UNIVERSIDAD NACIONAL DE COLOMBIA

Métodos de reducción de dimensión de variables para la clasificación de muestras de datos de expresión en células unitarias

Dimension reduction methods for Single cell Rna-seq data classification

Julián Hernández Castañeda

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá, Colombia
2021

Trabajo de Final

Julián Hernández Castañeda

Trabajo de final para optar al título de:
Magister en estadística

Director(a):
Ph.D. Liliana López Kleine

Línea de Investigación:
Estadística Genómica
Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá, Colombia
2021

Agradecimientos

Quiero agradecer a todas las personas que de manera directa o indirecta apoyaron este proyecto que hoy culmina un etapa importante. Agradezco a mi familia, hermanos, padres y a mi amada Camila que me han acompañado en este proceso con paciencia y cariño.

Quiero agradecerle a la Universidad Nacional de Colombia y a sus profesores por darme las herramientas que hoy guían mi futuro, en especial quiero agradecer al profesor Arun cuya humanidad y acompañamiento fue esencial para iniciar mi vida investigativa. No existiría este trabajo de no haber sido motivado por Arun a seguir mi sueño de la investigación, sus consejos y enseñanzas llegaron en el momento que más lo necesitaba.

Finalmente, quiero agradecer a mi directora Liliana, quien sin duda alguna fue la persona más importante en la construcción de este documento. Su compromiso y vocación fueron esenciales para sentirme cómodo trabajando con ella, su sabiduría y comentarios siempre fueron pertinentes. Gracias por darme la oportunidad y espero que así como me siento contento de entregar este trabajo, Liliana junto con los anteriormente nombrados, también sientan agrado por este documento y lo que significa.

Resumen

El estudio de datos de expresión en células unitarias ha venido creciendo en los últimos años dada su gran utilidad, ya que permite entender el funcionamiento de los sistemas biológicos a nivel molecular. Estos datos son muy extensos en términos informáticos por lo que es importante usar un método de reducción de dimensión adecuado para poder interpretar y visualizar la información. Actualmente, hay varios métodos y algoritmos que realizan esta labor. Sin embargo, carecen de buenos resultados o sustentos teóricos estadísticos fuertes.
Por medio de simulaciones se comparan los métodos más populares, analizando sus fortalezas, debilidades y limitaciones. Se plantea un método de reducción de dimensión basado en un modelo lineal mixto, tratando de capturar toda la información importante para datos de single cell RNA sequencing. Además, se propone una metodología particularmente fácil de implementar, que permite destacar los genes influyentes de un proceso biológico. Esta metodología es implementada en datos de oligodendrogliomas, mostrando 3 vías metabólicas que pueden ayudar a entender la heterogeneidad celular de este tipo de tejido.
Palabras clave: SCseq, envelopes, reducción de dimensión, clasificación de muestras, modelo lineal mixto

Abstract

The study of single cell expression data has been growing in recent years given its great utility since it allows us to understand how the biological systems work in a molecular level. These data are very extensive in computational terms, then it is important to use an adequate dimension reduction method to be able to interpret and visualize the information. Currently, there are several methods and algorithms that perform this work. However, they lack good results or strong statistical theoretical support.
With simulations, it is proposed to compare the most popular methods, analyzing its strengths, weaknesses and limitations. It is proposed a dimension reduction method based on a mixed linear model that aims to capture all the important information of single cell RNA sequencing data. Moreover, it is proposed a particularly easy-to-apply methodology that let the researcher mark the influyent genes in a biological process. This methodology is applied to oligodendrogioma data, showing 3 methabolic pathways that can lead to a better understanding of the celular heterogenity of this Tissue.
Keywords: SCseq, envelopes, dimention reduction, classification of samples, mixed linear model

Contenido

1 Introducción

Los datos en el mundo real, en general, están incluidos en un espacio de alta dimensionalidad. Esto quiere decir que a un mismo individuo se le pueden medir muchas variables diferentes. Históricamente, este número de variables de medición está limitado por la capacidad humana, no obstante, con las mejoras tecnológicas actuales esta capacidad humana se ha ido extendiendo, lo cual implica un aumento en el número de mediciones. Sin ir tan lejos, a un humano promedio se le miden numerosas variables todos los días con el objetivo de predecir su comportamiento: se miden sus datos demográficos, su comportamiento dentro de aplicaciones tecnológicas, sus tiempos para realizar actividades, su ubicación, el número de clicks que hace, dónde los hace, entre muchas otras variables. Sin embargo, que se tengan cientos de variables no implica necesariamente una mejor comprensión del fenómeno, a veces, el tener más variables hace más confuso el fenómeno pues se incrementa el ruido que no aporta nada a la comprensión del fenómeno mismo. Es por eso que a través de herramientas matemáticas y computacionales se vienen creando métodos cuyo objetivo es reducir el tamaño de dimensión, i.e., del número de variables a utilizar[2][7][18], dejándole al investigador únicamente un conjunto de variables más pequeño respecto al inicial que sirven para comprender el fenómeno que se esta estudiando, pues se elimina la redundancia y el ruido, seleccionando únicamente variables informativas.

Este conjunto de variables importantes suele ser llamado Dimensión intrínseca[4]. Cabe aclarar que esta dimensión intrínseca no necesariamente está conformada por variables que pertenecen al conjunto de variables iniciales, sino que usualmente son funciones de estas, por lo que se vuelve más complicada su identificación.

Los datos de alta dimensión que se estudiarán a través de este trabajo provienen del área de la transcriptómica. La transcriptómica es un área de investigación cuyos objetivos son: 1. Catalogar la transcripción de los genes de todas las especies y 2. Cuantificar o medir el cambio de los niveles de expresión de cada transcripción durante el desarrollo de una enfermedad o bajo diferentes condiciones[15]. Esta medición de los genes se hace por medio de la cuantificación de RNA mensajero bajo las diferentes condiciones.

El presente trabajo se centrará en el segundo objetivo de la transcriptómica mencionado en el párrafo anterior, es decir, en el uso de datos de transcripción obtenidos a partir de células (Secuenciación de de RNA en células unitarias en ingles Single cell RNA sequencing) para clasificar muestras y así entender la heterogeneidad celular y procesos moleculares como por ejemplo, la progresión de enfermedades[17].

La secuenciación de RNA en células es una de las tecnológicas más prometedoras para la

transcriptómica de células. El número de artículos, investigaciones y bases de este tipo de datos viene en crecimiento exponencial dada su importancia. Su fácil obtención gracias a las mejoras informáticas y tecnologías que se lograron en las últimas décadas, permite que se puedan secuenciar cualquier subconjunto de cualquier genoma con alta precisión y relativo bajo costo [15].

Por poner un ejemplo, el genoma humano tiene al rededor de 3200 millones de pares de bases [1], por lo que un pequeño pedazo del genoma secuenciado puede contener miles de bases ,i.e., miles de datos en términos informáticos. Teniendo en cuenta que la recolección de estos datos viene creciendo, los datos genómicos se han convertido en uno de los dominios de big data más grandes en la actualidad [17], por lo cual, es de suma importancia encontrar métodos adecuados para su tratamiento y análisis.

Como se había mencionado antes, se puede buscar una dimensión intrínseca y afortunadamente la evidencia parece indicar que esta existe,es decir, que con menos de la totalidad de los genes cuya transcripción es cuantificada, sería posible caracterizar la muestra y por ende la condición a la que pertenece ese individuo. Adicionalmente, esta dimensión intrínseca es mucho más pequeña a la dimensión original lo que permite que los datos de secuenciación se puedan visualizar en 2 o 3 dimensiones [4]. Es decir, a pesar de que se obtiene la transcripción de RNA de inclusive miles de genes y se conocen distintas características de su procedencia, existen muchos genes o características que son invariantes o innecesarios para la comprensión del sistema biológico, de una enfermedad y su progresión.

Si bien existen métodos para lograr la reducción de dimensión, los algoritmos y métodos actualmente usados, mencionados en los antecedentes y desarrollados brevemente en el marco teórico, tienen grandes limitaciones. Por ejemplo, el análisis de componentes principales no funciona bien dado que no mantiene las estructuras de baja dimensión de los datos, esto causa que no se haga una buena clasificación de los datos y consiguientemente no se logre comprender la progresión de la enfermedad; el t-SNE no cuantifica si se hizo una buena o mala reducción de dimensión [4] por lo que son inciertos los resultados, no se puede generalizar y es posible que las conclusiones que se saquen a partir de esta información sean inválidas. Adicionalmente, la mayoría de los métodos propuestos para realizar esta labor en los últimos años, son redes neuronales similares al T-SNE, cuya limitación principal, como se mencionaba, es que carecen de sustento teórico para confiar en sus resultados pues no hay forma de conocer si la reducción fue o no efectiva.

Para dar una respuesta teórica al naciente problema de la reducción de dimensión de datos de expresión genómica, se proponen dos métodos, el primero es usar la teoría de los envelopes desarrollada por Denis Cook, la cual proporciona una fuerte herramienta para la reducción de dimensión en problemas con datos abundantes y dimensión intrínseca pequeña[2] y la segunda es un modelo de reducción de dimensión sencillo basada en un modelo lineal mixto. Con estos dos métodos se busca solucionar el problema de la incertidumbre acerca de la buena o mala reducción de dimensión, permitiendo encontrar procesos que identifiquen el tamaño de la dimensión intrínseca junto con valores p que sustenten la toma de decisiones o

inclusive facilitando a través del modelo reconstruir las dimensiones intrínsecas con pequeños subconjuntos de genes. Esto quiere decir que de encontrar buenos resultados y una descripción de la progresión de una enfermedad, los resultados van a ser confiables y fácilmente replicables tanto en el ámbito de la estadística genómica como en otras áreas.
Ahora, como la efectividad del método se basa en su correcta clasificación de los datos, que a su vez está determinado por la correcta identificación de la dimensión intrínseca es necesario corregir el error recurrente en el que caen algunos métodos actuales. El abuso del supuesto de muestra aleatoria. A pesar que usualmente en simulaciones se crean datos con diferentes dispersiones, estas estructuras de varianza no son tenidas en cuenta a la hora de correr los algoritmos; por lo cual, difícilmente estas estructuras de baja dimensión van a ser identificadas. De aquí nace la propuesta de un método adicional.
Por lo dicho anteriormente, se pretende corregir este error, analizando los datos de secuenciación por medio de una reducción de dimensión basada en un modelo lineal mixto, es decir, incluyendo una estructura de varianza y covarianza a las observaciones. Es importante evaluar qué factores deben ser determinados como aleatorios y qué factores deben ser determinados fijos, pues como se menciona en el marco teórico, los factores aleatorios son los que alteran la estructura de covarianzas ergo su correcta identificación son un paso importante para el éxito de la reducción de dimensión.
Posteriormente, se analiza el rendimiento de los envelopes respecto a los métodos revisados en el marco teórico a través de indicadores, en este caso lo importante es la correcta clasificación de los datos, por lo cual se hace necesaria una simulación y los resultados deben ser medidos bajo ese criterio. Se deben probar diferentes escenarios dado el desconocimiento de cómo en realidad son las relaciones entre las variables, por lo que en la metodología se proponen dos simulaciones, cada una con tres escenarios, uno donde la dimensión intrínseca es producto de relaciones lineales, un escenario donde la dimensión intrínseca es producto de relaciones no lineales y un escenario mixto. Las dos simulaciones se diferencian en la inclusión de un factor aleatorio.
Una vez es determinado su desempeño, se tendrán en cuenta datos reales que ya han sido analizados con otros métodos, en específico, datos de oligodendriogliomas reportados en la metodología de este trabajo.
Con la información aportada por la simulación y los resultados de los datos reales, se concluye acerca de la pertinencia de usar los envelopes y el modelo propuesto, abriendo así la posibilidad de hablar acerca del camino a seguir para el análisis de estos datos desde el punto de vista teórico, pues ya existen textos [7] que extienden los envelopes permitiendo reducciones de dimensión no lineales.

2 Marco teórico

En este capítulo se presentarás brevemente resúmenes de los dos métodos más utilizados para reducir dimensiones, así como de la teoría que soporta el método propuesto. Adicionalmente, se presentan los datos con los cuales se propone trabajar.

2.1. Métodos de reducción de dimensión

2.1.1. Análisis de componentes principales

El análisis de componentes principales es el método más utilizado a nivel mundial para reducir dimensiones. No solo por su fácil interpretación e implementación, sino también por su soporte teórico que a su vez no requiere conocimientos de matemáticas avanzadas.
Para entender cómo funciona este método se sigue el acercamiento geométrico enunciado en [12]. Suponga que tiene una muestra de n vectores de tamaño p $y_1, \ldots, y_n$ cuyas p componentes están correlacionadas (variables de alta dimensión), se pueden encontrar combinaciones lineales $z_i = a_i^T y$, $z_j = a_j^T y$ de tal manera que la proporción de varianza en z_i, z_j sea máxima y z_i sea ortogonal a z_j.
z_i, z_j son llamadas componentes principales y son lo que en este trabajo se menciona como variables de baja dimensión.
Lo que buscan las componentes principales, es hacer una rotación de los ejes de tal manera que el eje o componente principal explique la mayor cantidad de varianza, permitiendo así que la variabilidad representada en otros componentes sea despreciable. Adicionalmente, se usa la restricción de que las componentes sean ortogonales lo que permite un mejor entendimiento de los datos y permite repartir la varianza explicada por las componentes evitando la redundancia en la información.
Usualmente, la decisión de cuántas componentes principales dejar, la toma el investigador con base a la varianza explicada, lo cual está relacionado directamente con los autovalores de la matriz de covarianza de los datos.

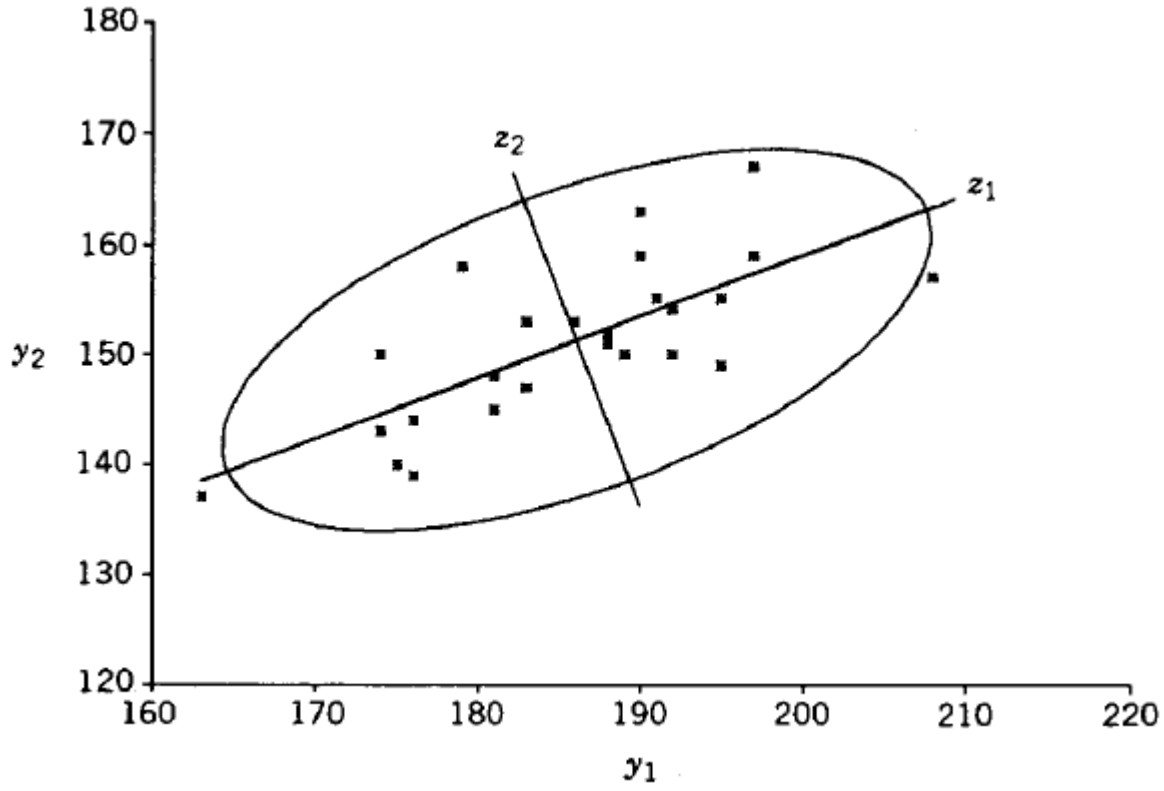

Figura 2-1: Representación de primeras dos componentes principales para un caso bivariado [12]

En la figura **2-1** se puede observar la rotación de los ejes para un caso bivariado. Es claro que la primera componente principal z_1 es en donde más varían las observaciones y z_2 es ortogonal a z_1; mientras que y_1 y y_2 son variables correlacionadas. Si la variación en z_2 fuera pequeña, su análisis se podría omitir.
La gran ventaja de utilizar el análisis de componentes principales es que no se hace ningún supuesto distribucional, sin embargo, necesita que las variables tengan relaciones lineales. Su otra desventaja, es que busca explicar la varianza global, omitiendo información importante de variables regresoras.
Finalmente su ventaja mas importante es que se puede demostrar que la solución del problema de reducción de dimensión se puede obtener a través de los vectores propios de la matriz de covarianza o correlación de los datos y la cantidad de varianza explicada se determina a través de los autovalores correspondientes.

2.1.2. T-distributed stochastic neighbor embedding (T-SNE)

A pesar que el análisis de componentes principales, ACP o PCA en inglés, es el método de reducción de dimensión más usado, el T-SNE viene cogiendo mucha fuerza en la última

década debido a su gran utilidad en diferentes áreas de análisis de datos.
Para entender de qué trata este algoritmo se sigue lo planteado en [8]. T-SNE surge como alternativa no lineal del PCA y diferentes métodos de reducción de dimensión que no tienen buenos resultados. T-SNE es un método no lineal que permite preservar las estructuras globales de los datos, es decir, si dos puntos en altas dimensiones son cercanos, en baja dimensión también lo serán.
El procedimiento para reducir la dimensión comienza convirtiendo las distancias entre puntos en probabilidades condicionales como sigue:

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-||x_i - x_k||^2/2\sigma_i^2)} \tag{2-1}$$

Siendo x_i la i-ésima observación del proceso en alta dimensión.
Esto preserva las distancias pues $p_{i|j}$ es alto cuando x_i, x_j son cercanos. Para las observaciones en baja dimensión (y_i) se procede de manera similar.

$$q_{j|i} = \frac{\exp(-||y_i - y_j||^2)}{\sum_{k\neq i}\exp(-||y_i - y_k||^2)} \tag{2-2}$$

Si las similitudes son bien modeladas $p_{i|j} = q_{i|j}$. Como $q_{i|j}$ es desconocido, el problema se reduce a encontrar $q_{i|j}$ que represente de la mejor manera a $p_{i|j}$, para lo cual se utiliza como criterio para minimizar la diferencia entre las probabilidades en alta y baja dimensión, la divergencia de Kullback-Liebler

$$C = \sum_i KL(p_i||q_j) = \sum_{ij} p_{j|i}\log(\frac{p_{j|i}}{q_{j|i}}) \tag{2-3}$$

C es llamada la función de costo, y a través de ella es posible estimar los vectores de baja dimensión minimizando el costo.
Para el método T-SNE es necesario definir $p_{ij} = \frac{p_{i|j}+p_{j|i}}{2}$, $q_{ij} = \frac{q_{i|j}+q_{j|i}}{2}$ y se vuelve a definir la función de costos usando la misma lógica.

$$C = \sum_i KL(p_i||q_j) = \sum_{ij} p_{ij}\log(\frac{p_{ij}}{q_{ij}}) \tag{2-4}$$

Finalmente, se encuentra la solución de manera iterativa usando $y_1, \ldots, y_n$ una muestra normal con media cero y varianza 10^{-4}.
Sus principales ventajas son que en simulaciones ha funcionado relativamente bien, es intuitivo y tiende a preservar las distancias entre los puntos, sin embargo, no es posible determinar qué tan bueno fue su rendimiento pues no hay ninguna medida arrojada por el método que indique el éxito que tuvo, por ejemplo para el PCA, los autovalores de la matriz de covarianza indican que tanta varianza (información) se esta preservando, otros métodos proporcionan inclusive medidas de verosimilitud de la reducción de dimensión, etc. Por último, en el T-SNE no se garantiza que el algoritmo converja a un óptimo global y es sensible a la "maldición de la alta dimensionalidad".

2.1.3. Reducción de dimensión suficiente

Autores como [4] aseguran que el T-SNE no conserva bien las estructuras de baja dimensión, al igual que resaltan que el hecho de no tener una cuantificación del éxito del procedimiento. Por lo tanto es necesario buscar una alternativa teórica para reducir la dimensión, lo que significa entrar a mirar la reducción de dimensión suficiente.
La reducción de dimensión suficiente proviene de la idea de crear combinaciones lineales de las variables (así como en componentes principales) y la idea de suficiencia que formalizó Fisher en 1922, la cual decía que para unos datos D un estadístico t es suficiente para θ, si $D|(\theta, t) \sim D|(t)$. Es decir, que toda la información posible de θ esta condensada en el estadístico t. Siguiendo el desarrollo realizado en [2] es necesario definir el subespacio central, para así encontrar la reducción de dimensión suficiente.

El subespacio central

Para empezar, es necesario traer a colación la idea de suficiencia de Fisher, pero en el contexto que es de interés. Sea Y una matriz que representa las variables respuesta y X su matriz diseño correspondiente.

> **Definición 1:** Una proyección $P_S : \mathbb{R}^p \to S \subset \mathbb{R}^p$ en un subespacio S q-dimensional es una reducción suficiente lineal si satisface alguna de las siguientes tres condiciones:
>
> 1. Reducción inversa, $X|(Y, P_S X) \sim X|P_S X$
> 2. Reducción hacia adelante, $Y|X \sim Y|P_S X$
> 3. Reducción conjunta, $Y \perp\!\!\!\perp X|P_S X$
>
> Entonces, el subespacio S es llamado subespacio de reducción de dimensión.

Esta definición describe lo que se quiere realizar. Se busca un subespacio donde esté incluida toda la información de X. Sin embargo, si $S \subset S_1$, S_1 es también un espacio de reducción de dimensión, por lo que si existe S, existen infinitos subespacios de reducción de dimensión.
El objetivo de esta metodología es reducir la dimensión. Por lo tanto, se busca el más pequeño de los espacios de reducción de dimensión.

> **Definición 2:** La intersección de todos los subespacios de reducción de dimensión, cuando a su vez también es espacio de reducción de dimensión, es llamado el subespacio central y notado como $S_{Y|X}$

El subespacio central no siempre existe, pero sí bajo ciertas condiciones. Así mismo, en las últimas décadas se ha estudiado la forma de estimar este subespacio central bajo ciertas condiciones, siendo las más frecuentes linealidad y homocedasticidad.

2.1.4. Envelopes

De acuerdo con [2], la metodología de envelopes se puede ver como una forma especializada de SDR (Sufficient Dimension Reduction) que es aplicable en análisis basados en el modelo, que puede ser utilizado a su vez para mejorar la eficiencia de SDR sin modelo.
Aquí nuevamente se tiene que hablar de un modelo lineal que en general tiene la forma.

$$Y = \alpha + \beta X + e \tag{2-5}$$

Con Y la matriz de respuestas, X la matriz diseño, β la matriz de efectos y e una matriz aleatoria generalmente con distribución normal.
Lo que se busca con los envelopes en este caso, es encontrar combinaciones lineales de Y que son invariantes en X. Si estas existen, se puede reducir la dimensión del modelo.
El modelo envelopes es una reparametrización del modelo multivariado (2-5) en términos del subespacio $\xi \subset \mathbb{R}^r$ y Q la proyección en este subespacio con las propiedades que:

$$(i) Q_\xi Y|(X = x_1) \sim Q_\xi Y|(X = x_2) \text{ para todo } x_1, x_2$$

$$(ii) P_\xi Y \perp\!\!\!\perp Q_\xi Y|X \tag{2-6}$$

La primera condición dice que la distribución de $Q_\xi Y$ no depende de quién sea la observación x, y la segunda garantiza que las variables $P_\xi Y, Q_\xi Y|X$ sean independientes. Siendo $P_\xi = I - Q_\xi$. Lo cual es bastante similar a las condiciones para el subespacio central

Definición 3: Un subespacio $R \subset \mathbb{R}^r$ se dice que es un espacio reductor de $M \in S^{r\times r}$ si R descompone a M como $M = P_R M P_R + Q_R M Q_R$. Si R es un espacio reductor de M, se dice que R reduce a M.

Definición 4: Sea $M \in S^{r\times r}$ y sea $B = span(M)$. Entonces el M-envelope de B, denotado como $\xi_M(B)$ es la intersección de todos los subespacios de M que contienen a B

En términos prácticos el interés se centra en hallar $\xi_\Sigma(B)$ siendo $B = span(\beta)$ y de esta manera se reparametriza el modelo con $span(\Gamma) = \xi_\Sigma(B)$ y $span(\Gamma_0) = \xi_\Sigma^\perp(B)$.
$Y = \alpha + \Gamma\eta X + e$

Con $\Sigma = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T$ y $\beta = \Gamma\eta$

Para hallar de forma gráfica el objetivo de los envelopes, se puede observar la figura **2-2** dónde se encuentran los datos del engorde de diferentes vacas con dos dietas en dos tiempos

diferentes (variables respuestas correlacionadas). Si se observa el diagrama de dispersión de las respuestas, pareciera que ambas dietas son similares en cuanto al engorde del ganado, sin embargo, al encontrar el envelope estimado y proyectar los datos ahí, es fácil ver que el engorde de la dieta representada por el color azul es superior al de la dieta representada con el color rojo.

