Choose File	genes.tsv	Genes
Choose File	matrix.mtx	Matrix
<input type="button" value="Cargar datos"/>		

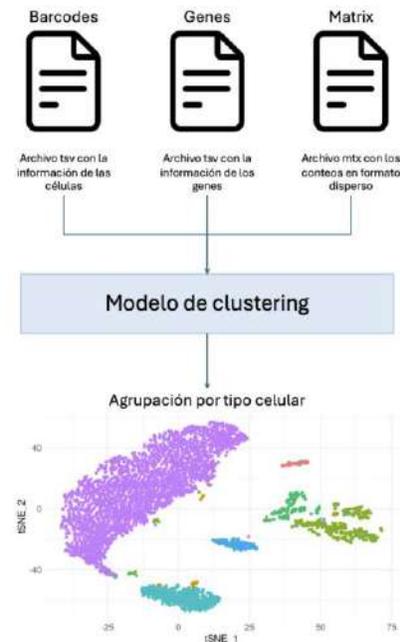
¡Carga exitosa! Se procesó el experimento de scRNA-seq con 33694 células y 4340 genes.

Seleccione la metodología

Basada en grafos

Basada en redes neuronales

Basada en modelos Bayesianos

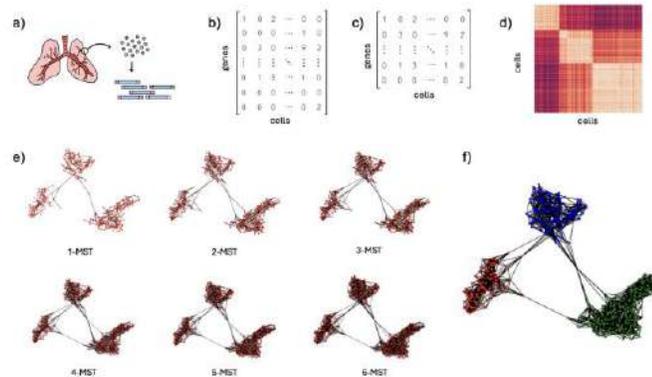


Formatos de entrada unificados

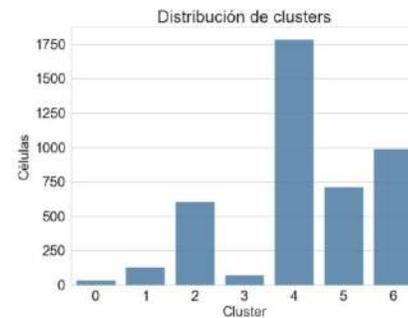
Demo de la herramienta

Aproximación basada en grafos

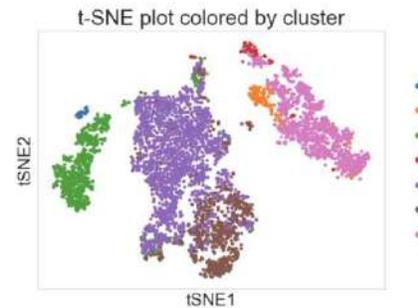
El proceso comienza con la extracción de datos de células individuales (a), seguido por la consolidación de la matriz de conteos a nivel de gen por cada célula (b). Luego, se realiza la normalización de los datos para mantener un conteo homogéneo en todas las células de la muestra y asegurar una comparación justa. Posteriormente, se lleva a cabo el cálculo de correlaciones de Pearson (d) entre las distintas células. Con esta información, se procede a la construcción del grafo (e), donde cada nodo representa una célula y las aristas representan las correlaciones. El grafo resultante es el k-MST del grafo completo de correlaciones. Finalmente, se realiza la detección de comunidades con el algoritmo de Louvain (f) dentro del grafo para identificar grupos de células que comparten características similares.



Correr modelo



Descargar Resultados

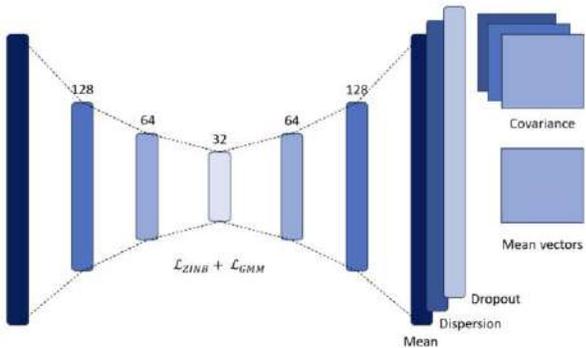


Gráficos de correlaciones y distribuciones de células en clusters detectados.

Demo de la herramienta

Aproximación basada en aprendizaje profundo

La red neuronal se preentrena utilizando la función de pérdida de log-likelihood de la distribución ZINB, comúnmente utilizada para modelar datos de Single Cell. Posteriormente, en la etapa de ajuste, se optimiza la representación latente mediante una pérdida basada en el Modelo de Mezcla Gaussiana. Durante este proceso, la red neuronal aprende los parámetros de cada gaussiana, que incluyen una probabilidad prior π , la media μ y la matriz de covarianza Σ para cada grupo. El resultado que se obtiene es una probabilidad de cada célula de pertenecer a cada uno de los grupos generados. Para la generación de los grupos es necesario conocer el número de clusters a generar.