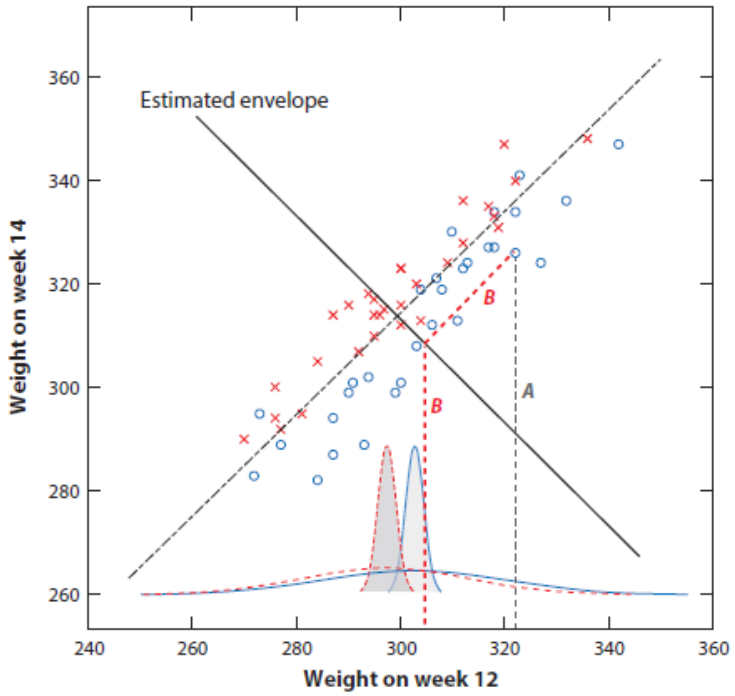

Figura 2-2: Gráfico de los datos de dos variables respuesta Peso en la semana 14 y 16 con un factor fijo [2]. La linea punteada es el subespacio donde la respuesta es invariante a cambios en las variables regresoras. Los vectores A,B representan la transformación que se le hace a un punto para pasar de alta dimensión a baja dimensión.

Si se usara un método usual como ACP, la primera componente principal sería estimada aproximadamente como la línea puntuada ortogonal a la linea que representa el envelope, ya que en esta dimensión o eje, la variabilidad es alta. No obstante, en este caso esa dimensión representa el subespacio donde la respuesta es invariante a observaciones en la variable regresora, i.e., el subespacio donde no hay información de interés, por lo cual se confundirían aun más los resultados pues no se tuvo en cuenta la información condicional de un modelo, llegando a conclusiones equivocadas dado que no se podria diferenciar entre las dos poblaciones.

Estimación del subespacio central

El éxito de la clasificación de los datos, se basa casi que únicamente en la efectividad del algoritmo de hallar el subespacio central y por lo tanto los envelopes. Por lo cual la elección de como estimarlos es de suma importancia.
En este caso se selecciona el algoritmo Envelope component screening, desarrollado en el artículo [2], el cual usa el algoritmo $1D$ y según el autor mejora la forma de maximizar la verosimilitud del modelo. Estos se encuentran resumidos en [3] y son presentados en las tablas **2-1** y **2-2**

Algoritmo 1: El algoritmo 1D (Cook y Zhang, 2016)

Sean $g_k \in \mathbb{R}^p$, $k = 1, \ldots, u$ las direcciones secuenciales obtenidas. Sea $G_k = (g_1, \ldots, g_k)$
Sea (G_k, G_{0k}) una base ortogonal para $\mathbb{R}^p$ y fije el valor inicial $g_0 = G_0 = 0$
Para $k = 0, \ldots, u-1$, repita el paso 1 y 2 como sigue:

1. Sea $G_k = (g_1, \ldots, g_k)$ y sea (G_k, G_{0k}) una base ortogonal de $\mathbb{R}^p$.
Defina $N_k = [G_{0k}^T(M+U)G_{0k}]^{-1}$, $M_k = G_{0k}^T M G_{0k}$ y la función sin restricciones $\phi(w) = \log(w^T M_k w) + \log(w^T N_k w) - 2\log(w^T w)$

2. Obtenga la solución $w_{k+1} = argmin\phi(w)$ Luego la dirección del $(k+1)$ ésimo envelope es $g_k = G_{0k} w_{k+1}/||w_{k+1}||$

Tabla 2-1: Algoritmo 1D

Algoritmo 2: La proyección de componentes de envelopes (ECS)

1. Construya una descomposición de M como $M = \sum_{i=1}^{p} \lambda_i v_i v_i^T$ donde $v_i^T v_j = \delta_{ij}$ Siendo δ_{ij} un delta de kronecker.

2. Evalúe $f_i = F(v_i) = \log(\lambda_i) + \log(v_i^T(M+U)^{-1}v_i)$ y luego ordénelas $f_p \leq \cdots \leq f_1 \leq 0$ con los correspondientes v_i

3. Sea $A_0 = (v_1, \ldots, v_{p-d})^T$ y $A = (v_{p-d+1}, \ldots, v_p) \in \mathbb{R}^p$ con d especificado desde un principio.

4. Estime $\xi_M(U) \sim A\xi_{A^T MA}(A^T U A)$

Tabla 2-2: Algoritmo ECS

Los dos algoritmos anteriormente mencionados dependen de unos parámetros iniciales, según se enuncia en [2], se pueden tomar diferentes elecciones de parámetros iniciales, una de estas elecciones puede ser $M = S_{Y|X}$, $U = 0$ y $G = eigen(M)$, con el anterior se forma

una base ortogonal y a partir de los vectores columna de esta matriz se pueden construir $G_0, G_1, \ldots, G_{d-1}$, así mismo con los vectores restantes se puede construir $G_{00}, G_{01}, \ldots, G_{0(d-1)}$.
Es claro que como se está buscando maximizar una verosimilitud, se están haciendo supuestos distribucionales. Sin embargo, en el artículo [2] están demostradas las propiedades asintóticas del método y algoritmo. Su principal supuesto es que los residuales sean normales e independientes. Así como que las variables tengan relaciones lineales que permitan reducir la dimensión.

2.2. Modelos lineales mixtos

Una vez comprendida la parte de reducción de dimensión, es necesario incluir en el modelo estructuras de covarianza. Esto, debido a que en casos donde las observaciones no son totalmente independientes se podría llegar a conclusiones equivocadas al no cumplirse el supuesto de independencia.
En esta sección se hace una breve descripción de los modelos lineales haciendo un énfasis en la parte aleatoria y cómo afecta al modelo. El acercamiento que se emplea en esta sección se basa en lo descrito en [9].
Teniendo en cuenta que un modelo lineal es aquel que tiene la forma $Y = X_1B_1 + X_2B_2 + \ldots X_KB_K$, cuando se habla de modelos lineales con efectos fijos se piensa en la formula $E(Y) = X\beta$, donde β es un vector de constantes.
Lo que representa o busca este modelo es que a través de diferentes variables o factores, se pueda llegar a explicar de mejor manera la variabilidad que se encuentra en Y.
El modelo lineal con factores fijos más sencillo, es el modelo $Y = \mu + e$. Siendo μ la media y e un vector de residuales. Es claro que μ es una constante que puede ayudar a explicar la variabilidad de los datos.

Cuando se habla de un modelo lineal mixto, se refiere a un modelo lineal, e.g., $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$ dónde se tienen factores fijos y factores aleatorios.
Para entender el modelo planteado anteriormente, suponga que la variable respuesta hace referencia a la producción de cervezas en una fabrica que depende de la maquina i y el trabajador j. Si supone que solo hay dos máquinas, pero 100 empleados. μ, α_i serían factores fijos, mientras que β_j seria un factor aleatorio, puesto que la producción va a depender de que trabajador se seleccione.
En modelos lineales, los factores fijos tienen el objetivo de modelar la media, mientras que los factores aleatorios tienen como objetivo modelar la estructura de varianza de las observaciones. El crear un modelo mixto solo surge de la necesidad de simplificar la matriz de covarianza, pues de no incluirse los factores aleatorios, la estructura de esta matriz sería desconocida y difícil de estimar.

En general un modelo lineal mixto tiene la forma

$$E(Y|U = u) = X\beta + Zu \tag{2-7}$$

Donde X es la matriz diseño para la parte fija, β el vector de parámetros fijos y Z la matriz diseño para la parte aleatoria, siendo entonces u el vector de efectos aleatorios (en realidad la realización del vector aleatorio U).
Sin perdida de generalidad se puede decir que $E(u) = 0$, $var(u) = D$ y $var(Y|u) = R$. Por lo tanto el vector Y tiene distribución con media $X\beta$ y varianza $ZDZ^T + R$. i.e.

$$Y|X \sim (X\beta, ZDZ^T + R) \tag{2-8}$$

Mostrando que efectivamente, los factores fijos solo afectan la media y los aleatorios solo la varianza.
A pesar que la descripción del modelo lineal mixto parece ser muy sencilla, la complejidad nace al momento de aplicar la teoría, pues es necesario decir a priori qué factores deben ser considerados como aleatorios, lo cual es fácil o difícil de decidir dependiendo el tema que se esté tratando.

2.3. Datos de secuenciación de RNA

En las secciones anteriores se introdujeron los métodos de reducción de dimensión para datos en general, sin embargo, aún no se ha mencionado qué son los datos de secuenciación de RNA en células unitarias, su potencial, cómo se comportan y qué características tienen. A continuación se presenta un breve resumen de esta información para así poder comprender de mejor manera qué tipo de método se necesita realmente y por qué es importante estudiar estos datos.
Para hablar de datos de secuenciación de RNA en células individuales o unitarias, es necesario primero aclarar a qué hace referencia el término secuenciación de RNA.
En un principio los estudios de transcriptómica se basaban en microarreglos de hibridación, cuyo objetivo era entender totalmente las diversas moléculas de RNA que se expresan en los diferentes niveles del genoma, sin embargo, este tipo de tecnologías no tenían este alcance. Una vez se introdujeron las técnicas de secuenciación de siguiente generación (Next generation DNA sequencing NGS), se revolucionó la transcriptómica permitiendo el análisis de RNA a través de cDNA a escala masiva [11]. Esto dió origen a lo que hoy en día se conoce como secuenciación de RNA, que básicamente lo que hace es producir millones de pequeñas lecturas de la secuencia de un genoma, epigenoma o transcriptoma, catalogando y cuantificando las diversas moléculas de RNA, esta nueva tecnología trae consigo grandes ventajas con respecto a los anteriores métodos que trataban de entender los genomas. La principal ventaja es que para producir estas lecturas del genoma no se necesita ningún conocimiento a priori a diferencia del PCR y microarreglos, lo que trajo una facilidad para crear

nuevos conjuntos de datos y con esto abrir una posibilidad para comprender nuevos procesos biológicos [10].

2.3.1. Datos de secuenciación en células unitarias

La secuenciación de RNA en un principio era utilizada para analizar los genomas, epigenomas o transcriptomas diferenciando por individuos. No obstante, siguiendo el acercamiento que se hace en [13], estudiar la secuenciación de RNA mensajero en células unitarias puede cambiar la forma de entender la heterogeneidad celular dentro de un órgano, es decir analizar las secuencias en diferentes células dentro de un mismo paciente permite entender procesos biológicos con mayor claridad; en algunos casos encontrar células atípicas analizando muestras dentro de un mismo paciente puede servir para detectar el resultado de una infección, resistencia a un antibiótico o una reincidencia de un cáncer.
Una sola célula puede contener grandes cantidades de información, en específico una célula promedio humana contiene 6 mil millones de pares de bases de ADN y 600 millones de bases de RNA mensajero [5]. Esta información se puede recopilar a través de diferentes métodos, pero del que se hablará a continuación es a través de una técnica llamada transcriptoma de células unitarias.

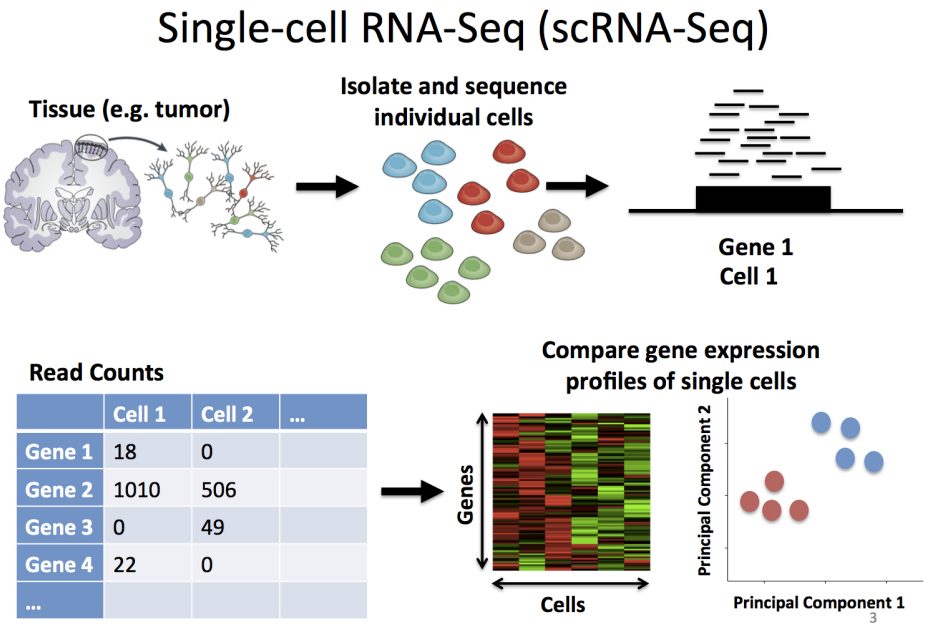

Figura 2-3: Procedimiento de secuenciación SC-RNA-seq [6]

Usualmente los estudios de transcriptomas se hacen a nivel de tejido, sin embargo, cuando se

trata de células madre o cáncer no se puede lograr este alcance, dejando estudios únicamente a nivel muestral.
Para hacer exitoso el procedimiento de transcripción es necesario hacer uso de dos técnicas, el aislamiento de diferentes células y la transformación de RNA en cDNA que básicamente permite el conteo de lecturas de genes. Este procedimiento se encuentra esquematizado en la figura **2-3**. Primero se selecciona un tejido, posteriormente se aíslan las células, se hace la secuenciación, lo que da como resultado una matriz de conteo y finalmente se hace el análisis estadístico.

Aislamiento de células

Para esta labor de aislar las células, potencialmente heterogéneas, existen varios métodos como Ordenamiento de células por activación de fluorescencia, manejo de células basadas en opto-fluidos, manejo de células basada en micro-fluidos y micro disección capturada por láser. No obstante, en este resumen solo se mencionará el primer método al ser el más común y el más utilizado.
El Ordenamiento de células por activación de fluorescencia es el más usado por sus bajos costos y su interfaz fácil de manejar. Lo que propone esta técnica es combinar la citometría de fluidos con un ordenamiento basado en un sistema de puerta de fluorescencia como se ve en la figura **2-4**. Por lo que este sistema de puerta permite aislar las células dependiendo de la fluorescencia. Estos anticuerpos fluorescentes pueden aislar las células requeridas de acuerdo a unos marcadores de terminados (actualmente se pueden usar hasta 17 marcadores simultáneos). Una vez aisladas las células se procede a leer la información que tengan disponible.

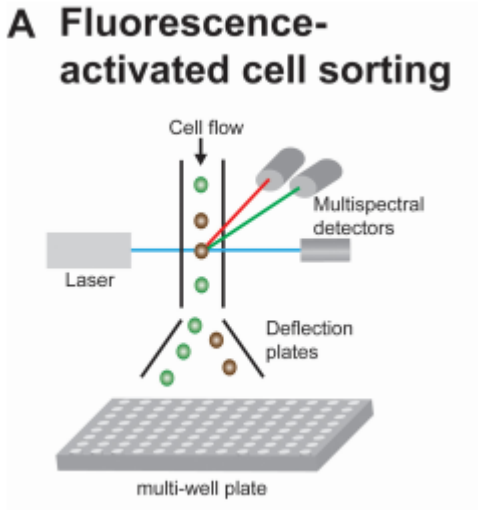

Figura 2-4: Ordenamiento de células por activación de fluorecencia [5]

Secuenciación de RNA en células unitarias

Como no es posible secuenciar las moléculas de RNA, lo que generalmente se hace, es capturar el transcriptoma de las células unitarias, esto es posible y se basa en tres pasos importantes.

1. Transcripción en reversa del RNA en una primera cadena de cDNA.
2. Segunda síntesis de cadena y amplificación del cDNA.
3. Secuenciamiento del cDNA con uso de nuevas tecnologías.

Diversas técnicas que incluyen esta serie de pasos logran el objetivo de capturar el transcriptoma de las células, 3 métodos comúnmente utilizados están esquematizados en las figuras **9-2** y **9-3** ubicadas en los anexos. Todas las técnicas dejan como resultado una librería de secuenciación que son los datos a tratar.

Importancia del estudio de la secuenciación del RNA en células unitarias

La importancia de utilizar estos datos radica en que inclusive tomando células de un mismo individuo, la secuenciación del RNA es tan variable que coeficientes de correlación pueden estar incluso por debajo de 0.5 entre la expresión de los genes de dos células, lo cual afirma que biológicamente hay diferencias entre expresiones de genes, encontrando que se pueden crear identidades de células y posiblemente también etapas.
Un tema de investigación actual es mirar cómo es el proceso de diferenciación en células madres. A través de la observación de este proceso se ha encontrado que hay diferentes subtipos de células que conforman los pulmones de algunos roedores.
También a través de la observación de la secuenciación de RNA en células unitarias se pueden observar etapas en el desarrollo de un embrión. Lo que lleva al objetivo más grande que es ver como las células se dividen y diferencian para crear un organismo completo[5].
Finalmente, los datos de secuenciación de células unitarias o individuales pueden ser vistas como un nuevo tipo de fuente de datos para la estadística genómica. Por lo tanto se necesita la creación de nuevos métodos y procesos que permitan analizar correctamente los datos. Hasta el momento se han mencionado dos aspectos muy importantes para incluir en estos procesos, siendo el primero que es necesario un método de reducción de dimensión, pues es difícil analizar poblaciones de individuos (células) cuando no se pueden visualizar. Es por eso que esa secuenciación de cientos de genes, reflejada en dos o tres componentes puede dar enormes cantidades de nueva información.
El segundo aspecto que se debe tener en cuenta para el análisis de datos de SCSeq es que debe incluirse un modelo para la reducción de dimensión, principalmente porque la información que tienen los investigadores en estas áreas debe ser incluida y no se puede malgastar el conocimiento que ya se ha conseguido en las últimas décadas acerca de poblaciones celulares; además porque este tipo de datos no deberían ser independientes debido a que diferentes

células dentro de un mismo paciente deben tener en común mucha información y al compararlas con células de otros pacientes, esta información repetida tiene que ser tenida en cuenta.

3 Antecedentes: reducción de dimensión para SCseq

Actualmente son utilizados principalmente el análisis de componentes principales (ACP) y el T-distributed stochastic neighbor embedding (t-SNE) para hacer reducción de dimensión en datos de secuenciación de RNA.
Es necesario recalcar que este tipo de datos genómicos generalmente se usan para clasificar muestras y así hacer diagnósticos, por lo tanto estos métodos, más allá de si reducen o no la dimensión deben estar centrados en guardar la información de la variabilidad de las muestras ya que solo con esta información va a ser posible clasificar bien los datos.

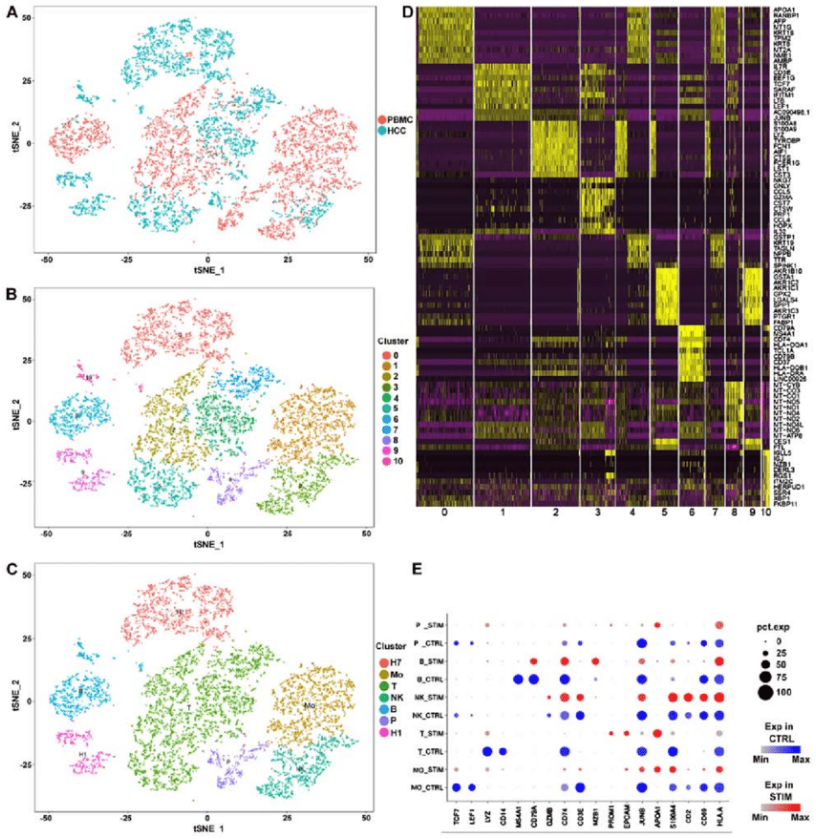

Figura 3-1: Visualización de estructuras en dimensiones reducidas obtenidas en [16]

Un ejemplo de lo mencionado anteriormente se puede ver en la imagen C de la figura **3-1**

que es resultado de un estudio que hace uso del algoritmo t-SNE. En este, se puede ver como una reducción de dimensión permite visualizar y clasificar tipos de células. En este caso se tiene la secuenciación de algunos genes en células de 360 pacientes con cáncer de hígado y tumores, lo cual posteriormente se relaciona con el paciente y por lo tanto con el estado de su cáncer. Adicional, por medio de la reducción de dimensión, se pueden encontrar genes relacionados con el cáncer y posteriormente como en la imagen E de la figura **3-1** observar qué porcentaje de células (tamaño del punto) con una alta expresión del gen en cuestión (eje x) tiene cada individuo o grupo de personas (eje y).
Otro estudio más acorde a lo que plantea este trabajo es [4], en el cual proponen el método SCVIS, una modificación del t-SNE, y a partir de una simulación comparan los rendimientos de los dos métodos como se puede ver en **3-2**

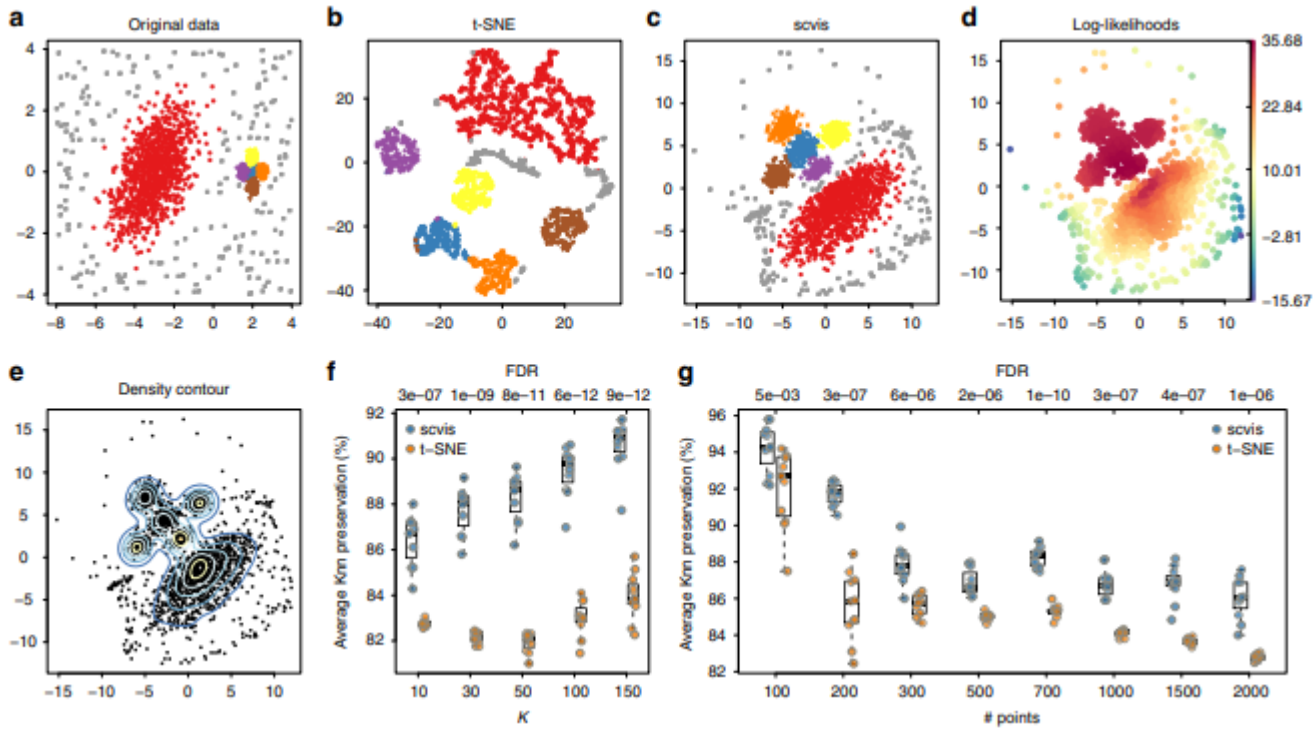

Figura 3-2: Simulación y resultados obtenidos en [4]

En este estudio simulan una estructura de baja dimensión como se observa en la imagen 'a'. Posteriormente, se crean nuevas variables como funciones de las variables de baja dimensión y después, con estos nuevos datos, se utilizaron los métodos para reducir la dimensión (imágenes 'b' y 'c'). Mantener la estructura de baja dimensión hace referencia a que el algoritmo capture las formas que se tenían en baja dimensión, en este caso 6 nubes de puntos y unos puntos de ruido uniformemente distribuidos en el fondo (los grises). Si el algoritmo captura bien esas formas y logra representarlas nuevamente, es de esperarse que posteriormente se pueda clasificar de manera más precisa. Bajo esta simulación, el algoritmo SCVIS se com-

portó mejor pues la estructura que obtuvo es más parecida a la original que la del t-SNE, lo cual se vio reflejado en una mejor clasificación como se puede observar en las imágenes 'f' y 'g' donde se ve el porcentaje de puntos correctamente clasificados haciendo uso del algoritmo vecinos más cercanos y variando el número de vecinos más cercanos entre 10 y 150.
Aparte de los resultados a través de las simulaciones, es necesario aplicar los métodos a bases de datos reales que ya hayan sido estudiadas. El procedimiento es tomar los datos y reducir la dimensión para posiblemente encontrar clusters o agrupaciones de datos que pueden tener diferentes interpretaciones, e.g., nubes de puntos que determinen cuáles células son malignas y cuáles no. Un ejemplo de esto se puede ver en la imagen 'c' de la figura **3-3**, dónde se hace un análisis de componentes principales para los datos de cada individuo buscando identificar los tipos de células que tiene el paciente y posteriormente, relacionarlo con la progresión del cáncer. Estos datos adicionalmente fueron trabajados en [4] por lo cual va a ser posible comparar resultados haciendo uso de diferentes métodos.