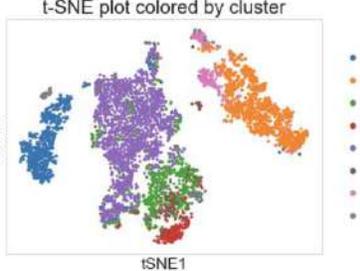
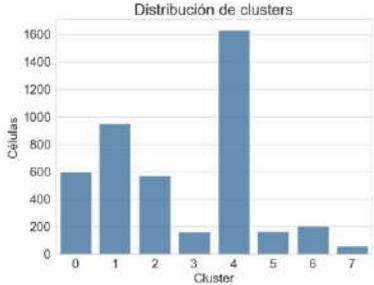
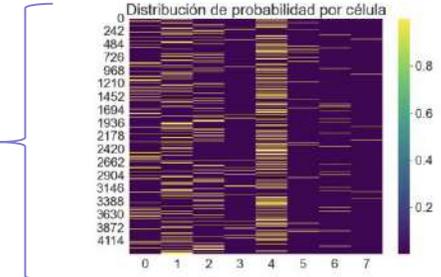


Número de clusters a generar

8

Correr modelo

Gráficos de las distribuciones de probabilidad por célula y distribución de células por cluster

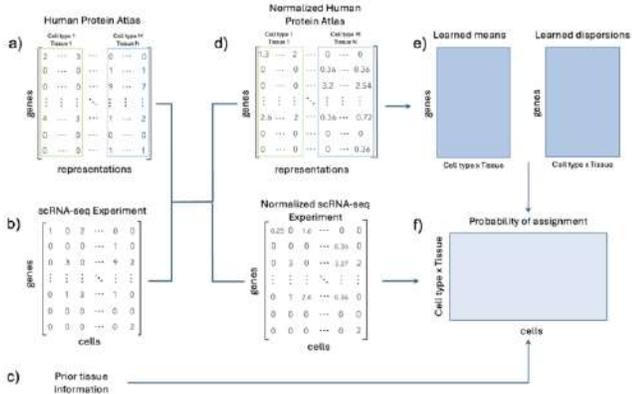


Descargar Resultados

Demo de la herramienta

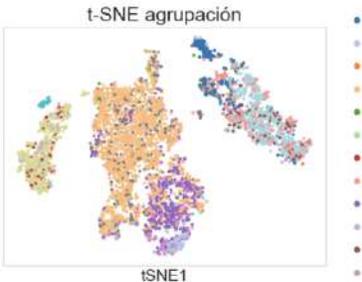
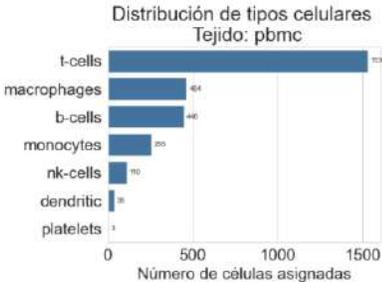
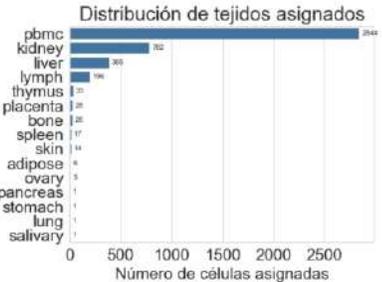
Aproximación basada en estadística Bayesiana

El proceso comienza con la extracción de datos de células individuales (a), seguido por la consolidación de la matriz de conteos a nivel de gen por cada célula (b). Luego, se realiza la normalización de los datos para mantener un conteo homogéneo en todas las células de la muestra y asegurar una comparación justa. Posteriormente, se lleva a cabo el cálculo de correlaciones de Pearson (d) entre las distintas células. Con esta información, se procede a la construcción del grafo (e), donde cada nodo representa una célula y las aristas representan las correlaciones. El grafo resultante es el k-MST del grafo completo de correlaciones. Finalmente, se realiza la detección de comunidades con el algoritmo de Louvain (f) dentro del grafo para identificar grupos de células que comparten características similares.



Selección del tejido secuenciado

Gráficos de la distribución de tipos celulares y tejidos en los resultados



[Descargar Resultados](#)

Discusión

Metodología	Ventajas	Desventajas
k-MST	<ul style="list-style-type: none">• No se utiliza PCA.• Se crea un grafo más denso que el utilizado por Seurat.• No requiere conocer el número de clusters previamente.	<ul style="list-style-type: none">• Puede hacer overclustering
AE-GMM	<ul style="list-style-type: none">• Utiliza la distribución inherente de los datos.• Permite clusters más flexibles (elipsoides).	<ul style="list-style-type: none">• Se debe conocer el número de clusters.
NB – Negative Binomial	<ul style="list-style-type: none">• Se asignan directamente los nombres de los tipos celulares.	<ul style="list-style-type: none">• Funciona únicamente para humanos• Algunas células pueden no coincidir con ninguna clase.

- La complejidad inherente de los datos de scRNA-seq impide que un modelo se destaque consistentemente en todos los conjuntos de datos.
- Los resultados con datos reales podrían estar influenciados por la forma en que se asignan las etiquetas de referencia
- Esto resalta la necesidad de continuar investigando y desarrollando nuevas metodologías de agrupación para el análisis de datos de secuenciación.

¡Gracias!



Referencias de imágenes:

<https://nin.nl/about-the-brain/neuron/>

<https://learn.gencore.bio.nyu.edu/single-cell-rnaseq/>

https://www.flaticon.es/icono-gratis/diagrama-de-flujo_12098213

<https://www.deeplearningnerds.com/deep-learning-how-autoencoders-work/>

<https://www.advancinganalytics.co.uk/blog/2022/6/13/get-started-with-clustering-the-easy-way>

<https://www.infosalus.com/salud-investigacion/noticia-desarrollan-nueva-herramienta-informatica-investigar-complejidad-genoma-20240411105125.html>