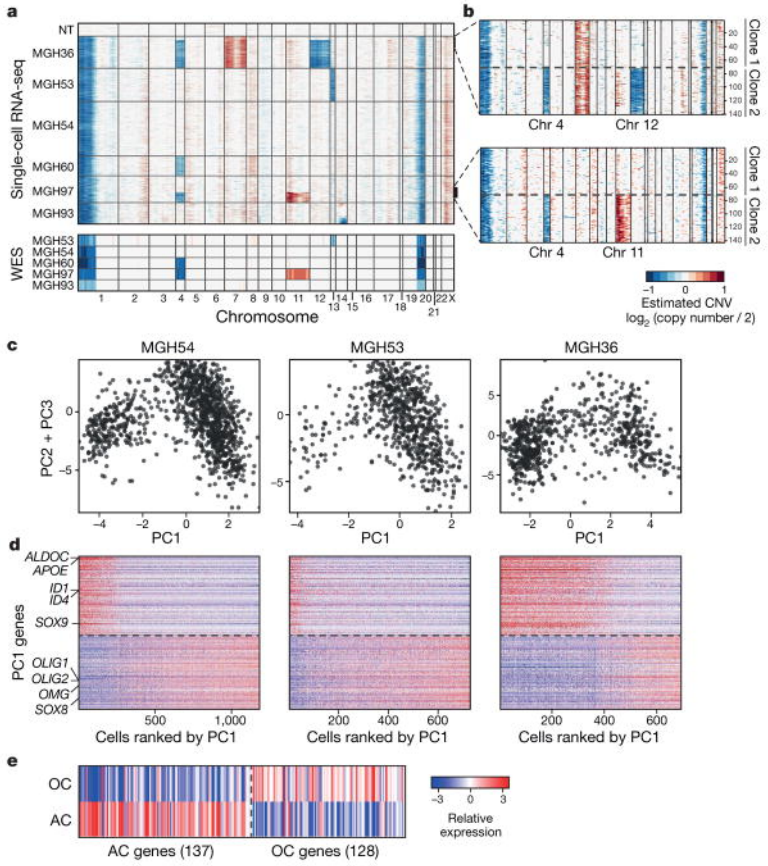

Figura 3-3: Datos de oligodendroglioma obtenidos en [14]

4 Metodología

El trabajo realizado puede ser dividido en 3 secciones principales:

1. Simulación de datos de secuenciación en células unitarias
2. Comparación de Métodos
3. Aplicación de Envelopes y modelo propuesto a bases de datos reales

En la primera, se realiza la creación de bases de datos según se entiende el problema de los datos de células individuales, en el segundo se evalúan los métodos mencionados en el marco teórico para reducir la dimensión de los datos y evaluar según tasas de alta clasificación y finalmente en el tercero, se utilizan los envelopes para mostrar su comportamiento en bases de datos reales que ya han sido trabajadas con diferentes métodos, para así concluir si su uso es adecuado en el área de la estadística genómica.

4.1. Simulación de datos de SCseq

En el artículo [4], presentan simulaciones de datos de secuenciación de RNA en células individuales bajo la idea de que hay una dimensión intrínseca de los datos (baja dimensión) donde se pueden ver diferentes clusters. Estos clusters suponen una diferenciación en la expresión de diferentes genes según las diferentes condiciones en la muestra i.e. un cluster podría representar las diferentes células de un individuo, también podría representar un tipo de células entre los diferentes individuos, o inclusive cosas que en un principio no se tengan en cuenta como género, etnicidad, etc.
Para las simulaciones, se va a suponer que los clusters están conformados únicamente por el tipo de célula, e.g, cancerígena y no cancerígena, mas no por el individuo. Esto, debido a que si se pretende comprender un proceso biológico debería este ocurrir proporcionalmente en diferentes individuos y no ocurrir de diferentes maneras en los diferentes organismos.

4.1.1. Simulación 1

Para la primera simulación se usa el supuesto más simple, que los clusters dados por los diferentes tipos de células tienen una dispersión similar en los diferentes individuos, es decir,

que si se miraran los clusters conformados en cada individuo, estos deberían ser similares. Para la creación de estos datos se sigue el siguiente procedimiento:

1. Se crea una matriz diseño X
   $X = (1_I \otimes I_T \otimes 1_R)$
   Donde I representa el numero de individuos, T el tipo de célula y R las replicaciones que en este caso son la célula. Donde $I = 50$, $T = 3$ y $R = 5$.
   Nótese que esta matriz tiene $T = 3$ columnas donde cada supuesta célula tiene un 0 o 1 dependiendo de si es o no del supuesto tipo de célula correspondiente

2. Se implanta una semilla para poder replicar los resultados (2021 para este caso)

3. Se generan tres vectores de parámetros $\theta_1, \theta_2, \theta_3$ los cuales representaran a los efectos.

4. Se generan las variables de baja dimensión (dimensión intrínseca) como $Z_i = X\theta_i + e_i$, siendo e_i un error con distribución normal de media cero y con la estructura de covarianza $4 \times I$

5. Se genera las variables de alta dimensión según el escenario (Esc):

Esc	Y1	Y2	Y3	Y4	Y5	Y6
I	$0.5Z_1$+$0.9Z_2$	$1.93Z_1$-$0.75Z_2$	$1.69Z_1$- $3.4Z_3$	$8.7Z_1$-$6.3Z_2$ -$2.7Z_3$	-$1.77Z_2$+$3.22Z_3$	Z_1-$3Z_2$+Z_3
II	$3Z_1Z_2$	$\exp(Z_1 + Z_2)$	$\frac{Z_1}{2Z_3}$	$\frac{Z1+Z_2}{Z_3}$	$Z_1 * Z_2 * Z_3$	$Z_1^2 + Z_2^2$
III	$Z_1 - Z_2$	$Z_1 + 2Z_3$	$Z_1 - Z_2 + Z_3$	$Z_1^2 + Z_2$	Z_1Z_2	$\exp(Z1 + Z_2)$

   Estas variables son contaminadas levemente con un error normal independiente de media cero ($\sigma^2 = 0{,}1$).

4.1.2. Simulación 2

Es de suponer que aunque el proceso biológico sea el mismo, la medición entre diferentes individuos debe cambiar. Por lo tanto el individuo tiene que ser un factor importante, no obstante, este no debe hacer que los clusters cambien su forma o que la combinación de la expresión de los genes cambie. Consiguientemente, se supone que el individuo es un factor aleatorio, lo cual cambia la dispersión de los datos más no la expresión media de los diferentes genes.
Para producir un set de datos con estas condiciones se usa el siguiente procedimiento:

1. Se crea una matriz diseño X
   $X = (1_I \otimes I_T \otimes 1_R)$
   Donde I representa el número de individuos, T el tipo de célula y R las replicaciones

que en este caso son las diferentes células. Donde $I = 5$, $T = 3$ y $R = 50$.
Nótese que esta matriz tiene $T = 3$ columnas donde cada supuesta célula tiene un 0 o 1 dependiendo de si es o no del supuesto tipo de célula correspondiente

2. Se implanta una semilla para fácil replicación (2021 para este caso)

3. Se generan tres vectores de parámetros $\theta_1, \theta_2, \theta_3$ los cuales representaran a los efectos.

4. Se generan las variables de baja dimensión (dimensión intrínseca) como $Z_{ijk} = X\theta_j + e_{ijk} + \epsilon_i$. Siendo e_{ijk} un error con distribución normal de media cero y con la estructura de covarianza $4 \times I$ y ϵ_i un error con distribución normal con estructura de covarianza $2I$. La estructura de covarianza de los vectores Z seria $2 \times (I_I|1_T 1_T^T|1_R 1_R^T)$

5. Se genera las variables de alta dimensión según el escenario (Esc):

Esc	Y1	Y2	Y3	Y4	Y5	Y6
I	$0.5Z_1$+$0.9Z_2$	$1.93Z_1$-$0.75Z_2$	$1.69Z_1$- $3.4Z_3$	$8.7Z_1$-$6.3Z_2$ -$2.7Z_3$	-$1.77Z_2$+$3.22Z_3$	Z_1-$3Z_2$+Z_3
II	$3Z_1Z_2$	$\exp(Z_1 + Z_2)$	$\frac{Z_1}{2Z_3}$	$\frac{Z1+Z_2}{Z_3}$	$Z_1 * Z_2 * Z_3$	$Z_1^2 + Z_2^2$
III	$Z_1 - Z_2$	$Z_1 + 2Z_3$	$Z_1 - Z_2 + Z_3$	$Z_1^2 + Z_2$	Z_1Z_2	$\exp(Z1 + Z_2)$

Estas variables son contaminadas levemente con un error normal independiente de media cero ($\sigma^2 = 0{,}1$).

Esta simulación pretende ser más aterrizada, en el sentido que se tiene en cuenta la diferencia entre individuos y también por el número de replicas (alto) frente al número de individuos (bajo) que se usa en la realidad.

4.2. Comparación de métodos

Una vez son generadas las variables, se corren los 3 diferentes algoritmos (Envelopes, PCA, T-SNE).
Es importante destacar que por la forma en que fueron construidos los datos, existe un modelo lineal establecido, para el cual sus estimaciones por máxima verosimilitud serían bastante superiores a los resultados con los algoritmos a comparar; sin embargo, no es incluido este modelo en el análisis porque no tiene sentido ponerlo a competir debido a que si así fueron creados los datos es lógico que sea el mejor. El objetivo de esta sección es comparar los 3 algoritmos mencionados para datos que se entienden como de secuenciación de RNA en células individuales.
El procedimiento realizado se puede resumir asi:

1. Generar los datos según sea el escenario y simulación

2. Correr los 3 algoritmos con los datos: el resultado de esto son 3 vectores por método, pues se supone que la dimensión intrínseca es 3 (así se construyeron los datos).

3. Se pone una semilla para fácil replicación

4. Se corre el algoritmo k-nearest-neighbors ($k = 15$) con datos iniciales seleccionados aleatoriamente (muestra de entrenamiento).
   La muestras de entrenamiento están conformadas con el 50 % de los datos.

5. Se calcula la tasa de clasificación evaluando los datos en la muestra de prueba. Se coloca 1 si efectivamente el dato fue clasificado en el cluster correcto, 0 si no fue clasificado correctamente.

6. Se presentan los clusters y se usa la prueba de Kruskal-Wallis para ver si hay diferencias entre las tasas de clasificación, esta prueba de hipótesis general sera notada como (PHG) posteriormente.

7. De existir diferencia, se usa la prueba de Wilcoxon para evaluar las hipótesis de si existen mejores tasas en algún algoritmo (PH1,PH2,PH3).

   $$PH1: \quad H_0 : TC_{ENV} \leq TC_{PCA} \quad vs \quad H_A : TC_{ENV} > TC_{PCA}$$

   $$PH2: \quad H_0 : TC_{ENV} \leq TC_{TSNE} \quad vs \quad H_A : TC_{ENV} > TC_{TSNE}$$

   $$PH3: \quad H_0 : TC_{PCA} \leq TC_{TSNE} \quad vs \quad H_A : TC_{PCA} > TC_{TSNE}$$

   $TC_{método}$ hace referencia a la tasa de correcta clasificación obtenida por el método.

8. Se miran las correlaciones lineales más altas de cada método con las dimensiones de los datos de baja dimensión

4.3. Procedimiento simulación

Anteriormente, se explica como se creó la base de datos simulada, sin embargo no se puede concluir con el rendimiento de una simulación, por lo tanto es necesario hacer replicas de cada simulación para ver el rendimiento de los diferentes algoritmos.
El procedimiento general para cada escenario en las dos simulaciones es el siguiente:

1. Se determina la semilla 2021.

2. Para i en $\{1, 2, 3, \cdots, 1000\}$.

   2.1 Se crean los datos de baja dimensión según la simulación (1 o 2) .

2.2 Se crean los datos de alta dimensión según el escenario.

2.3 Para método en (Envelopes, PCA, T-SNE)

2.3.1 Aplicar método de reducción de dimensión

2.3.2 Para las dimensiones 1,2,3, guardar la correlación más alta que se encuentre entre los resultados del método y la dimensión de los datos originales.

2.3.3 Usar K.N.N ($k = 15$) con 500 muestras de entrenamiento diferentes.

2.3.4 Guardar las 500 tasas de clasificación

2.4 Realizar test de Kruskal-Wallis (PHG).

2.5 Guardar el valor p de la prueba PHG[i]

2.6 Realizar el test de Wilcoxon para evaluar las hipótesis PH1,PH2 y PH3.

2.7 Guardar los valores p PH1[i], PH2[i], PH3[i].

3. Analizar los resultados obtenidos por correlaciones y pruebas de hipótesis.

4.4. Datos reales

Como se mencionó en un principio, es necesario usar una base de datos que haya sido usada para estudios que usen una metodología con reducción de dimensión, por lo cual es seleccionada la base de datos de oligodendriogliomas de humanos sacada de [14]. Esta es analizada con componentes principales en [14] y es analizada con un algoritmo no lineal en [4].
Esta base de datos esta conformada por 6 pacientes, de los cuales se dividen en dos grupos de tres pacientes, haciendo referencia a la profundidad de secuenciación i.e. el número de células secuenciadas.

4.4.1. Tratamiento de los datos

El set de datos de oligodendrogliomas a pesar de ya tener normalizaciones, también necesita con tratamiento para poder ser incluido en un análisis. De los 23703 genes que se encuentran descritos en el set de datos, se eliminaron los genes que no tenían lecturas en más del 81 % de las células. Posteriormente se eliminaron las células que no tenían lectura en ningún gen. Quedando así un set de datos conformado por 4347 células y 963 genes.
Una vez finalizado el control de calidad se hace una reducción de dimensión con los diferentes métodos, se clasifican los datos, se obtienen los genes que más le aportan a la reducción de dimensión y finalmente, se hace un enriquecimiento a través de la herramienta DAVID (The Database for Annotation, Visualization and Integrated Discovery).

5 Resultados

5.1. Datos simulados

Para poder comparar los métodos se crearon bases de datos con diferentes condiciones. Los resultados fueron los siguientes

5.1.1. Simulación 1

Para la simulación 1, se realizó la creación de muestras aleatorias cuyo modelo para la baja dimensión o dimensión intrínseca era $ZF_{ijk}^{d} = \alpha_j + e_{ijk}$. Siendo α_j el único efecto del modelo quien representa el tipo de célula y e_{ijk} observaciones de una variable aleatoria normal $i.i.d$ con $\mu = 0$ y $\sigma^2 = 4$.
El resultado de una simulación es el presentado en la figura **5-1**

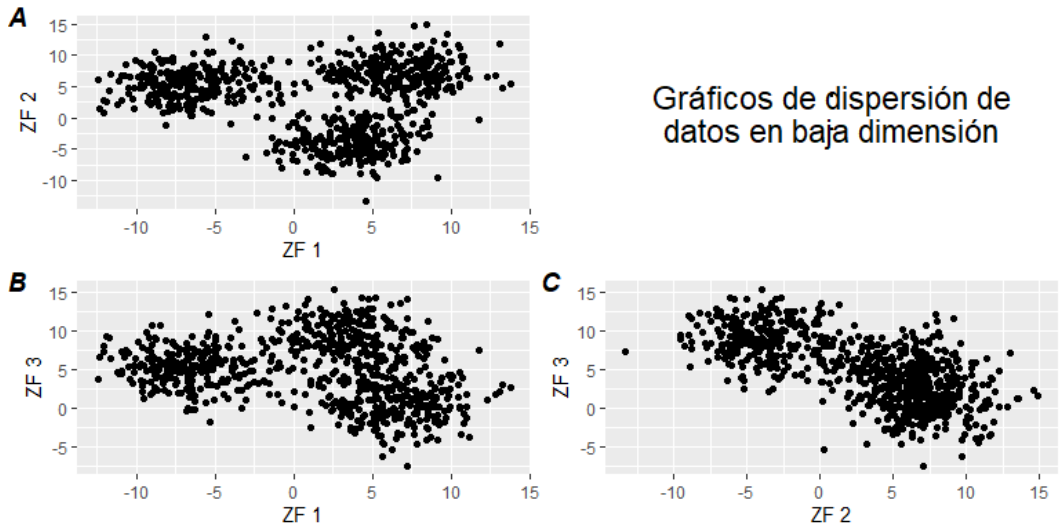

Figura 5-1: Representación gráfica de variables de baja dimensión

Cabe recordar que para esta simulación, los clusters de los diferentes individuos son similares. En las figuras **5-1 B** y **C** se puede ver lo que en unos datos no simulados se entendería como el proceso biológico, ya que se espera que haya una transición entre tipos de células. Es importante que cuando se utilicen los algoritmos, este tipo de fenómeno se siga visualizando.

Esta dimensión intrínseca es alterada según el escenario encontrando los resultados mostrados en las figuras **5-2**, **5-3** y **5-4**. Es necesario resaltar que en este caso la alta dimensión puede ser visualizada fácilmente en un gráfico como los recién mencionados. Sin embargo, esto es arduamente logrado cuando la dimensión de los datos no es 5 sino más de 20.000.
Lo primero que se puede ver es que así como se espera en los datos de secuenciación de ARN en células individuales, los datos graficados muestran altas correlaciones en el escenario 1, esto se puede pensar ya que existen diferentes factores biológicos que hacen que distintos genes tengan altas correlaciones en sus niveles de expresión causadas por relaciones funcionales, proteínas, condiciones, etc.

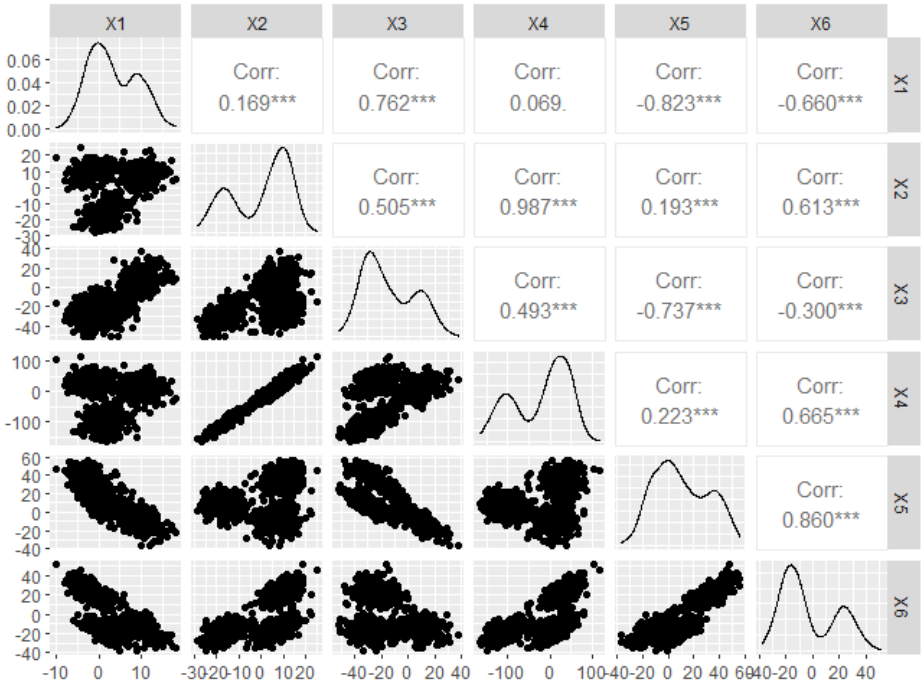

Figura 5-2: Representación gráfica de variables de alta dimensión (escenario 1)

Si se utilizara un método para clasificar los datos, se podría llegar al error de clasificar en solo 2 grupos así como se muestra en la figura **5-2** pues todos los histogramas parecieran indicar una mixtura entre dos poblaciones. No obstante, ya se conoce en este caso que son 3. Adicionalmente, es importante resaltar que en este caso parece evidente la clasificación, sin embargo, entre más alta se la dimensión, la probabilidad de que existan varios genes donde no se identifiquen los tipos de células aumenta, haciendo que los algoritmos puedan equivocarse por tomar variabilidad de genes que no tengan que ver con el proceso biológico de interés.

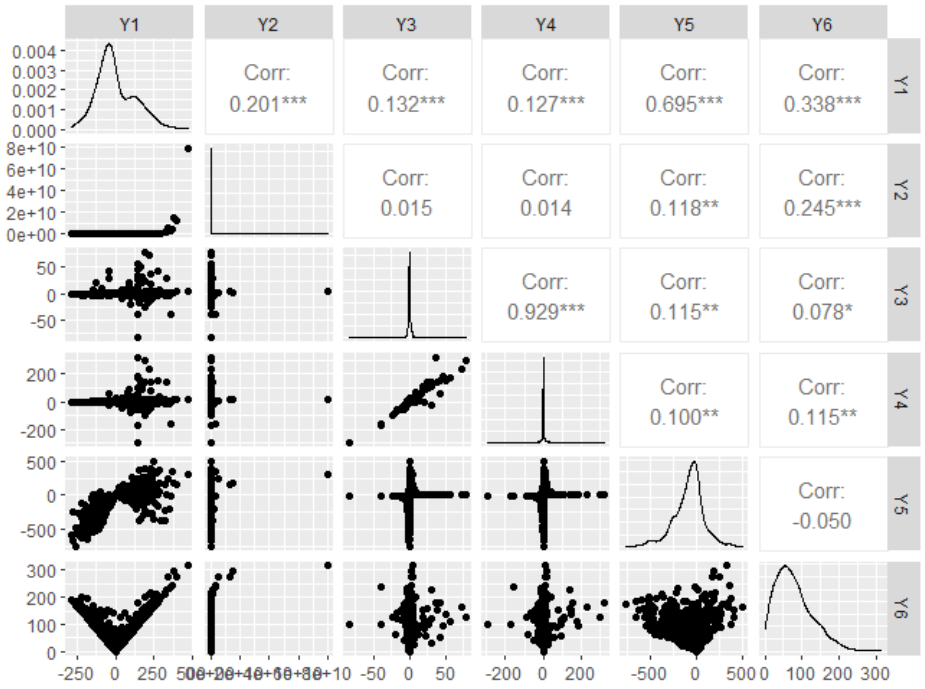

Figura 5-3: Representación gráfica de variables de alta dimensión (escenario 2)

En la figura **5-3** se puede ver lo que se estaba hablando anteriormente, las combinaciones no lineales hacen que aparezcan lo que serían genes dónde no se ven los subgrupos. Al ser alterada la baja dimensión por relaciones no-lineales también bajan las correlaciones lineales entre variables de alta dimensión y se hace evidente que es necesario un buen método que logre identificar las componentes de baja dimensión, pues inclusive si se lograran visualizar como en este caso todas las variables seria compleja la tarea de sacar subgrupos manualmente. Para este caso tan drástico, inclusive se podría concluir precipitadamente que no hay efecto entre los tipos de célula dado que los histogramas muestran una única población.
En cuanto al escenario 3 (figura **5-4**), se espera que sea el más parecido a la realidad, ya que es de esperarse que si existan relaciones no lineales entre genes, pero también que existan relaciones lineales. En ese sentido si se espera que existan genes que hagan perder los algoritmos como Y_6, relaciones como la existente entre Y_5 y Y_1, pero también relaciones como la que se puede visualizar entre las variables de alta dimensión Y_2 y Y_3. En ese sentido es interesante ver como funcionan los diferentes algoritmos, ya que, el PCA y los envelopes, son netamente lineales, ¿será suficiente que con algunas relaciones lineales se comporten a la par de un algoritmo como el T-SNE?.
Note que en la figura **5-4** tampoco es tan fácil hacer la clasificación de manera manual y que también los histogramas parecen indicar una mixtura de dos poblaciones.

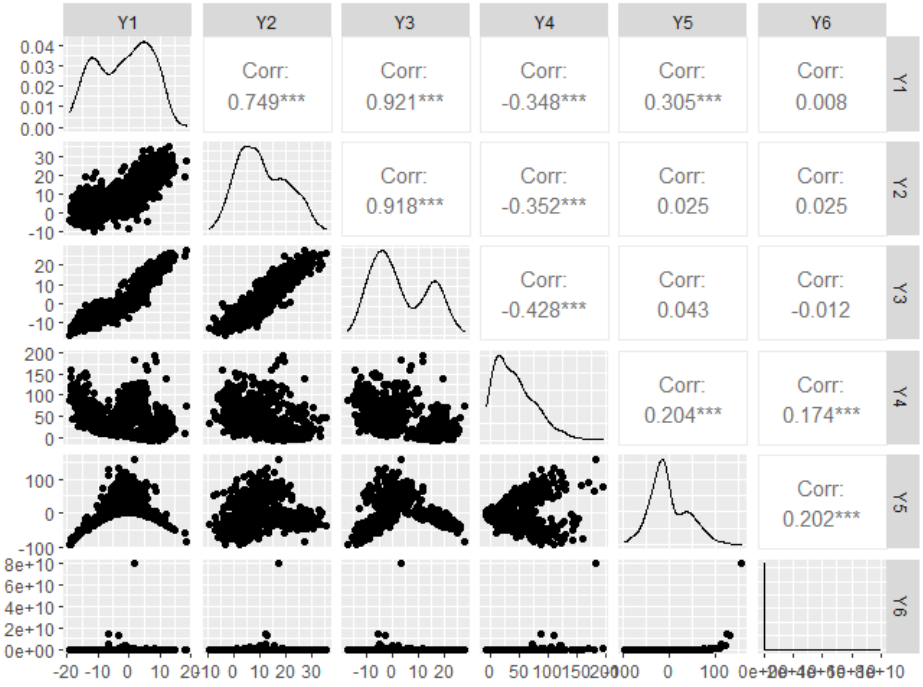

Figura 5-4: Representación gráfica de variables de alta dimensión (escenario 3)

5.1.2. Simulación 2

Para la simulación 2 se incluyo una variable muy importante que es el individuo de donde provienen las células. Por lo mencionado en la métodología se incluyo como un factor aleatorio, quedando el modelo $ZF^d_{ijk} = \alpha_j + \epsilon_j + e_{ijk}$ dónde el nuevo factor ϵ_i afecta únicamente la varianza de los datos, no la media. Esto se ve reflejado en cluster de diferente amplitud, lo que se traduce en unos datos más variables con clusters más difusos como se ve en la figura **5-5**. En este caso se podría ver que el individuo simulado 5 presentaría una variabilidad mayor que el individuo simulado 1, también, inclusive estando en baja dimensión, habrían puntos (células) difíciles de clasificar manualmente.
Es necesario aclarar que tanto los datos mostrados en la figura **5-1** como los datos mostrados en la figura **5-5** fueron creados con la misma semilla y parámetros, por lo tanto lo único que cambia entre ambos conjuntos de datos es el nuevo error que fue incluido en la simulación 2, i.e., los residuales e_j son exactamente los mismos para ambos conjuntos de datos.

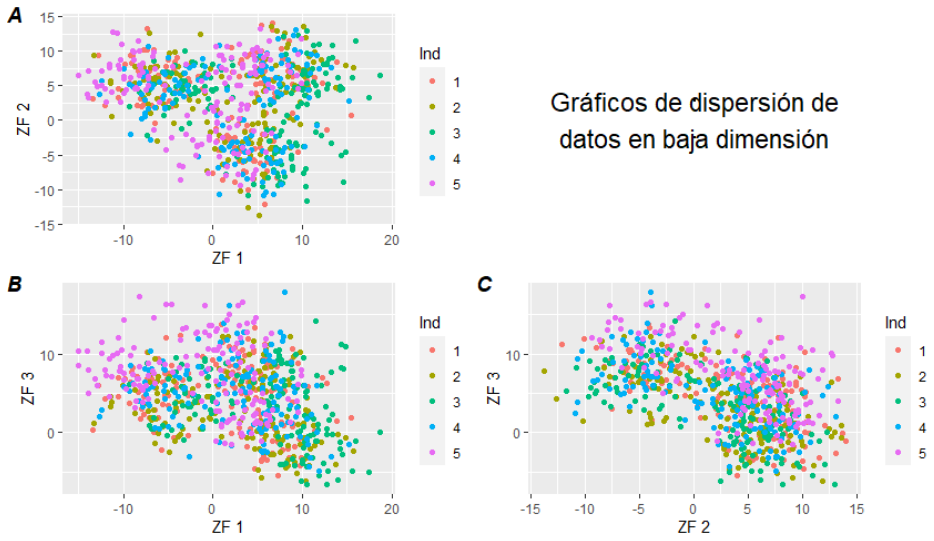

Figura 5-5: Representación gráfica de variables de baja dimensión Simulación 2

Las variables de alta dimensión se pueden visualizar en las figuras **5-6**, **5-7** y **5-8**. Para el caso de la simulación 2 se puede decir que se tiene lo mismo que en la simulación 1, salvo que el aumento en la variabilidad se traduce en grupos mas difíciles de identificar y correlaciones en general más bajas. En específico se ve un gran cambio para el escenario 1 (Figura **5-6**) pues donde la variabilidad del individuo fuera un poco mayor, sería complejo visualizar más de un grupo, inclusive en este caso, los histogramas identifican casi que solo una población. Lo cual quiere decir que este nuevo error que fue agregado propone un nuevo obstáculo para los algoritmos de reducción de dimensión y clasificación

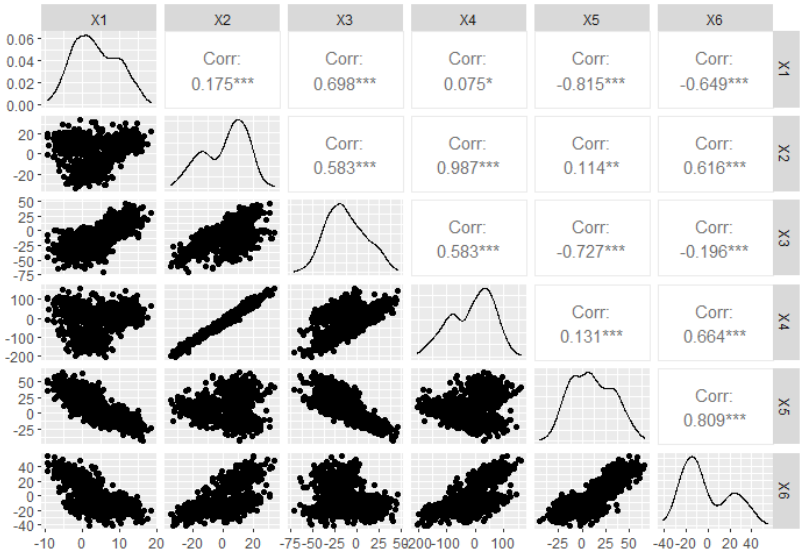

Figura 5-6: Representación gráfica de variables de alta dimensión (escenario 1) Simulación 2

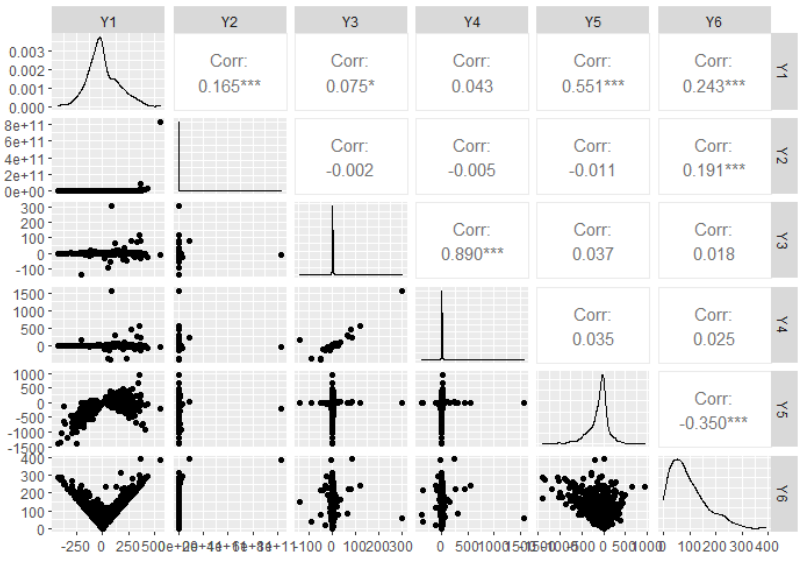

Figura 5-7: Representación gráfica de variables de alta dimensión (escenario 2) Simulación 2

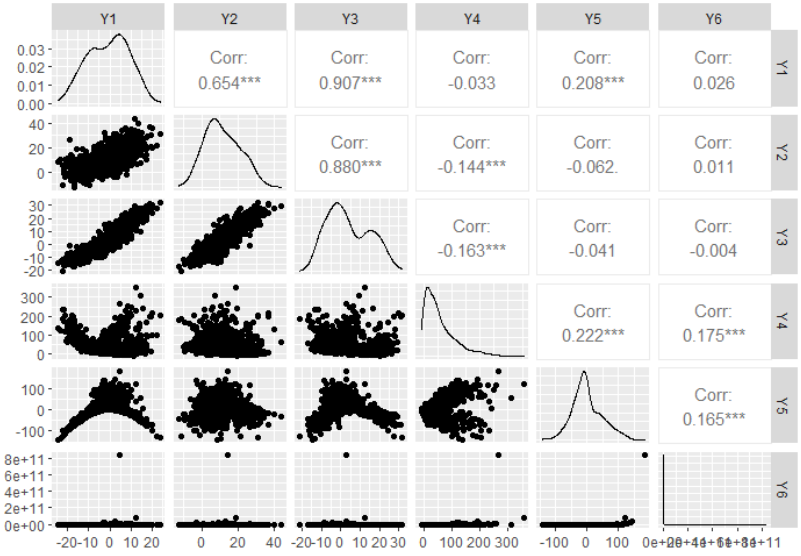

Figura 5-8: Representación gráfica de variables de alta dimensión (escenario 3) Simulación 2

5.2. Comparación de métodos

En esta sección se muestran los resultados de la comparación de los métodos Envelopes, PCA y T-SNE para el caso de los datos simulados.

5.2.1. Resultados gráficos de las simulaciones

Lo primero que hay que tener en cuenta es que los datos de baja dimensión generados tienen 3 componentes las cuales teóricamente tienen los mismos pesos. Por otro lado, los métodos que se usan en este trabajo intentan plasmar la mayor cantidad de información en una dimensión, haciendo que la importancia decaiga con las componentes adicionales. Esto puede entenderse fácilmente con el análisis de componentes principales ya que en este el porcentaje de varianza explicada de la tercera componente no puede ser mayor que el de la segunda, quien simultáneamente no puede tener un porcentaje de varianza explicada mayor al de la primera componente principal.
Es por eso que la primera componente de los datos de baja dimensión no tiene que coincidir con la primera componente detectada con los métodos. En la figura **5-9** se puede observar este fenómeno pues el envelope 1 si es el más correlacionado con la dimensión 1 (componente 1 de los datos de baja dimensión), pero para el caso del PCA la componente más correlacionada con la dimensión 1 es la componente principal 2, lo cual es en este caso una casualidad y depende netamente de la aleatoriedad de la simulación ya que las componentes iniciales son en teoría igualmente importantes, sin embargo, puede que debido a su generación una de estas incluya mayor o menor ruido.

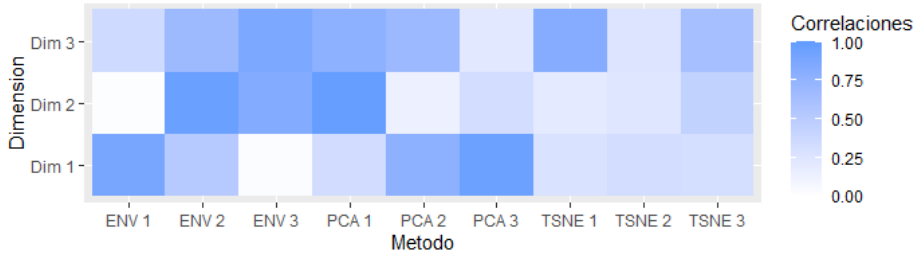

Figura 5-9: Correlaciones con las variables de baja dimensión originales (escenario 1)

Para esta simulación en particular (corriendo con la semilla 2021) se encuentra que bajo linealidad el T-SNE no se comporta a la par de los métodos lineales y esto no solo se puede ver en las correlaciones sino también en el mantenimiento de las estructuras de baja dimensión (figura **5-10**). Se observa que en las subfiguras **A, B, C** que hacen referencia a los datos originales, los envelopes y las componentes principales, las estructuras (formas y distancias) se mantienen de manera proporcional, es decir, son bastante similares; por otro lado en la subfigura D, que hace referencia a las componentes halladas por el T-SNE hubo una deformación de los clusters

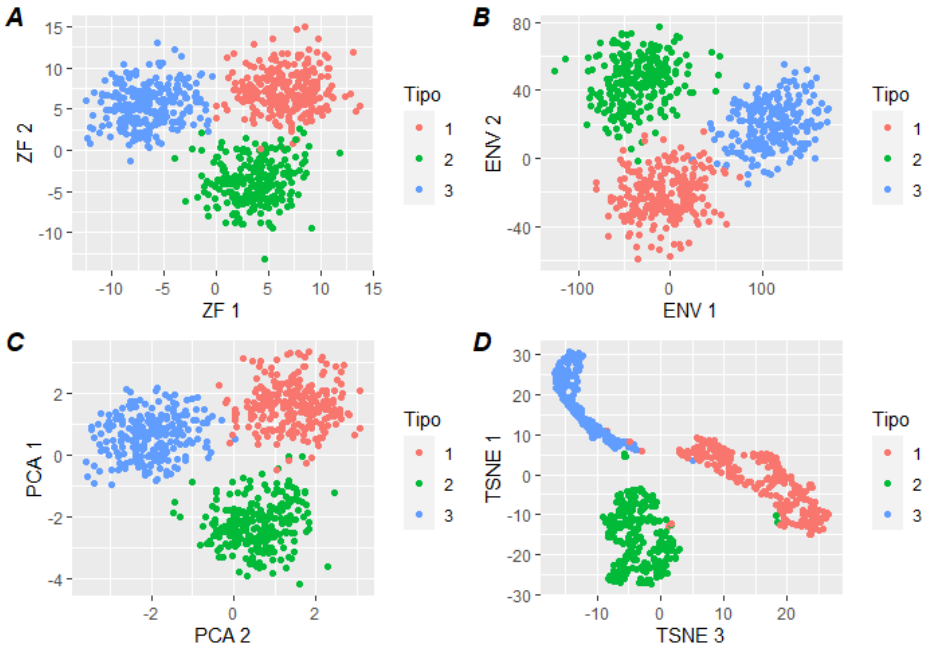

Figura 5-10: Comparación primeras dos dimensiones (escenario 1)

Para el caso del escenario 2, donde todas las combinaciones son no lineales sucede algo bastante particular. Si se observa la figura **5-11** se puede notar que el método que mejor capturó las componentes originales es el PCA, ya que tiene las correlaciones más altas con la dimensión 1 y 3; en la dimensión 2 el método que mejor logró capturar la información fueron los envelopes. No obstante, el T-SNE a pesar de no destacarse, sí logró capturar parte del comportamiento de la dimensión 1.

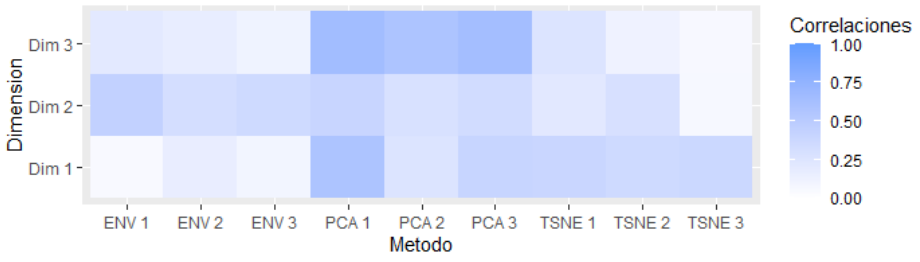

Figura 5-11: Correlaciones con las variables de baja dimensión originales (escenario 2)

Respecto al mantenimiento o captura de las estructuras de baja dimensión, ningún método logra la tarea de manera satisfactoria como si ocurría en el anterior escenario (ver figura **5-12**). No obstante, los envelopes, en específico el envelope 2, logró capturar la información necesaria para mantener una proporcionalidad de los grupos que se encontraban en baja dimensión. En este sentido, el PCA fué el peor método y el T-SNE a pesar de hacer un

trabajo no tan malo con las células simuladas tipo 3 y 2, en cuanto a las células tipo 1 deformó totalmente el grupo.

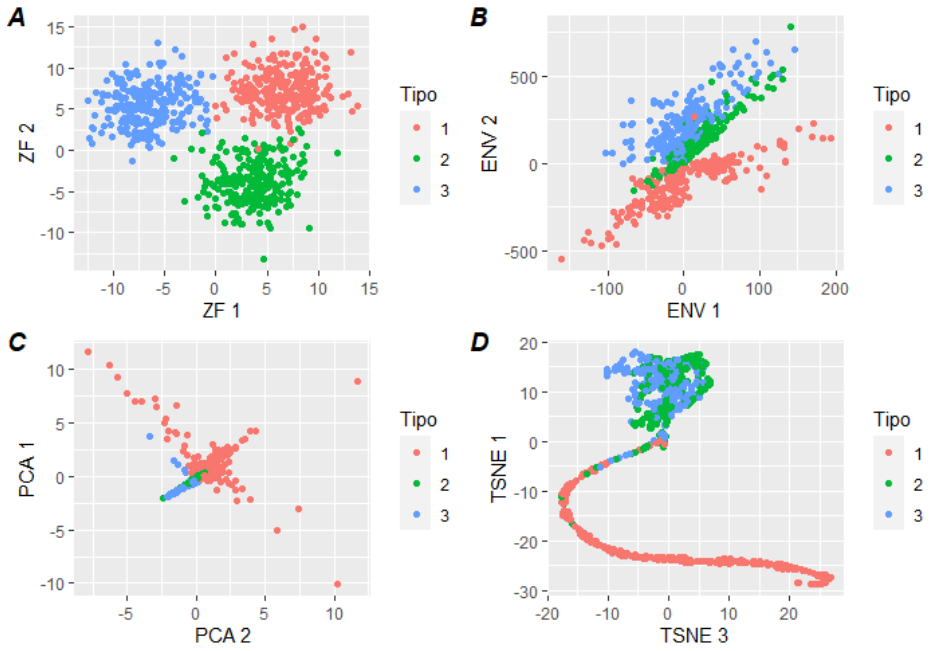

Figura 5-12: Comparación primeras dos dimensiones (escenario 2)

Aquí es posible ver la importancia del mantener las estructuras de baja dimensión, porque en SCseq no solo es importante clasificar bien, sino entender qué sucede con los datos y para eso es necesario mantener las estructuras originales. Adicionalmente, esta cualidad no solo va a aumentar la interpretación de los datos, sino que también va a permitir elevar las tasas de correcta clasificación, esto se ve claramente en la figura **5-12** D, ya que para un algoritmo va a ser totalmente difícil clasificar las células tipo 2 y 3, pues estas están una encima de la otra. Para los datos de la simulación 2 (con individuo como factor aleatorio) se encuentran resultados similares.

En primer lugar, para el escenario 1 (ver figura **5-13**), los envelopes y componentes principales parecen hacer un gran trabajo pues todas las dimensiones están correlacionadas con alguna componente obtenida en los métodos. En cuanto al T-SNE solo se encuentra una correlación mediana para la dimensión 3.

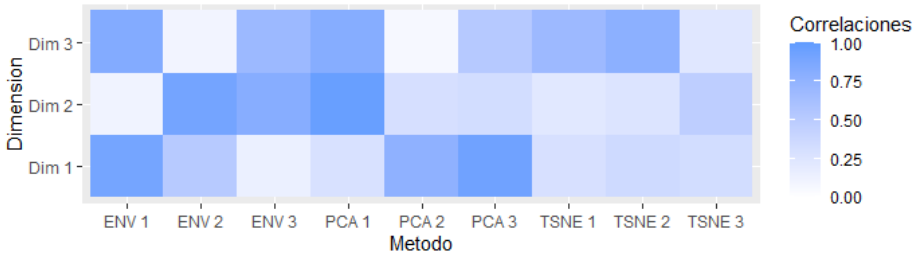

Figura 5-13: Correlaciones con las variables de baja dimensión originales (escenario 1) Simulación 2

Cabe recordar que las correlaciones no son lo más importante por lo tanto es de gran interés observar la figura **5-14**; para este caso se encuentra que el único método que guarda la proporcionalidad de las distancias i.e. mantiene las estructuras de baja dimensión, parece ser los envelopes. En el caso del PCA se forma un gran cluster y para el caso del T-SNE sucede lo mismo que en los casos anteriores, una deformación de los grupos.

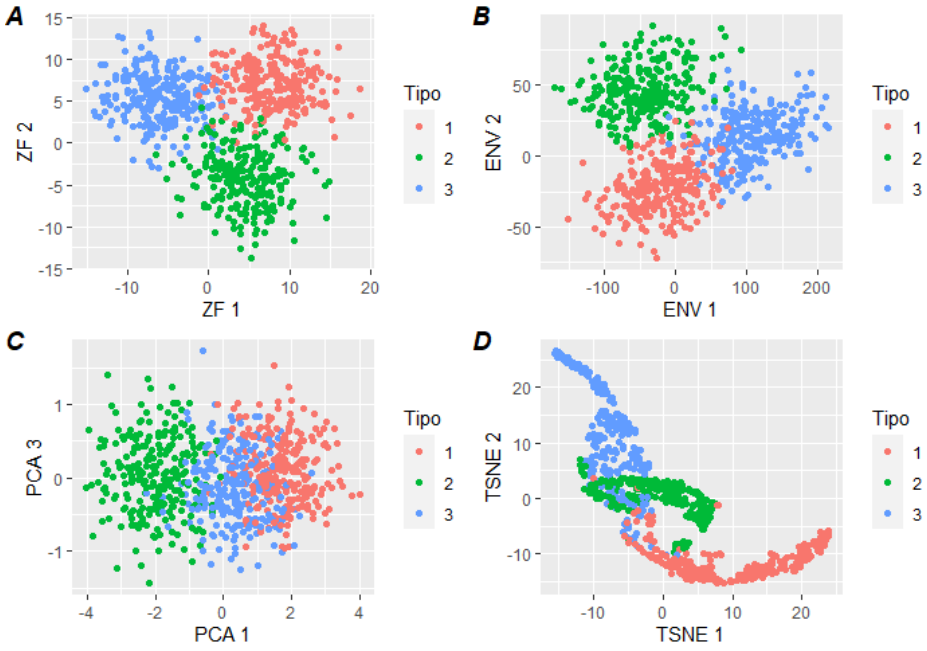

Figura 5-14: Comparación primeras dos dimensiones (escenario 1) Simulación 2

Para el segundo escenario, se encuentra lo que se esperaba, un buen desempeño del T-SNE. En la figura **5-15** se puede ver que las componentes originales están correlacionadas de manera baja con los envelopes, de manera media con las componentes principales y de manera medianamente alta con las componentes obtenidas con el T-SNE. En un principio se esperaba que este fuera el comportamiento en el escenario 2 ya que es el único método creado

para lidiar con la no linealidad.

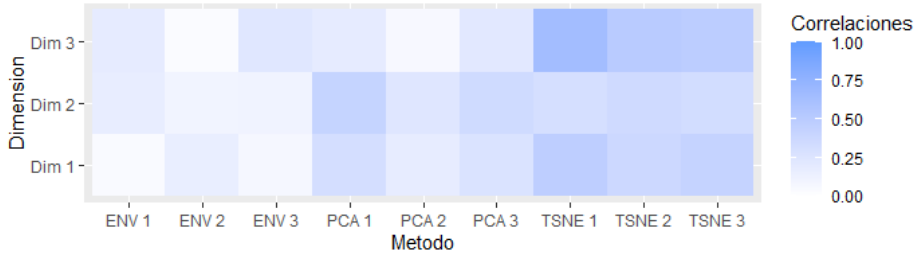

Figura 5-15: Correlaciones con las variables de baja dimensión originales (escenario 2) Simulación 2

En cuanto a mantener las estructuras de baja dimensión (figura **5-16**), ninguno de los métodos parece destacarse, los clusters en los envelopes parecen cruzarse, en las componentes principales existen datos mal clasificados como atípicos y finalmente para el T-SNE se ve una deformación total de los datos dónde por ejemplo las células 2 y 3 están totalmente mezcladas.

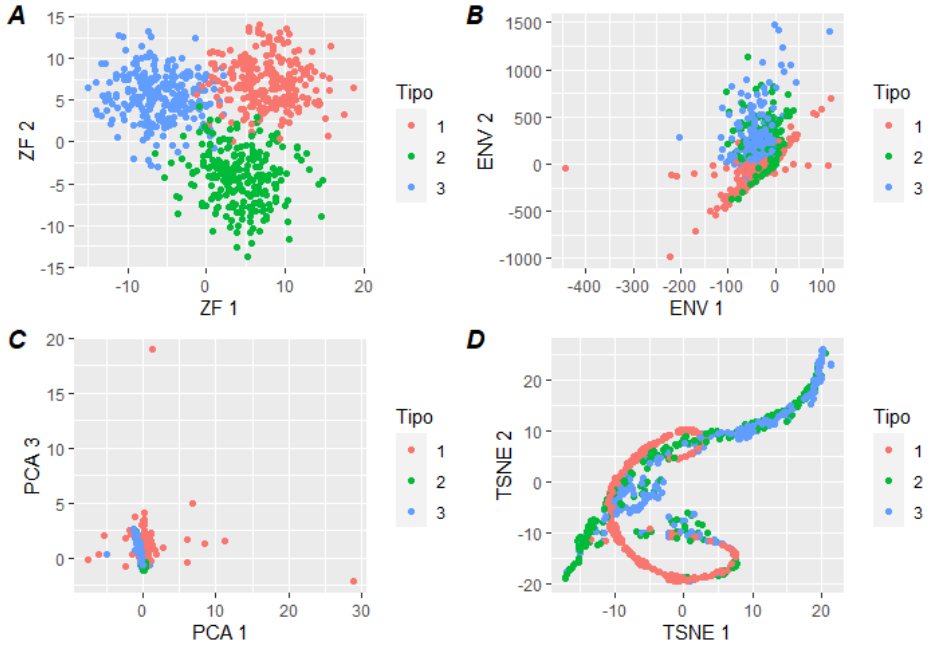

Figura 5-16: Comparación primeras dos dimensiones (escenario 2) Simulación 2

Finalmente para el escenario 3 se obtiene algo similar a lo obtenido en la simulación 1. En la figura **5-17** se observa que las dimensiones 1 y 2 son identificadas por los envelopes; las dimensiones 1,2 y 3 son en gran parte capturadas por la componente principal 1, mientras

que para el T-SNE se encuentra que las dimensiones 1 y 3 fueron identificadas en gran medida. Los tres métodos se comportan similar en cuanto a correlaciones se refiere, por lo tanto es necesario mirar si mantienen o no las estructuras de baja dimensión.

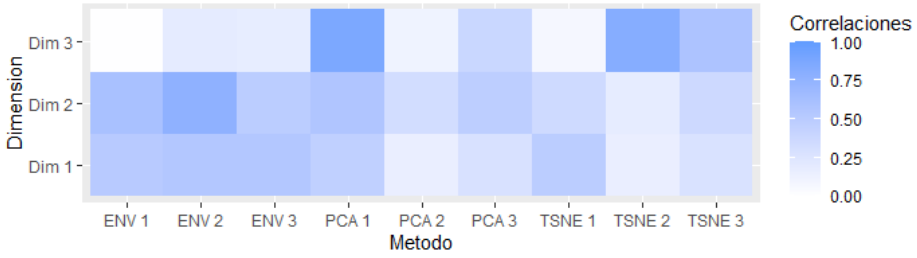

Figura 5-17: Correlaciones con las variables de baja dimensión originales (escenario 3) Simulación 2

En la figura **5-18** se pueden ver los diagramas de dispersión que se obtienen con los diferentes métodos. En este gráfico se puede observar que los tres métodos se destacaron en diferentes cosas. Los envelopes mantienen la variación de los datos mejor que los otros dos métodos. El PCA a pesar de mostrar un dato muy atípico en su primera componente mantiene la proporcionalidad de las distancias y las figuras o distribución global; mientras que el T-SNE hace una diferenciación de los clusters a pesar de deformarlos.

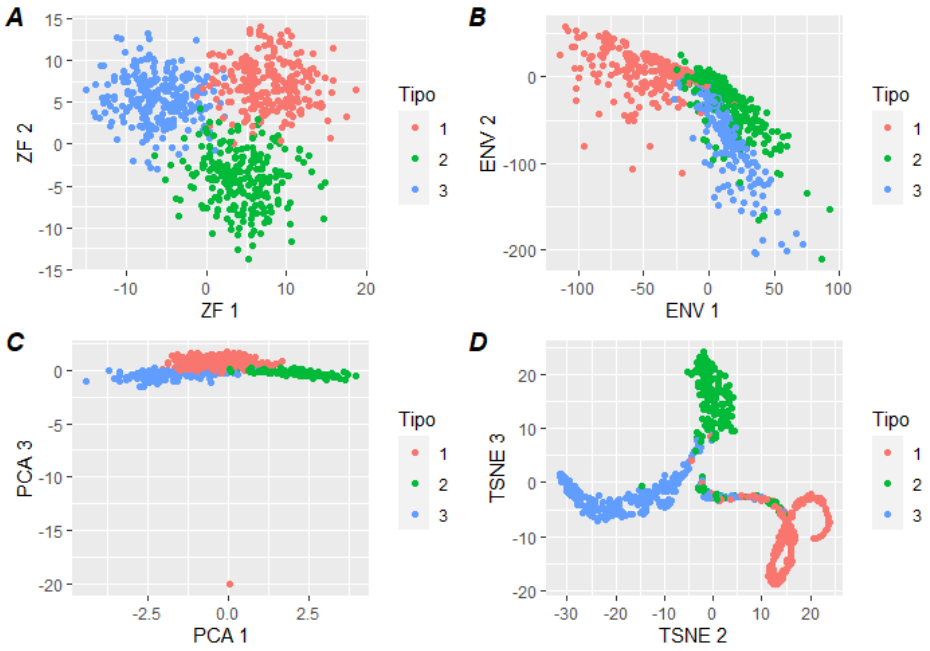

Figura 5-18: Comparación primeras dos dimensiones (escenario 3) Simulación 2

5.3. Máximas correlaciones y tasas de correcta clasificación

En la sección anterior se observaron las correlaciones de la estructura original con las estructuras estimadas por los métodos, sin embargo, los resultados obtenidos son particulares a la simulación determinada con la semilla 2021. Para entender mejor como se están comportando se corren 100 simulaciones de cada caso y escenario como se explicaba en la metodología y los resultados se pueden ver en las figuras **5-19** y **5-20** en su parte derecha.
Las componentes o dimensiones detectadas por los métodos no están relacionadas de manera secuencial con las componentes de los datos de baja dimensión; es por eso que los gráficos están divididos por escenario y subdivididos por método. Por cada método hay 3 boxplots, el primer boxplot resume al información de las correlaciones más altas halladas en valor absoluto entre las componentes del método y la dimensión 1 de la estructura de baja dimensión; el segundo boxplot resume la información de las correlaciones más altas halladas en valor absoluto entre las componentes del método y la segunda dimensión de la estructura de baja dimensión y el tercer boxplot resume consecuentemente las relaciones con la tercera dimensión. Es evidente que ningún método se destaca por tener mejores correlaciones que los demás, pues los tres son estadísticamente iguales en los diferentes casos. Es de suma importancia resaltar que en el escenario 2 de ambas simulaciones se encuentra que el T-SNE no se desempeña superior que los otros dos métodos; esto es interesante ya que seria de esperarse que un método no lineal se comporte mejor en todo sentido en escenarios no lineales y al menos en este aspecto no lo hace.
Otra característica importante que se alcanza a observar, es que la inclusión de relaciones no lineales hace que aumente la variabilidad del desempeño del método medido con correlaciones, pues en el escenario 1 la variabilidad de los diferentes métodos es la menor de los tres escenarios, en el segundo la mayor y en el tercer escenario es la intermedia. Asimismo la inclusión de relaciones no lineales también afecta la media de las máximas correlaciones.

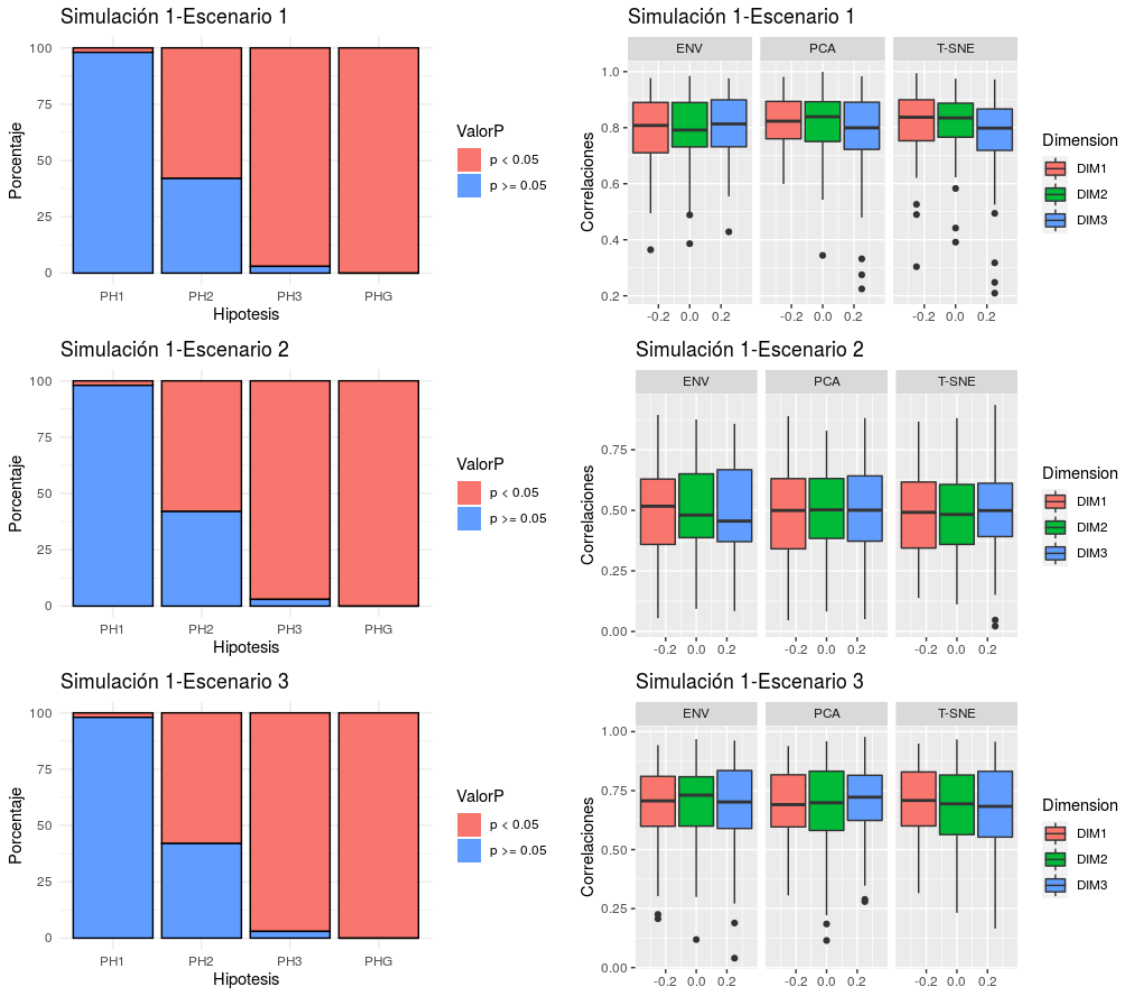

Figura 5-19: Evaluación de hipótesis y máximas correlaciones por método en Simulación 1

En cuanto al desempeño por tasas de correcta clasificación, fue medido a través de las hipótesis mencionadas en la metodología. Para cada simulación (conjunto de datos) se corrieron las 4 hipótesis, por lo tanto se resumieron los 100 valores p obtenidos en las figuras **5-19** y **5-20** parte izquierda.

Se observa que la hipótesis que hace referencia a si los 3 métodos tienen el mismo rendimiento en tasas de correcta clasificación, fue rechazada 100 de 100 veces. Para la simulación 1 (sin efecto aleatorio) son muy pocas las veces que se rechaza la hipótesis que la tasa de correcta clasificación de los envelopes es menor que la del PCA, mientras que la mayoría de veces sí se rechazan las hipótesis 2 y 3, lo que quiere decir que la tasa de los envelopes y del PCA es mejor que las del T-SNE.

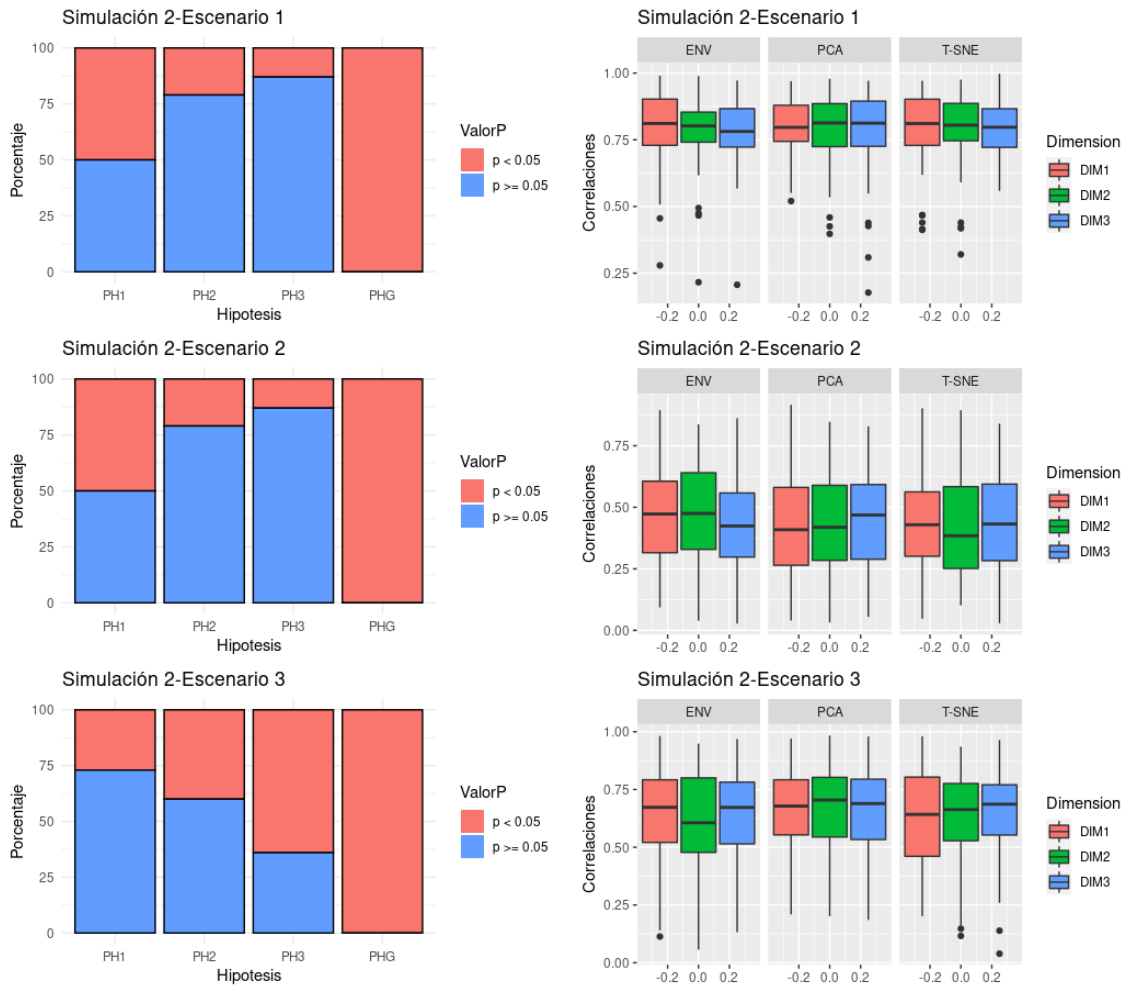

Figura 5-20: Evaluación de hipótesis y máximas correlaciones por método en Simulación 2

Para la segunda simulación (con factor aleatorio) se tiene que la hipótesis general (tasas iguales para los tres métodos) se rechaza 100 de 100 veces y se muestra una mejoría en cuanto al desempeño de los envelopes respecto al PCA; en el escenario 1 y 2 los envelopes no tienen mejores tasas que el T-SNE, al igual que sucede con el PCA que no tiene mejores tasas que el T-SNE. Sin embargo, en el escenario 3 el PCA tiene mejores tasas que el T-SNE.

5.4. Modelo propuesto

Ninguno de los tres métodos propuestos anteriormente incluye un efecto aleatorio, es decir, ninguno tiene en cuenta las diferentes dispersiones que pueden tener cada grupo, es por eso que en este trabajo se propone un modelo lineal sencillo de reducción de dimensión para incluir estos efectos aleatorios y tener buenos resultados según se entienden hasta este punto

los datos de secuenciación de RNA de células individuales.

$$Z = YC = XB + e \tag{5-1}$$

$Z_{N\times bd}$ hace referencia a los datos de baja dimensión (bd), $Y_{N\times ad}$ a los datos de alta dimensión(ad), $C_{ad\times bd}$ los coeficientes o cargas de cada variable en el momento de reducir la dimensión, $X_{N\times p}$ la matriz diseño, $B_{p\times bd}$ la matriz de efectos fijos y finalmente, $e_{N\times bd}$, los residuales estructurados del modelo.
Para llevar a cabo la estimación se supone que $e \sim N(0, \Sigma)$, lo cual nos deja una función de log-verosimilitud:

$$L = -\frac{bd}{2}\ln(2\pi) - \frac{bd}{2}\ln(|\Sigma|) - \frac{1}{2\sigma}tr[\Sigma^{-1}(YC - XB)(YC - XB)^T] \tag{5-2}$$

Para poder simplificar las cuentas es necesario adecuar los datos, i.e., organizarlos por individuos ya que este es el factor aleatorio que se quiere tener en cuenta.

$$Y_{N\times ad} = [Y_1^T|Y_2^T|\cdots|Y_I^T]^T \text{ , siendo } [Y_i^T]_{n_i\times ad} = [(y_1^T)^{(i)}|(y_2^T)^{(i)}|(y_{n_i}^T)^{(i)}]^T \tag{5-3}$$

Dónde $(y_j^T)^{(i)}$ es un vector fila que representa las mediciones de expresión para j-ésima célula en el i-ésimo paciente.
Asimismo, se puede particionar la matriz diseño

$$X = [X1^T|X_2^T|\cdots|X_I^T] \tag{5-4}$$

Como los individuos son independientes la matriz de covarianza Σ se puede suponer con la siguiente estructura

$$\Sigma = \left[\begin{array}{c|c|c|c} (I_{n_1}\sigma + J_{n_1}^T\theta) & 0 & \cdots & 0 \\ \hline 0 & (I_{n_2}\sigma + J_{n_2}^T\theta) & \cdots & 0 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \cdots & (I_{n_I}\sigma + J_{n_I}^T\theta) \end{array}\right] = \left[\begin{array}{c|c|c|c} S_1 & 0 & \cdots & 0 \\ \hline 0 & S_2 & \cdots & 0 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \cdots & S_I \end{array}\right] \tag{5-5}$$

Siendo I_{n_i} la matriz identidad de tamaño $n_i \times n_i$ y J_{n_i} una matriz de $n_i \times n_i$ llena de 1's.
Note que σ recoge la variación que existe entre células y θ la variación adicional que hay entre individuos.
Usando la formula de Sherman Morrison se obtiene que:

$$S_i^{-1} = \frac{1}{\sigma}I_{n_i} - \frac{\frac{\theta}{\sigma^2}I_{n_i}J_{n_i}I_{n_i}}{(1 + 1_{n_i}^T 1_{n_i}\frac{\theta}{\sigma})} = \frac{1}{\sigma}(I_{n_i} - \frac{\theta}{\sigma + n\theta}J) \tag{5-6}$$

Adicionalmente usando la identidad del determinante de Sylvester se obtiene:

$$|\Sigma| = |S_1| \times \cdots \times |S_I| = \prod_{i=1}^{I}\sigma^{n_i}|(I_{n_i} + 1_{n_i}1_{n_i}^T\frac{\theta}{\sigma})| = \prod_{i=1}^{I}\sigma^{n_i-1}(\sigma + n_i\theta) \tag{5-7}$$

Por lo tanto la log-verosimilitud se puede reescribir como:

$$L = -\frac{bd}{2}Ln(2\pi) - \frac{bd}{2}((N-I)\ln(\sigma) + \sum_{i=1}^{I}\ln(\sigma+n_i\theta)) - \frac{1}{2\sigma}\sum_{i=1}^{I} tr[S_i^{-1}(Y_iC - X_iB)(Y_iC - X_iB)^T] \tag{5-8}$$

De esta manera se pueden obtener la derivada parcial respecto a C

$$\begin{aligned}\frac{\partial L}{\partial C} &= \frac{-1}{2\sigma}\frac{\partial}{\partial C}\sum tr[S_i^{-1}(Y_iCC^TY_i^T - 2Y_iCB^TX_i^T + X_iBB^TX_i^T)] \\ &= \frac{-1}{2\sigma}\sum C^TY_i^TS_i^{-1}Y_i - B^TX_i^TS^{-1}Y_i = 0\end{aligned} \tag{5-9}$$

Por lo que se obtiene:

$$\begin{aligned}C^T\sum_i Y_i^TS_i^{-1}Y_i &= B^T\sum X_i^TS_i^{-1}Y_i \\ (\sum_i Y_i^TS_i^{-1}Y_i)C &= (\sum Y_i^TS_i^{-1}X_i)B\end{aligned} \tag{5-10}$$

Note que la anterior igualdad es equivalente a $Y^T\Sigma^{-1}YC = Y^T\Sigma^{-1}XB$ debido a que Σ^{-1} es una matriz particionada por bloques de cuyos bloques fuera de la diagonal son matrices de ceros.
Por otro lado, la derivada parcial de la log-verosimilitud con respecto a B esta dada por la siguiente ecuación:

$$\begin{aligned}\frac{\partial L}{\partial B} &= B^TX^T\Sigma^{-1}X - C^TY^T\Sigma^{-1}X \\ &= B^T\sum_i X_i^TS^{-1}X_i - C^T\sum_i Y_i^TS^{-1}X_i\end{aligned} \tag{5-11}$$

En general las matrices diseño pueden ser reducidas o construidas de manera tal que sean rango columna completo i.e $(X^T\Sigma^{-1}X)^{-1}$ existe. Si existen las I inveras de $(X_i^T\Sigma^{-1}X_i)$ es mejor usarlas por eficiencia computacional, sin embargo no es necesario, debido a su equivalencia con $(X^T\Sigma^{-1}X)$. Por lo tanto

$$\begin{aligned}\hat{B}^T &= C^T(\sum_i Y_i^TS_i^{-1}X_i)(X^T\Sigma^{-1}X)^{-1} \\ \hat{B} &= (X^T\Sigma^{-1}X)^{-1}(\sum_i X_i^TS_i^{-1}Y_i)C\end{aligned} \tag{5-12}$$

De donde se obtiene

$$\begin{aligned}(\sum_i Y_i^T S_i^{-1} Y_i)C &= (\sum_i Y_i^T S_i^{-1} X_i)B \\ &= (\sum_i Y_i^T S_i^{-1} X_i)(X^T\Sigma^{-1}X)^{-1}(\sum_i X_i^T S_i^{-1} Y_i)C \\ Y^T\Sigma^{-1}YC &= (Y^T\Sigma^{-1}X)(X^T\Sigma^{-1}X)^{-1}(X^T\Sigma^{-1}Y)C \\ 0 &= (Y^T\Sigma^{-1}Y - Y^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y)C \\ 0 &= [Y^T\Sigma^{-1/2}(I - \Sigma^{-1/2}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1/2})\Sigma^{-1/2}Y]C\end{aligned} \tag{5-13}$$

Nótese que $\Sigma^{-1/2}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1/2}$ es una matriz de proyección. Por lo tanto, $(I - \Sigma^{-1/2}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1/2})$ también es una matriz de proyección (proyecta en el espacio ortogonal), es decir, se busca una matriz C cuyas columnas sean la base de un subespacio donde la proyección de $Y^T\Sigma^{-1}Y$ coincida con la de $Y^T\Sigma^{-1/2}P_{X^*}\Sigma^{-1/2}Y$. $X^* = \Sigma^{-1/2}$ y P_{X^*} la matriz de proyección en el subespacio formado por las columnas de X^*
Si $P_{Y^*} = \Sigma^{-1/2}Y(Y^T\Sigma^{-1}Y)^+Y^T\Sigma^{-1/2}\Sigma^{-1/2}Y$
Entonces $\hat{C} = (Y^T\Sigma^{-1}Y)^+Y^T\Sigma^{-1}Y$ minimiza la ecuación ya que:

$$\begin{aligned}0 &\approx [Y^T\Sigma^{-1/2}(I - P_{X^*})\Sigma^{-1/2}Y]C = [Y^T\Sigma^{-1/2}(I - P_{X^*})\Sigma^{-1/2}Y](Y^T\Sigma^{-1}Y)^+Y^T\Sigma^{-1}Y \\ &\approx Y^T\Sigma^{-1/2}P_{Y^*}\Sigma^{-1/2}Y - Y^T\Sigma^{-1/2}P_{X^*}P_{Y^*}\Sigma^{-1/2}Y \\ &\approx Y^T\Sigma^{-1/2}[\Sigma^{-1/2}Y - P_{X^*}P_{Y^*}\Sigma^{-1/2}Y] = Y^T\Sigma^{-1/2}[\Sigma^{-1/2}Y - \hat{y}]\end{aligned} \tag{5-14}$$

Siendo $\hat{y}$ la estimación de $\Sigma^{-1/2}Y$.
Si $\Sigma^{-1/2}Y$ pertenece al subespacio conformado por $\Sigma^{-1/2}X$ y $\Sigma^{-1/2}Y$ (las columnas seleccionadas que transforman a baja dimensión) entonces $P_{X^*}P_{Y^*}\Sigma^{-1/2}Y = \Sigma^{-1/2}Y$ y por lo tanto la igualdad a 0 se cumple.
Para obtener a $\Sigma^{-1/2}$ no es necesario hacer ningún proceso computacionalmente costo, pues al ser una matriz particionada diagonal, basta con encontrar las matrices $S_i^{-1/2}$

$$S_i^{-1/2}S_i^{-1/2} = S_i^{-1} = \frac{1}{\sigma}(I - \frac{\theta}{\sigma + n\theta}J) \tag{5-15}$$

Es claro que $S_i^{-1/2}$ debe ser de la forma $\frac{1}{\sqrt{\sigma}}(I - xJ)$, luego

$$(I - xJ)(I - xJ) = (I - 2xJ + x^2JJ) = (I - 2xJ + nx^2J) = (I - \frac{\theta}{\sigma + n\theta}J) \tag{5-16}$$

Es decir solo hace falta resolver una ecuación cuadrática sencilla $x^2n - 2x + \frac{\theta}{\sigma+n\theta} = 0$ De dónde se obtiene $S_i^{-1/2} = \frac{1}{\sqrt{\sigma}}(I - \frac{1+\sqrt{1+\frac{n\theta}{\sigma+n\theta}}}{n}J)$

Finalmente se deriva la log-verosimilitud con respecto a los parámetros de varianza obteniendo las siguientes estimaciones:

$$\begin{aligned}\hat{\sigma} &= \frac{tr[J(YC-XB)(YC-XB)^T](tr[(YC-XB)(YC-XB)^T]-N\times bd)}{N^2\times bd^2}\\ \hat{\theta} &= \frac{tr[J(YC-XB)(YC-XB)^T](2\times N\times bd - tr[(YC-XB)(YC-XB)^T])}{N^3\times bd^3}\end{aligned} \tag{5-17}$$

Como la estimación de C, B depende de los parámetros σ, θ entonces toca realizar la maximización de log verosimilitud iterativamente, sin embargo, esto no es un gran impedimento pues los parámetros desconocidos en el algoritmo son solo σ y θ.

B es un estimador insesgado puesto que

$$E[(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}YC] = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}E[YC] = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}XB = B \tag{5-18}$$

Su varianza es:

$$\begin{aligned}V(\hat{B}) &= (X^T\Sigma^{-1}X)^{-1}(X^T\Sigma^{-1})V[YCC^TY^T]\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}\\ &= (X^T\Sigma^{-1}X)^{-1}(X^T\Sigma^{-1})\Sigma\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} = (X^T\Sigma^{-1}X)^{-1}\end{aligned} \tag{5-19}$$

5.5. Datos de oligodendrogliomas

Es claro que los procesos que ocurren en la secuenciación de células individuales es un tema de constante investigación, por lo tanto los escenarios descritos anteriormente no necesariamente son similares a los datos reales. Es por esto que surge el interés de aplicar los diferentes métodos a un conjunto de datos; para este caso se toman los datos de oligodendrogliomas que ya han sido analizados en [4] y [14].
Los datos presentados a continuación, pertenecen a una muestra de 6 individuos con oligodendrioglioma, el cual es un tumor primario del sistema nerviosos central. Existen dos grados, el bajo implica un crecimiento y expansión lento al tejido normal cercano, usualmente se forma varios años antes de que sea detectado. El grado alto, o maligno (canceroso) hace referencia a cuando esta expansión se hace de manera rápida, se suele llamar oligodendroglioma anaplásico. El estudio supone que este tipo de cáncer se debe a temas genéticos y mutaciones de los genes IDH1 o IDH2. Para verificar estos supuestos y entender la formación de estos gliomas, se tomaron datos de 6 individuos notados como MGH, por el hospital donde fueron obtenidos los datos, que tenían un oligodendroglioma de grado II, es decir bajo. La medición se hizo a través del método de secuenciación de RNA en células unitarias. En total se analizaron 4347 células diferentes.

Cabe aclarar que los autores reportan que no hubo ningún procedimiento estadístico involucrado en el calculo del tamaño de muestra, ni se aleatorizaron los experimentos.
Los datos presentados a continuación no son de conteos debido a que los niveles de expresión presentados en la base de datos hace referencia a $E_{ij} = \log_2(TPM_{ij}/10 + 1)$ siendo TPM_{ij} la transcripcion por millon del gen i en la muestra j, lo que significa que estos datos están normalizados para hacer las muestras comparables y eliminar el sesgo de profundidad de secuenciación y longitud del gen.
Para obtener estas transcripciones por millón es necesario primero dividir el número de conteos por la longitud del gen (kilobases), contar todos las lecturas por millones de kilobases y finalmente dividir estos números por un millón.
A pesar que los datos presentados ya tenían un control de calidad. Así como se mencionó en la metodología, para este trabajo se realizó un filtro adicional quedando únicamente 4347 células con lecturas en 963 genes, en vez de 4347 células con lecturas en 23703 genes.

5.5.1. Análisis descriptivo

En la figura **5-21** se puede observar la distribución de la expresión promedio de los genes. Es de suma importancia notar que los boxplot son muy similares (las distribuciones no son exactamente iguales). Por lo tanto no es necesario hacerle ninguna normalización o transformación adicional a los datos.

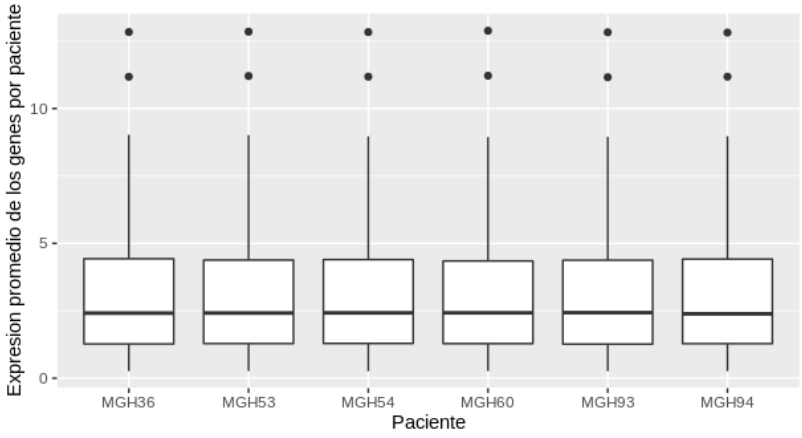

Figura 5-21: Distribución de la expresión de los genes promedio por paciente

Por otro lado, en la figura **5-22** se observa que al promediar las expresiones de los genes por células se obtienen resultados casi idénticos en los 6 pacientes, lo cual va a facilitar el análisis de los datos. Anteriormente, se mostraba cómo los 6 pacientes tenían una expresión promedio casi idéntica, esto tiene sentido pues los 6 tienen la misma condición y el objetivo no es identificar la diferencia entre pacientes; en este caso se busca identificar la diferencia en

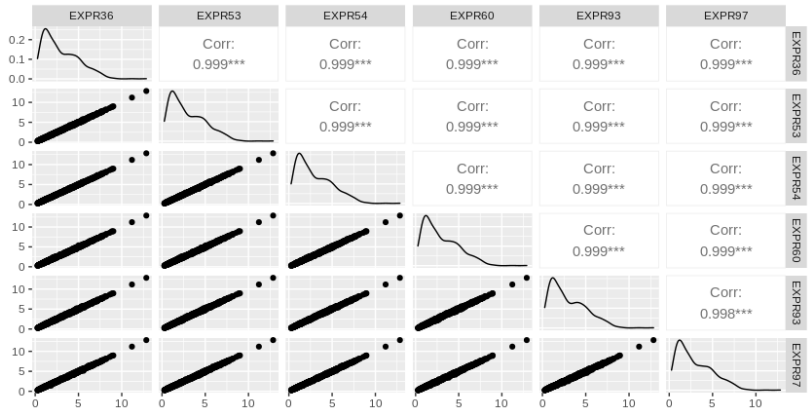

Figura 5-22: Relación entre muestras

la expresión entre las células, las cuales en general tienen coeficientes de expresión similares. Sin embargo, en la figura **5-23** se puede observar que hay algunas células que tienen expresión diferente al resto de las células (líneas blancas o de colores verde claro). Esto podría ser un indicio de que si hay una posible expresión diferencial entre los diferentes tipos de células.

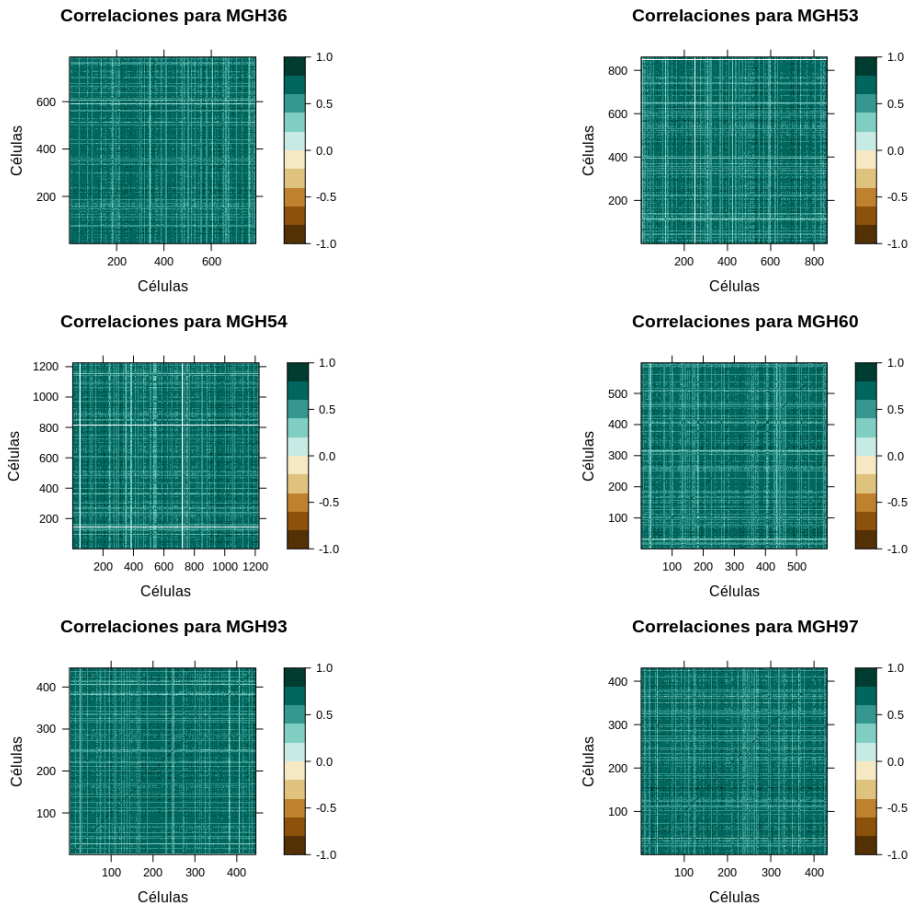

Figura 5-23: Correlación entre células

5.5.2. Reducción de dimensión

Al ver que los datos están bien, en el sentido que los pacientes y genes son comparables, se procede a hacer la reducción de dimensión. Los datos no son independientes como era de esperarse, así que hacer uso de los envelopes, T-SNE y PCA no es completamente adecuado (esto se puede ver por los altos coeficientes de autocorrelación y autocorrelación parcial. Ver la figura **9-4** y **9-5** en la sección de anexos). En primer lugar es importante resaltar que las primeras tres componentes principales coinciden con los envelopes, esto es un indicio de que las variables expuestas en la matriz diseño no están aportando a la reducción de dimensión. Cabe aclarar que esto no significa que las variables no sean de utilidad, sino que no lo son al momento de reducir la dimensión. Es decir no hay ruido innecesario.
Una vez se hace la reducción de dimensión es posible hacer una clasificación. El número de clusters, es decir, de poblaciones que se van a identificar se da una vez se ven las estructuras de baja dimensión. Se usa el algoritmo EM y se eligen 3 clusters únicamente. Debido a que con 3, las clasificaciones con diferentes semillas son más consistentes.

En un principio parece que solo hay dos poblaciones, es decir, que los datos deberían clasificarse en 2 grandes grupos, sin embargo, la elección de 3 se hace evidente después de tener en cuenta la información de la profundidad de la secuenciación.
Dependiendo del método adoptado anteriormente para reducir la dimensión de los datos, se van a tener diferentes componentes (componentes sacadas por PCA, TSNE o el modelo propuesto) y dependiendo de estas componentes, los resultados de la clasificación van a variar. Para los resultados que se muestran a continuación (figura **5-24**) se utiliza la clasificación desprendida de la reducción basada en el modelo propuesto. Sin embargo, si se quisieran observar los resultados obtenidos con diferentes métodos, es decir, los resultados que se obtendrían con clasificaciones basadas en PCA y TSNE, estos se pueden observar en las figuras **9-6** y **9-7** ubicadas en los anexos.

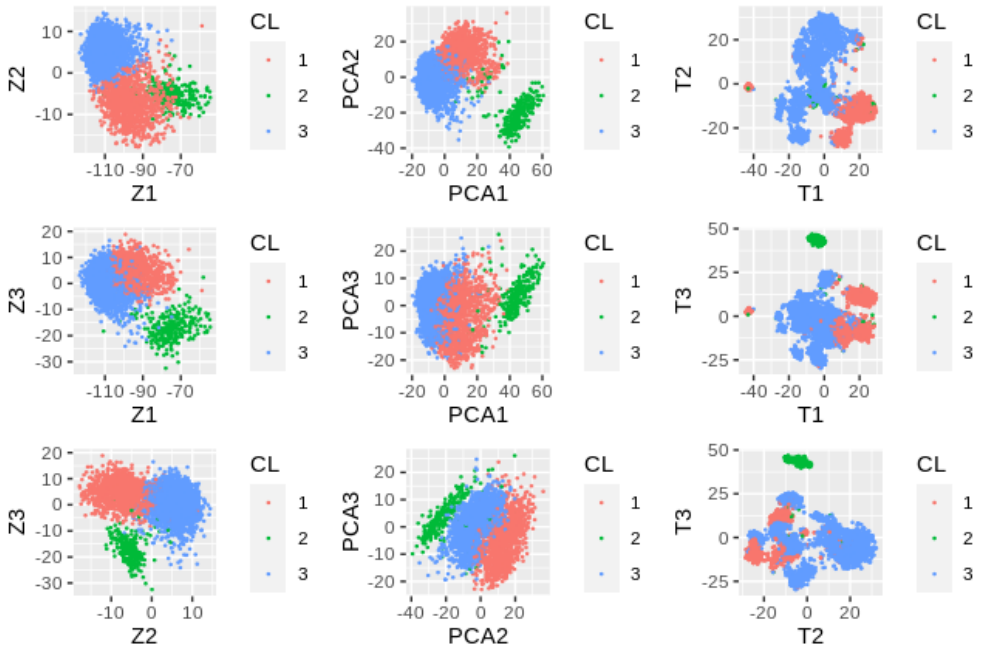

Figura 5-24: Datos en baja dimensión para los distintos modelos

Utilizando la clasificación mencionada se encuentra una gran separación de la nube de datos representada con color verde para el modelo propuesto, PCA y Envelopes, para el T-SNE este cluster o subconjunto de datos no es necesariamente identificado, solo se puede observar si se analiza la tercera componente. En cuanto a los subconjuntos denotados por los colores azul y rojo, estos se diferencian pero su discrepancia no es tan grande para los 4 métodos, de hecho, un pequeño porcentaje de estos datos va a variar su clasificación según el método y semilla del proceso aleatorio que se utilice. Es posible visualizar de mejor manera los 3 subgrupos si se tiene en cuenta la profundidad de la secuenciación (ver figura **5-25**). Cabe recordar que los datos presentados son de 6 pacientes divididos en 2 grupos con profundidad

de secuenciación diferente.

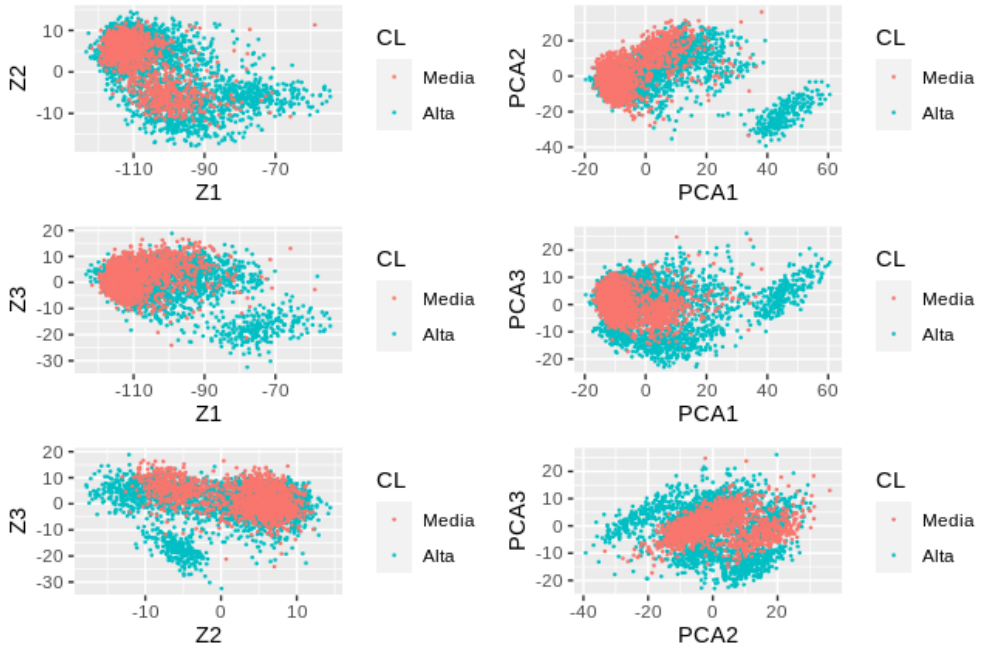

Figura 5-25: Visualización de la profundidad de secuenciación del experimento en los datos de baja dimensión

En la figura **5-25** también se resalta el hecho de que es necesario hacer uso de una secuenciación con profundidad alta para poder visualizar el tercer cluster que en la figura **5-24** se observa en verde.
Por lo mencionado anteriormente se decide reclasificar las muestras tomando por separado los datos de alta y media profundidad. Para el conjunto de datos de profundidad alta, se seleccionan 3 clusters a clasificar para que la única diferencia con los datos de profundidad media sea el cluster que se observaba en verde en la figura **5-24**. Para los datos de profundidad media se seleccionaron únicamente 2 clusters.
Los datos pertenecientes a individuos cuya profundidad de la secuenciación fue mayor están mostrados en la figura **5-26**, nótese que no solo se encuentra el cluster adicional del que ya se ha hablado, sino que también gracias a la profundidad de secuenciación se alcanzan a formar clusters o agrupaciones adicionales. En las componentes principales se alcanzan a diferenciar por la sobre-posición de los grupos, no obstante, estos clusters también pueden ser visualizados en las componentes de baja dimensión resultantes del modelo propuesto, con la diferencia de que no forman clusters sino que forman una franja horizontal que puede ser visualizada en la figura **5-26** entre la componente de baja dimensión Z3 o Z1 y a lo largo de la dimensión Z2.

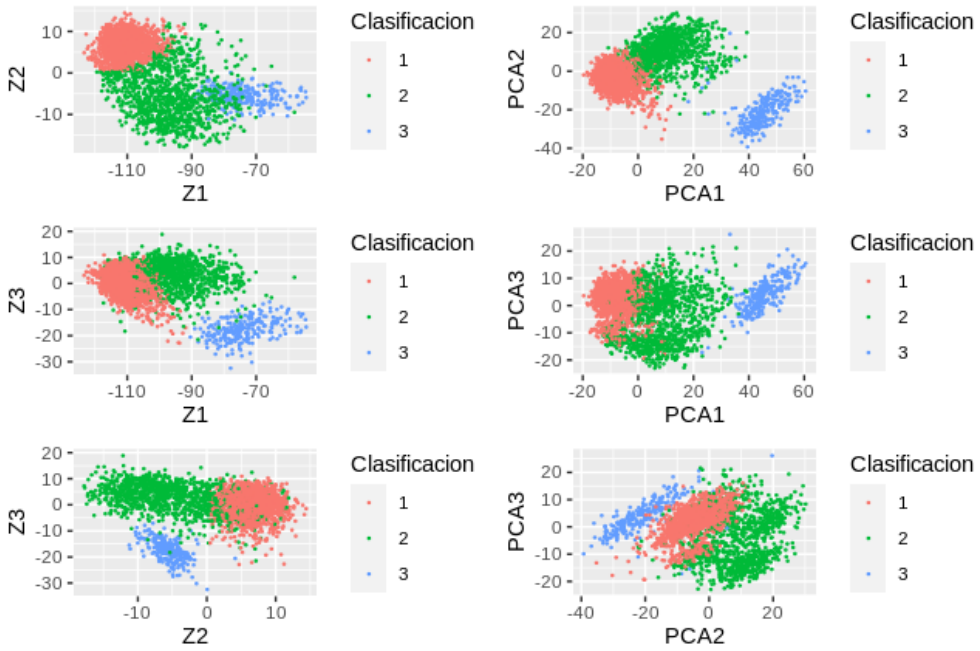

Figura 5-26: Clasificación de datos de baja dimensión en individuos con profundidad de secuenciación alta

5.5.3. Enriquecimiento

Una vez se hace el proceso de reducción de dimensión, lo primero que se debe hacer es mirar qué genes son los más importantes para no perder información en el proceso. Con estos genes es que se pretende entender los procesos biológicos relacionados con los oligodendrogliomas a través de la herramienta DAVID que proporcionará vías metabólicas importantes.

Identificación de Genes influyentes

Esto puede hacerse de diferentes maneras, no obstante, para fines de simplicidad, en este trabajo se propone realizar esta selección de genes a través de la distancia de Mahalanobis. Es importante recordar, que como las variables están estandarizadas desde antes de reducir la dimensión, entonces las filas con los coeficientes más altos en valor absoluto de la matriz C representaran las variables más importantes, i.e., los genes influyentes para identificar estas poblaciones de células y las filas con coeficientes más cercanos a cero, identificaran genes innecesarios a la hora de encontrar o identificar estas poblaciones de células.
La distancia de Mahalanobis permite identificar las variables (genes) que se encuentran más alejados de la media, es decir, los más importantes para construir cada una de las componentes de baja dimensión encontradas. En la figura **5-27** se pueden observar en color rojo los genes seleccionados de acuerdo a la atipicidad (o lejanía a la media) de sus cargas para las tres variables de baja dimensión. El valor para seleccionar los genes en este caso fue

$d_i > 6,8$ es decir, que la distancia de Mahalanobis de las cargas de un gen a las medias de las cargas, fuera de al menos 6.8. La elección de este valor se basa en que se querían hallar el 10 % de los genes más influyentes, lo cual daría un resultado de 96 genes. Sin embargo, para mayor simplicidad se selecciona este 6.8 ya que con este valor se encontraban los $100 \sim 10\,\%$ genes más influyentes e importantes para la reducción de dimensión.
Esta decisión de que porcentaje de genes más influyentes elegir, depende completamente del contexto ya que no hay una regla general establecida debido a que en algunos conjuntos de datos, puede llegar a ser únicamente un gen el que determina el proceso biológico y en otros casos, puede que no exista ningún gen con mayor importancia. En este caso, para entender el proceso biológico serían necesarios todos los genes.

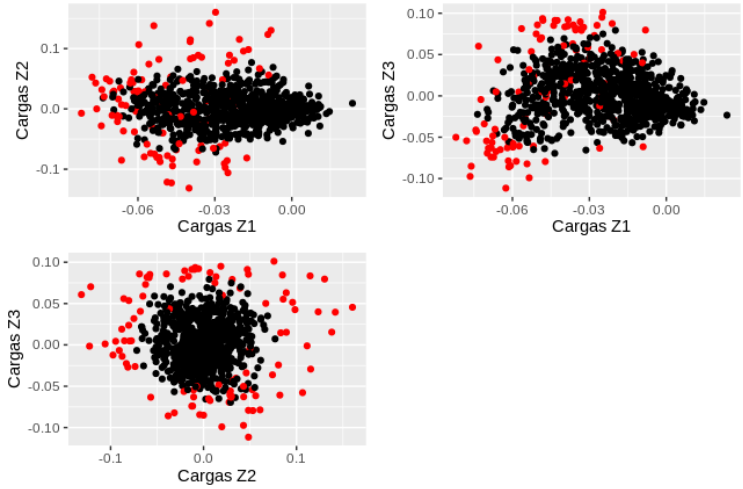

Figura 5-27: Distancia de Mahalanobis para las cargas de los genes

Los cinco genes con distancias más altas se pueden observar en la tabla **5-1**. Los 100 genes más importantes están en las tablas **9-1** , **9-2** y **9-3** ubicadas en los anexos.

Orden	Distancia	Gen
1	26.7603932996252	'ACBD6'
2	23.1739857279793	'ACN9'
3	20.970220832096	'ANKLE2'
4	20.7139853573505	'ARHGAP10'
5	20.5459156226671	'ARMCX5-GPRASP2'

Tabla 5-1: Genes con mayores distancias de Mahalanobis

Estos genes se traducen y se analizan por medio de la herramienta DAVID (Database for Annotation, Visualization and Integrated Discovery) obteniendo agrupaciones de estos genes destacados por algún motivo, proteínas relacionadas, enfermedades, pacientes, etc. Los

resultados para estos genes se puede visualizar en la tabla **5-2**.
Las agrupaciones encontradas se basan en las relaciones que tienen los grupos de genes G1, G2, G3, G4, G5 y G6.
G1={ACBD6, ANKRD34A, ANKLE2, ANKRD20A4, ANKRD30A, ANKRD30BL, ANKRD23, AGAP11, AGAP4, ANKEF1, ANKRD50}
G2={ ACBD6, ANKRD34A, ANKRD20A4, ANKRD30A, ANKRD23, AGAP11, AGAP4, ANKEF1, ANKRD50}
G3={ ANKRD30BL}
G4={AQP12B, AQP10, AQP8, AQP4, AQP2}
G5={ AQP10, AQP8, AQP4, AQP2}
G6={ADH4, ADH1A, ADH6}
Cada grupo parece tener relaciones muy marcadas, el primero cluster de genes conformado por los primeros 3 grupos (G1, G2 y G3) están relacionados con la repetición de ANK. El cluster 2 conformado por los grupos G4 y G5 hacen referencia a algo relacionado con el transporte de agua, y finalmente, el tercer cluster tiene relación aparente con alcohol deshidrogenasa.
En cuanto a los clusters y sus términos asociados, estos últimos tienen relacionado un valor p que hace referencia a un test exacto de Fisher que afirma que los genes de la lista que fueron identificados por su importancia en el modelo, tienen una relación con los términos expuestos por algo diferente al azar. Para los primeros dos clusters, los valores p ajustados de los términos son menores a un $\alpha = 0{,}02$. No obstante, para el tercer cluster los términos no son significativos, por lo cual se podría presentar el caso donde no salga nada interesante del análisis de estos genes en este trabajo.
En términos generales se podría decir que los clusters están representados por los conjuntos de genes G1, G4 y G6, respectivamente.
Para ver como esta representada su importancia en los procesos biológicos que se quieren analizar en este trabajo, se reconstruyen las estructuras de baja dimensión usando únicamente los genes de los conjuntos G1, G4 y G6 respectivamente. Esto quiere decir que C es una matriz $ad \times bd$ cuyas filas son 0's salvo por las filas que pertenecen a los genes de cada conjunto.
Si $C = [c_1^T|c_2^T|\dots|c_{ad}^T]^T$ siendo c_i vectores fila de tamaño $bd = 3$ que representan las cargas de el i-ésimo gen en las diferentes componentes de baja dimensión.
Entonces $C_{G_1} = [0^T|0^T|\dots|c_{j_1}^T|0^T|\dots|0^T|c_{j_n}^T|0^T|\dots|0^T]^T$ es la matriz de las cargas del conjunto G1, cuyas únicas filas no nulas son las filas $j_1, \dots, j_{11}$, que tienen los valores $\{c_{j_1}, \dots, c_{j_{11}}\}$ que hacen referencia a las cargas de los genes que pertenecen a G1. De la misma manera se definen C_{G4} y C_{G6}. De esta manera, la reconstrucción de las componentes de baja dimensión con cada conjunto de datos es simplemente $Z_{Gi} = YC_{Gi}$

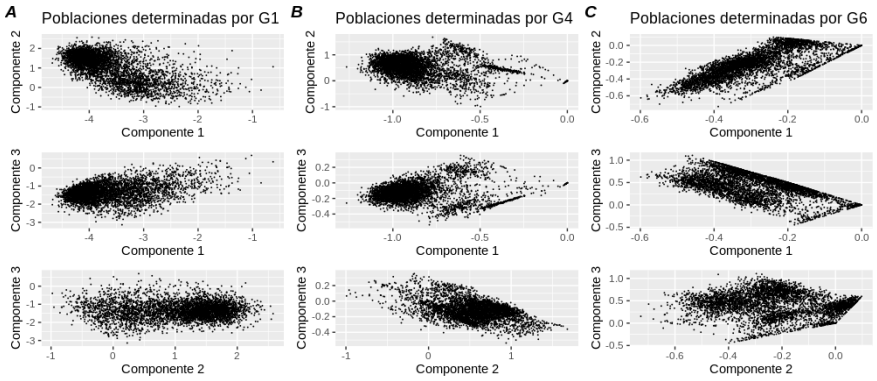

Figura 5-28: Replicación de estructuras de baja dimensión usando únicamente los genes de los conjuntos G1, G4 y G6

En la figura **5-28** se observan las reconstrucciones de las componentes de baja dimensión con base en los conjuntos de genes G1, G4 y G6. De aquí se puede observar que el conjunto G1 es suficiente para la construcción y diferenciación de 2 poblaciones, es decir, que con los datos de los genes pertenecientes a G1 se puede llegar a entender las poblaciones que se conformaban a lo largo de la componente de baja dimensión 2 presentada en la figura **5-26**. Si llegase a haber un proceso biológico relacionado con esta componente de baja dimensión 2, un primer paso sería entender que poblaciones o procesos biológicos pueden formarse en el marco de los genes pertenecientes a G1.
Asimismo se puede observar en las imágenes B y C de la figura **5-28** que la reconstrucción de los datos con base en los conjuntos de genes G4 y G6 permite identificar 6 posibles poblaciones de células adicionales, esto debido a que las 3 poblaciones que se logran visualizar en la imagen B, que hacen referencia a la reconstrucción basada en G4, no necesariamente son las mismas visualizadas en la imagen C, pues la separación entre las poblaciones se da en diferentes componentes de baja dimensión.

Annotation Cluster 1	Enrichment Score: 5.947616095807349							
Category	Term	Count	PValue	Genes	List Total	Pop Hits	Pop Total	Bonferroni
UP$_{KEYWORDS}$	ANK repeat	11	1.53E-7	G1	87	264	20581	2.49E-5
INTERPRO	IPR002110:Ankyrin repeat	11	2.27E-7	G1	85	255	18559	4.37E-5
INTERPRO	IPR020683:Ankyrin repeat-containing domain	11	3.24E-7	G1	85	265	18559	6.23E-5
PFAM	PF12796:Ankyrin repeats (3 copies)	10	1.77E-6	G2 y G3	83	239	17475	1.56E-4
UP$_{SEQFEATURE}$	repeat:ANK 1	9	1.0E-5	G2	87	244	20063	0.00285
UP$_{SEQFEATURE}$	repeat:ANK 2	9	1.03E-5	G2	87	245	20063	0.0029
Annotation Cluster 2	Enrichment Score: 5.655106862340335							
Category	Term	Count	PValue	Genes	List Total	Pop Hits	Pop Total	Bonferroni
UP$_{SEQFEATURE}$	short sequence motif:NPA 1	5	3.05E-7	G4	87	14	20063	8.56E-5
UP$_{SEQFEATURE}$	short sequence motif:NPA 2	5	3.05E-7	G4	87	14	20063	8.56E-5
GOTERM$_{MFDIRECT}$	GO:0015250 water channel activity	5	6.63E-7	G4	77	16	16881	1.42E-4
INTERPRO	IPR000425:Major intrinsic protein	5	6.82E-7	G4	85	16	18559	1.31E-4
PFAM	PF00230:Major intrinsic protein	5	7.85E-7	G4	83	16	17475	6.91E-5
INTERPRO	IPR023271:Aquaporin-like	5	8.88E-7	G4	85	17	18559	1.71E-4
GOTERM$_{BPDIRECT}$	GO:0006833 water transport	5	1.78E-6	G4	77	20	16792	7.62E-4
INTERPRO	IPR022357:Major intrinsic protein, conserved site	4	1.05E-5	G5	85	10	18559	0.002
GOTERM$_{MFDIRECT}$	GO:0015254 glycerol channel activity	4	1.87E-5	G5	77	12	16881	0.004
GOTERM$_{BPDIRECT}$	GO:0009992 cellular water homeostasis	4	2.47E-5	G5	77	13	16792	0.0105
GOTERM$_{BPDIRECT}$	GO:0015793 glycerol transport	4	2.47E-5	G5	77	13	16792	0.0105
Annotation Cluster 3	Enrichment Score: 2.3900620916188355							
Category	Term	Count	PValue	Genes	List Total	Pop Hits	Pop Total	Bonferroni
GOTERM$_{MFDIRECT}$	GO:0004024 alcohol dehydrogenase activity, zinc-dependent	3	2.97E-4	G6	77	6	16881	0.0615
GOTERM$_{MFDIRECT}$	GO:0004022 alcohol dehydrogenase (NAD) activity	3	4.14E-4	G6	77	7	16881	0.0848
INTERPRO	IPR002328:Alcohol dehydrogenase, zinc-type, conserved site	3	7.14E-4	G6	85	9	18559	0.1281
GOTERM$_{BPDIRECT}$	GO:0006069 ethanol oxidation	3	0.0013	G6	77	12	16792	0.4266
UP$_{SEQFEATURE}$	metal ion-binding site:Zinc 1; catalytic	3	0.0016	G6	87	14	20063	0.3631
INTERPRO	IPR013149:Alcohol dehydrogenase, C-terminal	3	0.0023	G6	85	16	18559	0.3612
INTERPRO	IPR013154:Alcohol dehydrogenase GroES-like	3	0.0023	G6	85	16	18559	0.3612
INTERPRO	IPR020843:Polyketide synthase, enoylreductase	3	0.0023	G6	85	16	18559	0.3612
PFAM	PF00107:Zinc-binding dehydrogenase	3	0.0025	G6	83	16	17475	0.1978
PFAM	PF08240:Alcohol dehydrogenase GroES-like domain	3	0.0025	G6	83	16	17475	0.1978
INTERPRO	IPR002085:Alcohol dehydrogenase superfamily, zinc-type	3	0.0029	G6	85	18	18559	0.4334
INTERPRO	IPR011032:GroES-like	3	0.004	G6	85	21	18559	0.53855
UP$_{SEQFEATURE}$	binding site:NAD	3	0.0225	G6	87	54	20063	0.9984
UP$_{SEQFEATURE}$	nucleotide phosphate-binding region:NAD	3	0.0463	G6	87	80	20063	1
UP$_{KEYWORDS}$	NAD	3	0.166	G6	87	175	20581	1
INTERPRO	IPR016040:NAD(P)-binding domain	3	0.1944	G6	85	179	18559	1

Tabla 5-2: Agrupación de Genes destacados

Los fenómenos anteriormente descritos destacan la importancia de los genes incluidos en los clusters, debido a que con pocos genes (11, 5 y 3) se pueden diferenciar distintas poblaciones. En la figura **5-29** se muestra la reconstrucción de las estructuras de baja dimensión seleccionando diferentes conjuntos. G8 es un conjunto de 15 genes seleccionados de manera aleatoria en R con el comando sample y la semilla 2021. G9 el conjunto de los 863 genes clasificados como menos importantes y G10 el conjunto conformado por los 100 genes clasificados como más importantes.
La reconstrucción de la baja dimensión con el conjunto de genes G8 es similar a cualquier reconstrucción basada en conjuntos de genes que no aportan al entendimiento de la heterogeneidad celular. Esta reconstrucción de la baja dimensión se muestra como una nube de puntos de una única población. Inclusive si la selección no son 15 genes sino 863 como en G9, el entendimiento de la heterogeneidad celular no es tan claro como si se tuvieran únicamente los 100 genes identificados en las tablas **9-1**, **9-2** y **9-3** ubicadas en los anexos. Ya que en la reconstrucción de la baja dimensión usando el conjunto de genes G10 se alcanza a visualizar con claridad 3 poblaciones diferentes, mientras que en la recontrucción de la baja dimensión usando el conjunto de genes G9 solo se alcanzan a identificar 2 poblaciones.

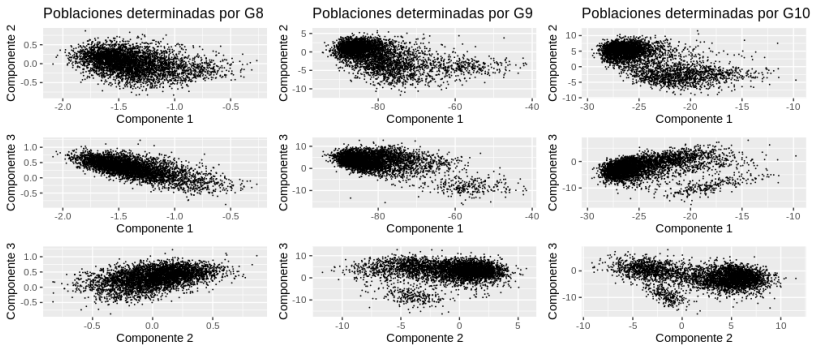

Figura 5-29: Reconstrucción estructuras con diferentes conjuntos de Genes

5.5.4. Resultados del modelo

Al ser una reducción de dimensión basada en un modelo, es importante conocer si este modelo es necesario o si por el contrario, no aporta al entendimiento de la expresión media de los genes en la población global que representan las células analizadas.
A pesar de que los resultados obtenidos son muy interesantes por sus formas de reconstruir esas estructuras de baja dimensión donde se pueden identificar poblaciones, las variables determinadas en el modelo, que son las que fueron presentadas por los autores de la base de datos, no son significativas. Es decir el resultado de la prueba de hipótesis H_4, fue de no rechazar la hipótesis nula (valor p de 0.7). Lo que quiere decir que las variables identificadas por los autores de la base de datos, incluyendo variables dummys que caracterizan si son

tumorosas o no las células, tienen un efecto nulo en las variables de baja dimensión que finalmente representan a la expresión media de las células. Este valor p se obtuvo usando la F de Rao para evaluar el lambda de Wilks.

$H_4 : \beta_1 = \beta_2 = \cdots = \beta_p$

Dónde β_i es la i-ésima fila de la matriz B. Esto se puede analizar visualmente en la figura **5-30**, pues las nubes de puntos no cambian mucho si se incluye o no los efectos fijos del modelo

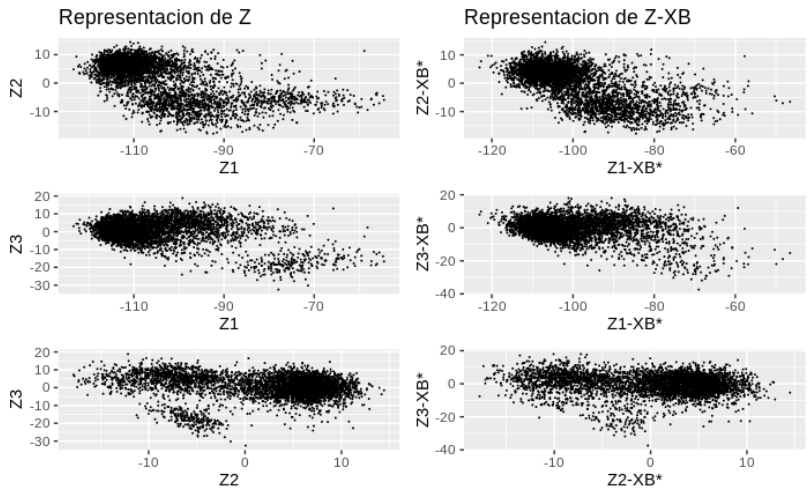

Figura 5-30: Residuales del modelo propuesto

Por otro lado, la importancia de este modelo es encontrar las estimaciones de los factores aleatorios. Para esta ocasión, las estimaciones de las desviaciones estándar son 0.626 y 2.313, lo cual quiere decir que hay una mayor variabilidad por individuo que entre células. No obstante el supuesto de normalidad no se cumple debido a la falta de información acerca de las poblaciones que pueden existir, aunque visualmente no parece descabellado que la distribución sea normal cuando las poblaciones estén bien identificadas, esto se puede ver en la figura **5-31**. Las poblaciones indentificadas tienen una densidad en dos dimensiones cercana a una elíptica, salvo por algunos picos y alargaciones que se dan por la mala identificación de las poblaciones e inclusive la no identificación de poblaciones.

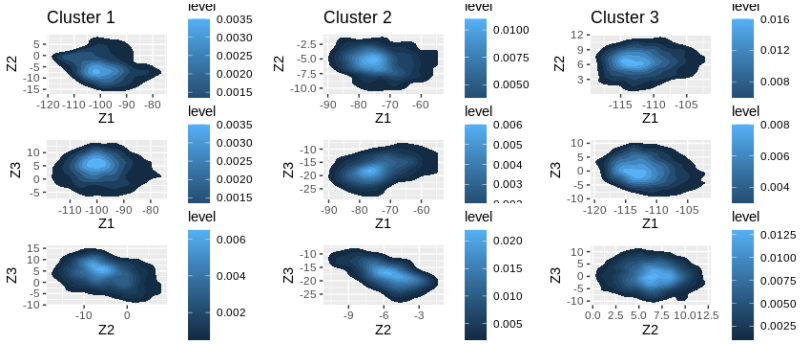

Figura 5-31: Residuales del modelo segragados por cluster

Finalmente, para verificar que no se tiene una distribución Normal, en este caso, se realiza el test de Mardia, cuyos resultados se pueden ver en la tabla 5.5.4

	Beta-hat	kappa	p-val
Skewness	1.40	363.08	0.00
Kurtosis	16.93	6.95	0.00
Skewness.1	1.43	56.86	0.00
Kurtosis.1	19.46	6.29	0.00
Skewness.2	1.25	534.25	0.00
Kurtosis.2	16.31	6.04	0.00

5.6. Prueba de hipótesis para igualdad de distribución de distancias entre puntos

Para observar el desempeño de los algoritmos se usó como medida la tasa de correcta clasificación, no obstante, en la parte inicial de esta sección de resultados, se pudo observar que dicha estadística no mide lo que se quiere, i.e., que las estructuras de baja dimensión se preserven lo mejor posible. Lo anterior, debido a que inclusive se pueden tener altas tasas de correcta clasificación sin necesidad de hacer una reducción de dimensión y que aun teniendo tasas altas de correcta clasificación, algunas poblaciones pueden estar mezcladas y eso evita que se comprenda el proceso biológico causante de los datos. Por otro lado, si se controla el mantenimiento de las estructuras de baja dimensión, se garantiza obtener buenas tasas de correcta clasificación.

Es por eso que en este trabajo se propone un enfoque diferente para próximas simulaciones e investigaciones, y es medir el desempeño de los algoritmos a través de los valores p subyacentes de la prueba de hipótesis basada en permutaciones mencionada en el anexo 1.

Los resultados se encuentran resumidos en la figura **5-32** para la comparación de los algoritmos T-SNE, PCA y envelopes en las distintas simulaciones. Cabe aclarar que, como la

prueba de hipótesis tiene una potencia tan alta, fue necesario multiplicar el estadístico T mostrado en el paso 3 de la prueba de hipótesis por 0.05 para lograr que no todos los valores p fueran nulos, pues es evidente que ningún algoritmo es perfecto y va a entregar las mismas estructuras. Lo que se quiere mostrar aquí es cuál algoritmo produce estructuras similares a las originales. Dicho esto, los valores p en la figura **5-32** no son en realidad probabilidades sino una estadística proporcional a probabilidades que denotan el éxito de la reducción de dimensión de los diferentes algoritmos.

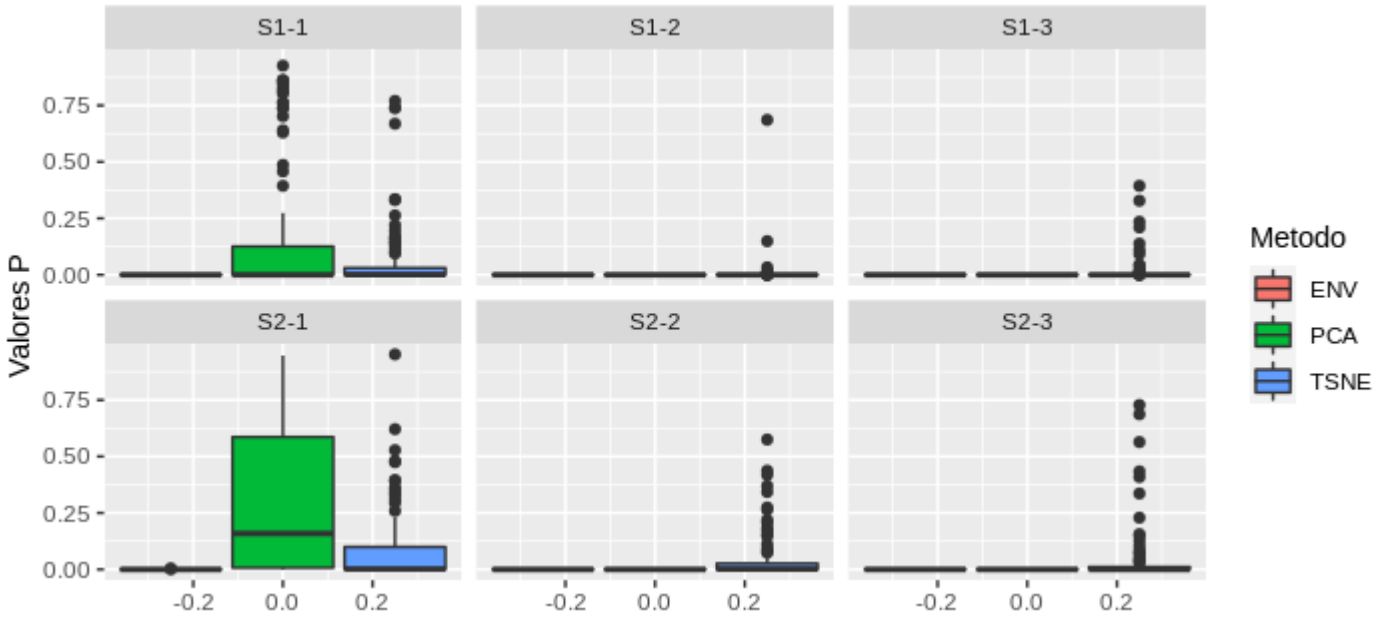

Figura 5-32: Valores p proporcionales para la evaluación de la hipótesis de igualdad de estructuras

En ese sentido, en casos dónde existe linealidad, el PCA sería evidentemente el mejor algoritmo para mantener las estructuras de baja dimensión. El T-SNE en casos de no linealidad, como era de esperarse, sería el mejor algoritmo para mantener las estructuras de baja dimensión y los envelopes no mostrarían ventaja alguna en ningún caso.
De esta manera, si las relaciones que forman la expresión de los genes son no lineales, sí se hace evidente la necesidad de métodos que reconozcan este tipo de relaciones. Lo cual no era tan evidente si se usaban estadísticos como tasas de correcta clasificación.

6 Discusión

Este trabajo pretende entender de mejor manera cómo son los datos de secuenciación de RNA en células individuales desde un punto de vista estadístico y de esta manera encontrar el método más adecuado para analizar estos datos. Se abordó desde un principio a partir de simulaciones para poder cuantificar el desempeño de los distintos algoritmos. Posteriormente, teniendo en cuenta las características de los datos, se propuso un método y se analizaron cuatro diferentes algoritmos en una base de datos real de oligodendrogliomas para posteriormente encontrar una forma de proceder al análisis de estos datos y el alcance que se tiene con estos procedimientos. En esta sección se discutirá acerca de los resultados obtenidos, de lo que se esperaba encontrar, nuevos resultados y la importancia que pretende tener el trabajo en la resolución de la problemática planteada.

6.1. Simulaciones

Con las simulaciones se lograron observar dos aspectos importantes, siendo el primero la necesidad de encontrar estadísticos que cuantifiquen el rendimiento de los distintos algoritmos en el sentido que se desea. En este trabajo, se utilizó en un principio la tasa de correcta clasificación debido a que diferentes autores utilizan este criterio; no obstante, los resultados no fueron los esperados respecto a la linealidad, a diferencia de los que se obtienen usando la prueba de hipótesis de misma distribución de distancias propuesta en los anexos.

El segundo aspecto importante es que la inclusión de factores aleatorios deteriora o cambia el rendimiento de los algoritmos usuales, pues estos no logran comprender el ruido adicional. Especialmente en las simulaciones que incluían relaciones no lineales, la variabilidad de las tasas de correcta clasificación aumentaba, haciendo que para un algoritmo como T-SNE no se tuviera certeza de si el resultado es bueno o malo. Respecto a la medición del éxito del algoritmo a través de la prueba de hipótesis propuesta, esta también se ve afectada por la inclusión de factores aleatorios, no obstante, la afectación es positiva, lo que puede deberse a que se comparan la estructura de los datos sin discriminar al individuo que pertenecen.

6.2. Datos de oligodendrogliomas

Lo primero que se debe resaltar de los resultados obtenidos al usar los datos de oligodendrogliomas, es que los envelopes y el PCA coincidieron siendo métodos diferentes. La igualdad de estos dos modelos indica que el conjunto de variables respuesta es posiblemente independiente al conjunto de variables regresoras (o que al menos las regresoras no aportan al modelamiento de la media), lo que hace que los envelopes no puedan encontrar un subespacio donde Y es invariante a los cambios de X y por lo tanto, el subespacio de reducción de dimensión coincide con las componentes principales.

Esto, sumado con que el conjunto de variables regresoras resultó no significativo para el modelo propuesto, indica que es altamente probable que falte mucho por investigar y entender acerca de la heterogeneidad celular en este tipo de tejidos. Por otro lado, los envelopes y componentes principales se diferenciaron de las componentes halladas por el modelo propuesto. De no ser necesarias las variables incluidas por los autores de la base de datos, estos tres métodos deberían coincidir, por lo tanto, es necesario ahondar en el análisis de este tipo de datos bajo la vista de un modelo de reducción de dimensión que incluya factores mixtos. De hecho, es importante aclarar que otras características de los datos de alta dimensión como ser células tumorosas o no, podrían incluirse como factores aleatorios, ya que como se mencionaba anteriormente, este tipo de características entregadas por los autores de la base de datos, salieron no significativas para modelar la media, sin embargo, la presencia de este tipo de variables fue esencial para poder hacer un proceso de reducción de dimensión que a través de pequeños conjuntos de genes unidos por vías metabólicas destacara y diferenciara distintas poblaciones celulares.

Otro aspecto importante, es que a través de este trabajo se logra destacar la utilidad de métodos de reducción de dimensión basados en modelos lineales, ya que, a pesar de no ser únicamente lineales las relaciones existentes en los datos de secuenciación de RNA en células individuales, estos captan la información suficiente para poder diferenciar las poblaciones celulares. La forma de proceder y analizar los datos en este trabajo, necesita de un método que permita conocer las funciones o la matriz de proyección usada para pasar de los datos de alta dimensión a los datos de baja dimensión; de lo contrario no es posible hacer las reconstrucciones de la baja dimensión con subconjuntos de genes. El no tener esta poderosa herramienta, implica que se tenga que depender de la separación de los datos en la baja dimensión y que el ruido que introducen los genes innecesarios sea pequeño para así poder visualizar los clusters formados. Adicionalmente, si no se entregan las matrices de coeficientes o de proyección va a ser más difícil entender que significa cada componente de baja dimensión, mientras que conociendo los coeficientes, las transformaciones y funciones realizadas para llegar a la baja dimensión se puede especificar qué genes están predominando cada componente y de esta manera, entender el proceso biológico que esta plasmando. Por eso,

para datos de secuenciación de RNA en células individuales, es necesario seguir proponiendo modelos que permitan la identificación de las relaciones e interacciones entre genes y que no se limiten a reducir la dimensión únicamente.

Artículos como [7] ya vienen introduciendo métodos de reducción de dimensión basados en modelos no lineales y envelopes no lineales. Es importante analizar la pertinencia de este tipo de modelos y ver si pueden ser utilizados para datos de secuenciación de RNA en células individuales. Un gran limitante es la forma de estimar los subespacios, que generalmente se hallan por medio de regresiones inversas, no obstante, para los datos presentados en este trabajo, no es posible estos métodos de estimación pues las variables regresoras son categóricas y dicotómicas.

En cuanto a los resultados obtenidos en este trabajo comparados con los obtenidos en [4] y [14]. A pesar de que los enfoques son diferentes, siendo el de estos autores un enfoque desde la biología. El trabajo presente logra agregar nueva información. En [14] se logran identificar 2 cluster y a partir de estos hacen un análisis biológico de la identificación de estas dos poblaciones, mientras que en [4] se encuentran más de 15 clusters, con la diferencia de que a pesar de que caracterizaron varios tipos de células y poblaciones celulares, no se encontró una coincidencia entre las características identificadas y los clusters obtenidos por el procedimiento de reducción de dimensión. En este trabajo se muestra que a través del modelo propuesto se puede llegar a un enriquecimiento que favorece las probabilidades de que los expertos en poblaciones celulares encuentren patrones y relaciones que permitan identificar la heterogeneidad celular que caracteriza al tejido.

7 Conclusiones

- A partir de simulaciones se observa que los métodos clásicos, lineales no deben ser descartados desde un principio, pues pueden inclusive comportarse mejor que métodos no supervisados no lineales dependiendo de las condiciones de los datos y el criterio que se utilice para compararlos. El caso particular de la simulación 2 escenario 3 logra mostrar que los métodos clásicos lineales pueden llegar a obtener excelentes resultados en escenarios de relaciones no lineales, siempre y cuando, existan relaciones lineales que le permitan a estos métodos identificar patrones de interés.
- En simulaciones es necesario definir un criterio para la comparación de los distintos algoritmos. En el marco de secuenciación de RNA en células individuales, la tasa de correcta clasificación no es la mejor estadística si lo que se quiere es comprender los procesos biológicos, debido a que lo que se quiere es el mantenimiento de la estructura de baja dimensión que puede ser medida bajo diferentes conceptos, siendo el propuesto en este trabajo uno de ellos.
- Es necesario que los investigadores y creadores de bases de datos de secuenciación de RNA en células individuales, informen de distintas condiciones de los pacientes, debido a que con lo visto en la base de oligodendrogliomas, ninguna variable reportada influyó en el modelamiento de la media, sin embargo, en el estudio no se reportó ninguna variable de la condición médica o social de los pacientes.
- La inclusión de factores aleatorios en la metodología para comprender los datos de secuenciación de RNA en células individuales mostró ser útil. Por lo tanto, es necesario evaluar qué factores deben ser incluidos en los nuevos modelos y cuáles son posiblemente fijos o aleatorios.
- En datos de secuenciación de RNA en tejidos no se conoce si existen fuentes de ruido externas al proceso a analizar. Es por eso que los métodos de reducción de dimensión basados en modelos son necesarios, puesto que permiten optimizar y enfocar la reducción de dimensión.
- La incorporación de un modelo lineal a la hora de reducir la dimensión de un conjunto de datos permite elaborar procedimientos fáciles de implementar con la ventaja de poder seleccionar genes importantes y analizar subgrupos de genes con mayor profundidad para entender fragmentos importantes de la heterogeneidad celular.

8 Perspectiva

A partir de este trabajo se destacan dos elementos muy importantes para el análisis de datos de secuenciación de RNA en células individuales. El primero y más importante, es que la evidencia hallada parece indicar que los factores aleatorios son cruciales para poder analizar satisfactoriamente este tipo de datos y el segundo elemento es la no linealidad que puede existir.

1. Por lo tanto es necesario crear investigaciones que se enfoquen en analizar qué variables deben ser reportadas en las bases de datos e identificar cuales de estas variables modelan la media o la varianza.

2. Adicionalmente, es necesario investigar métodos de reducción de dimensión que incluyan dichos factores aleatorios y para esto, es necesario crear estadísticas de rendimiento para los diferentes métodos.
   En este trabajo se propone una estadística que intenta cuantificar la eficiencia de un algoritmo para mantener estructuras de baja dimensión, no obstante, es importante aclarar que esta estadística debe ser analizada con mayor profundidad, además que no incluye conceptos clave como la discriminación por individuos. Es decir, se necesita una estadística de rendimiento que mida la similitud de estructuras en conjunto con una correcta agrupación de los datos.

3. Finalmente, con lo que se plantea en este trabajo se espera que se abra la puerta a la creación de nuevos métodos de reducción de dimensión. Métodos de reducción de dimensión basados en modelos lineales mixtos con un gran énfasis en la parte aleatoria debido que a pesar de que en general los factores fijos son de interés, en este tipo de datos parece que no aportan al entendimiento del problema.

4. A medida que se identifican las características importantes como se menciona en el primer numeral, y se desarrollan nuevos métodos lineales que incluyan estos factores aleatorios como se menciona en el tercer numeral, es necesario que se empiecen a llevar este tipo de métodos a la no linealidad para así confirmar si efectivamente hay relaciones de este tipo cuando se habla de secuenciación de RNA en células individuales.

9 Anexos

9.1. Prueba de hipótesis propuesta

En datos de secuenciación es crucial mantener las estructuras de baja dimensión, debido a que la identificación de patrones y procesos biológicos solo va a ser posible si los algoritmos que transforman los datos son capaces de mantenerla. Es por eso que las tasas de correcta clasificación no son la manera más adecuada de medir el rendimiento de un algoritmo para este tipo de datos, pues como se menciona en el trabajo, inclusive se pueden tener altas tasas de correcta clasificación sin incurrir a ningún método de reducción de dimensión.
En este trabajo se propone una forma alternativa de mirar el desempeño de los algoritmos y es a través de las distribuciones de las distancias entre puntos. Se parte del hecho que dos conjuntos de datos con la misma estructura de baja dimensión, tienen un conjunto de distancias entre puntos exactamente igual (el reciproco no es necesariamente correcto).
Inspirados en pruebas como Kolmogorov-Smirnov se entiende que una forma simple de ver si dos conjuntos de datos tienen la misma distribución es a través de la igualdad de las funciones de distribución empírica. Como no es tan eficiente calcular el área entre las dos funciones de distribución empírica, se pueden utilizar cuantiles para esta labor.
Como en este caso no va a haber una distribución de distancias teórica, se puede crear una distribución de las distancias entre puntos a través de permutaciones. De esta manera,la prueba de hipótesis
$H_0 : D_X = D_Y$ vs. $H_A : D_X \neq D_Y$
Siendo D_X la distribución de las distancias del conjunto X, y D_Y la distribución de las distancias del conjunto Y.
Se puede esquematizar como sigue:

1. Obtenga las distancias entre puntos del conjunto X, Y.
2. Estandarice las distancias
3. Calcule la siguiente estadística $T = \sum_{i=1}^{3} (q_{iX} - q_{iY})^2$

Siendo q_{iX} el i-ésimo cuartil de las distancias entre puntos del conjunto X

4. Tome K muestras de tamaño $0{,}9 \times N$ del conjunto de las distancias entre puntos de X
5. Divida en dos esa muestra y calcule la estadística T_k con estos dos nuevos conjuntos
6. A partir de estas estadísticas $T_1, \ldots, T_K$ cree una distribución del estadístico T
7. Evalúe la hipótesis calculando el valor $\mathrm{p} = \dfrac{\sum_{k=1}^{K} 1_{T_k > T}}{K}$

(9-1)

Para verificar que la prueba de hipótesis esté rechazando cuando debe se hacen 4 simulaciones para medir la confianza y potencia empírica.
Para la primera simulación se simulan 1000 conjuntos de puntos aleatorios dentro de un círculo y se compara con 1000 conjuntos de puntos aleatorios dentro de un triángulo equilátero. En esta simulación, al ser dos estructuras diferentes debería rechazarse. En la figura **9-1** se pueden ver los resultados de esta simulación como la linea azul clara que toma valores más altos, esto hace referencia a la potencia empírica. Para la segunda simulación se construyen conjuntos con estructuras iguales, en este caso dos conjuntos cuya estructura forma un triangulo equilátero, a partir de los rechazos de esta prueba de hipótesis se pretende estimar la significancia, es decir, la confianza empírica $(1 - \hat{\alpha})$. Para la tercera y cuarta simulación se hace lo mismo pero mezclando las estructuras, es decir se compara un conjunto de datos que tiene dos estructuras de triángulos con un conjunto que tiene un triángulo y un círculo para calcular la potencia empírica y se compara el conjunto con dos triángulos con otro con dos triángulos, para calcular la significancia empírica, estas ultimas dos se pueden ver en la figura **9-1** en color azul oscuro.
Como es de esperarse, al mezclar las estructuras (azul oscuro) la potencia disminuye, sin embargo, la significancia o error tipo 1 también lo hace. Adicionalmente, se observa que a partir de 1000 puntos la prueba comienza a funcionar bien. Con 200 su potencia es muy reducida.

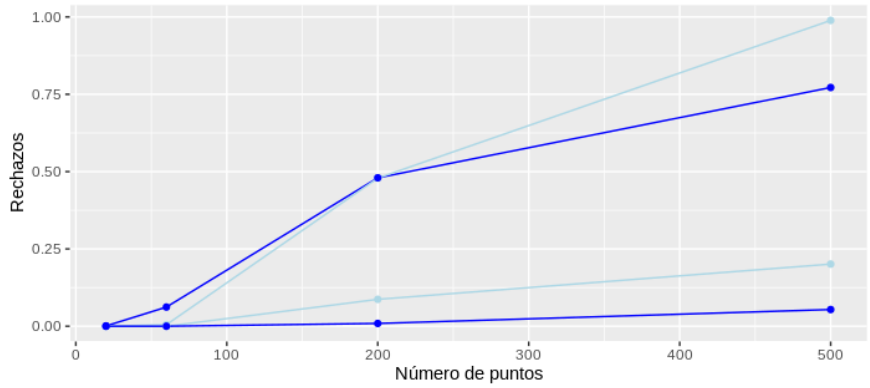

Figura 9-1: Porcentaje de rechazos según escenario. En azul oscuro resultados de significancia y potencia con estructuras dobles y en azul claro los resultados de significancia y potencia con estructuras sencillas.

9.2. Amplificación PCR, transcripción in vitro y amplificación de círculo rodante

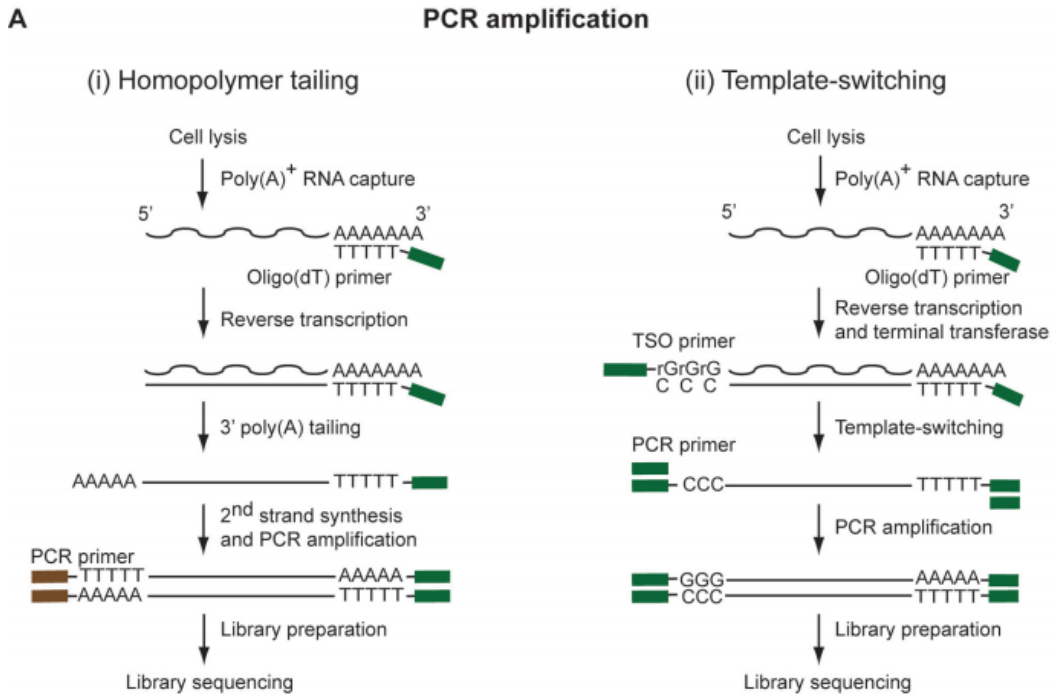

Figura 9-2: Amplificación PCR [5]

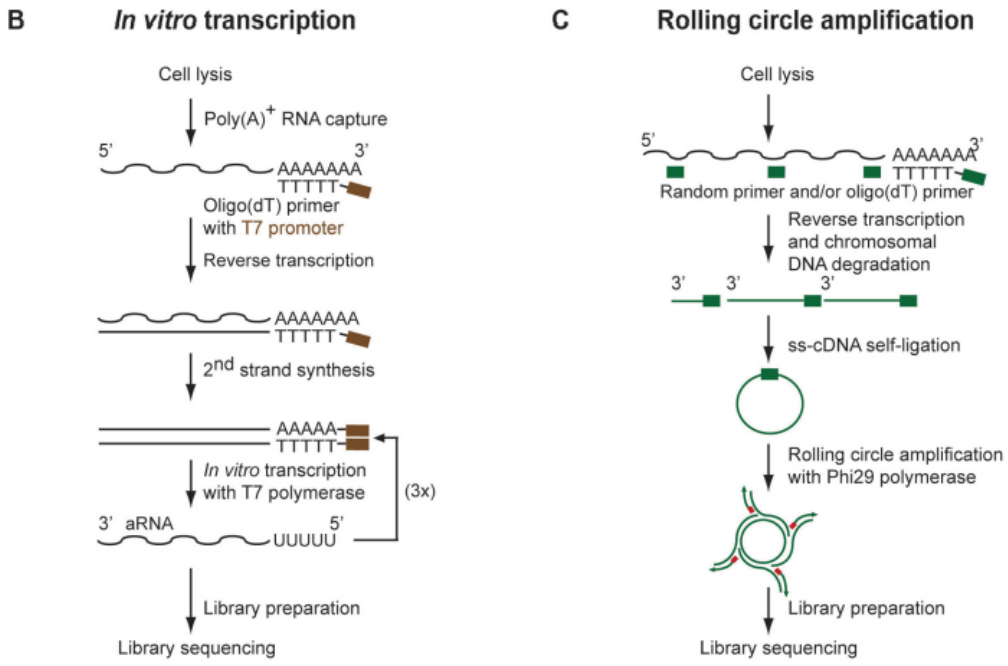

Figura 9-3: Transcripción in vitro y amplificación de círculo rodante [5]

9.3. Autocorrelaciones en PCA y envelopes

A continuacion se muestran las autocorrelaciones y autocorrelaciones parciales halladas en las componentes principales usando únicamente las células del paciente 1.

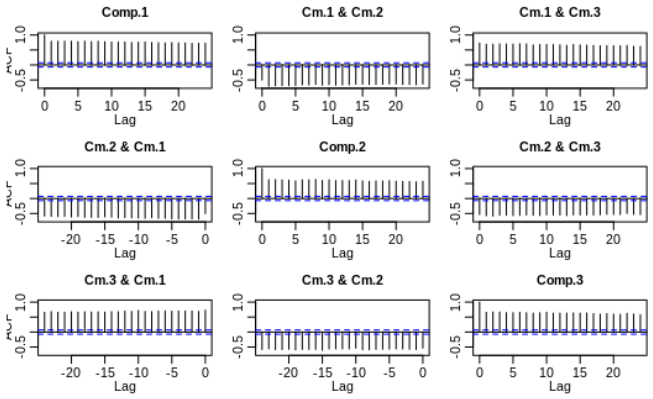

Figura 9-4: Autocorrelaciones para los scores de las componentes principales del paciente 1

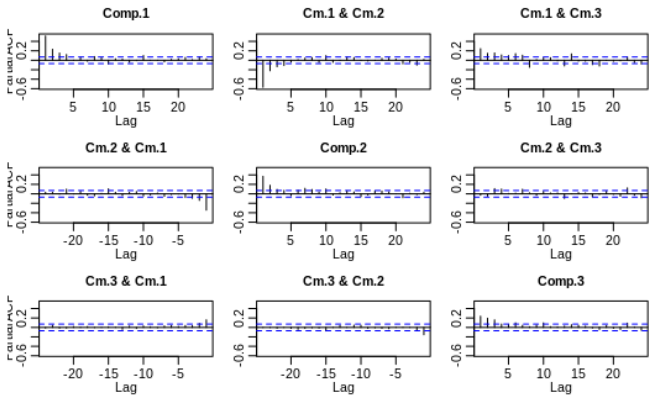

Figura 9-5: Autocorrelaciones parciales para los scores de las componentes principales del paciente 1

9.4. Clasificación según método seleccionado

Clasificación usando el PCA como método de reducción de dimensión

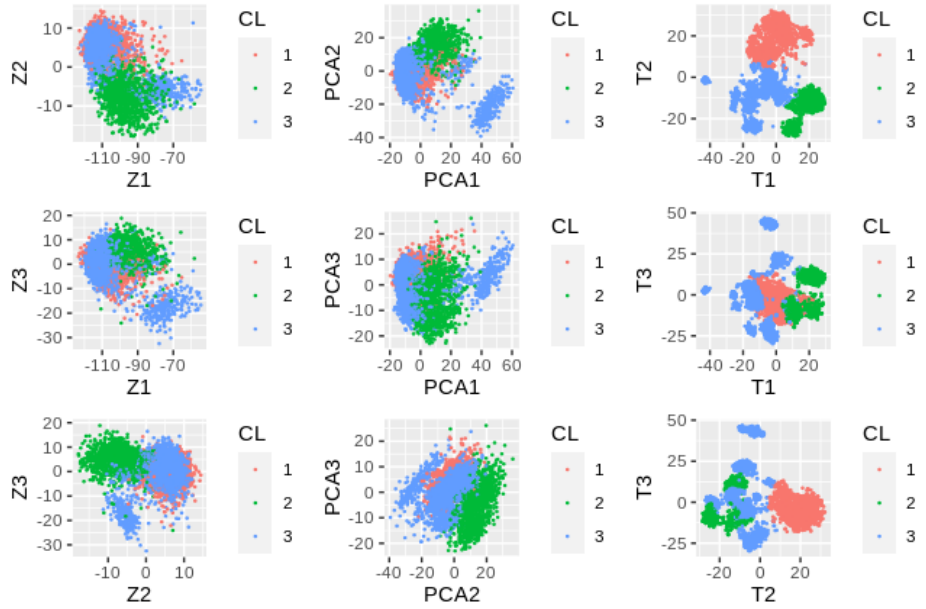

Figura 9-6: Datos en baja dimensión para los distintos modelos según T-SNE

Clasificación usando el T-SNE como método de reducción de dimensión

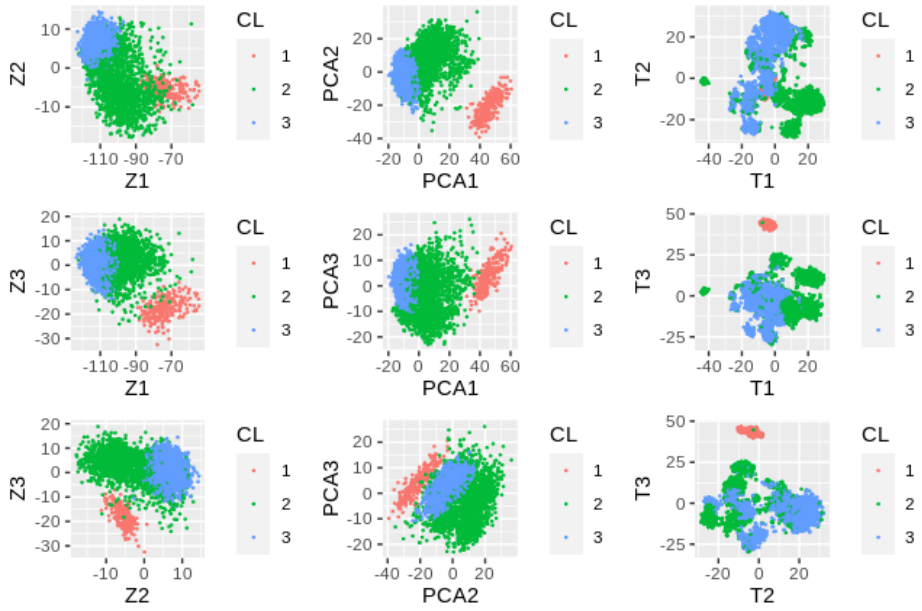

Figura 9-7: Datos en baja dimensión para los distintos modelos según PCA

9.5. Genes influyentes identificados

Orden	Distancia	Gen
1	26.7603932996252	'ACBD6'
2	23.1739857279793	'ACN9'
3	20.970220832096	'ANKLE2'
4	20.7139853573505	'ARHGAP10'
5	20.5459156226671	'ARMCX5-GPRASP2'
6	20.3257405926955	'ADAMTS9-AS2'
7	19.8645559589355	'ACTBL2'
8	17.6691401796171	'ACSM1'
9	17.4051815546951	'ACTRT2'
10	17.0554610256871	'ACP1'
11	16.8552546794008	'AGBL4'
12	15.6702658469868	'ADCYAP1R1'
13	15.3677426136971	'ALPK2'
14	14.23708931858	'ANKRD30BP2'
15	13.7024109747939	'ALKBH4'
16	13.3818559557611	'ANKRD23'
17	12.771283346924	'ANKRD20A4'
18	12.4591868991499	'ADAM29'
19	12.3685821627372	'ALG13'
20	12.1277648495615	'ACTR10'
21	11.9414830899051	'ABHD10'
22	11.7895820494122	'ALS2CR8'
23	11.7839059196184	'ALDH1L1-AS1'
24	11.7554705633373	'AADACL2'
25	11.7344719857034	'APOH'
26	11.494333899316	'GRK5'
27	11.2621800735778	'AKAP6'
28	11.2510222043162	'GABRQ'
29	11.0684930208547	'CEP78'
30	11.0082279143867	'AGBL2'
31	10.8871214090704	'ALOX15B'
32	10.8715140709841	'ADORA2A'
33	10.7702883440462	'AARD'
34	10.5059977486714	'AGAP11'
35	10.2770603202716	'IFNAR2'
36	10.2568355038039	'ADH1A'
37	10.2455544994333	'ANP32A'

Tabla 9-1: Genes más importantes Parte 1

Orden	Distancia	Gen
38	10.1982207674327	'ADAMTS14'
39	10.065469791535	'AACS'
40	10.0288207605556	'ANKRD30BL'
41	10.004985070577	'KIF3B'
42	9.97496065419872	'ALG1'
43	9.8383132008794	'ALLC'
44	9.82443771204134	'ACBD5'
45	9.74940300838143	'ACRC'
46	9.63222758082349	'HIST1H1A'
47	9.55557412212457	'FKBP14'
48	9.53712285338609	'ALDH1L1-AS2'
49	9.42367579864837	'ANKRD50'
50	9.27419753496283	'ADAD1'
51	9.25885693067171	'AGPAT9'
52	9.25389265880266	'ALPP'
53	9.01433922383534	'ACER2'
54	9.00349607251534	'ADRA2B'
55	8.93265862691012	'CACNA1C-IT3'
56	8.88773072870289	'AGO3'
57	8.85744486822037	'AQP4'
58	8.72906063174539	'ABCB1'
59	8.69163463474118	'ADC'
60	8.61880618012659	'ANP32E'
61	8.61337688566249	'APOBEC3G'
62	8.53464832039671	'ADAT2'
63	8.53461387417208	'ARHGAP26-AS1'
64	8.39878763231327	'AKR7A2'
65	8.38673121619587	'AQP10'
66	8.33008702834814	'AMBP'
67	8.28479980487085	'ABCA7'
68	8.26971835752979	'AGAP4'
69	8.1374506515392	'GLDN'
70	8.10707704203893	'APC'
71	8.03304727493213	'ANKEF1'
72	7.9831742937964	'ADAMTS15'
73	7.97311600931905	'ANAPC11'
74	7.94476289805619	'ANO5'
75	7.9206311351942	'ANKRD30A'

Tabla 9-2: Genes más importantes Parte 2

Orden	Distancia	Gen
76	7.87953560934077	'AQP2'
77	7.72869445159928	'LINC00472'
78	7.52786326817793	'ACTR1A'
79	7.49594107664681	'AHCTF1P1'
80	7.42664499929457	'ACTL6A'
81	7.42478173566089	'ABCB7'
82	7.37966573813131	'ACY3'
83	7.35046746012539	'ATP4B'
84	7.22759884315151	'ALX1'
85	7.20973307016795	'AQP12B'
86	7.17590436120032	'AMER2'
87	7.17448468581536	'HOXA-AS4'
88	7.14206391941434	'ALG14'
89	7.09247738057641	'A2MP1'
90	7.0825679293128	'APOF'
91	7.02037479350137	'AKR1C1'
92	7.01095733485178	'AMER1'
93	6.99081209529402	'ANKRD34A'
94	6.98137505117198	'GRK4'
95	6.9674023235833	'ADH6'
96	6.90269054163576	'ABHD5'
97	6.88312790481071	'ANGPTL6'
98	6.85935660149543	'AQP8'
99	6.85035337962985	'ADH4'
100	6.84903226026858	'AMMECR1L'

Tabla 9-3: Genes más importantes Parte 3

Bibliografía

[1] Chial, Heidi: DNA sequencing technologies key to the Human Genome Project. En: *Nature Education* 1 (2008), Nr. 1, p. 219

[2] Cook, R D.: Principal components, sufficient dimension reduction, and envelopes. (2018)

[3] Cook, R D. ; Zhang, Xin: Fast envelope algorithms. En: *Statistica Sinica* 28 (2018), Nr. 3, p. 1179–1197

[4] Ding, Jiarui ; Condon, Anne ; Shah, Sohrab P.: Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. En: *Nature communications* 9 (2018), Nr. 1, p. 1–13

[5] Eberwine, James ; Sul, Jai-Yoon ; Bartfai, Tamas ; Kim, Junhyong: The promise of single-cell sequencing. En: *Nature methods* 11 (2014), Nr. 1, p. 25–27

[6] Hicks, Stephanie. *Welcome to the World of Single-Cell RNA-Sequencing*

[7] Lee, Kuang-Yao ; Li, Bing ; Chiaromonte, Francesca [u. a.]: A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. En: *The Annals of Statistics* 41 (2013), Nr. 1, p. 221–249

[8] Maaten, Laurens van d. ; Hinton, Geoffrey: Visualizing data using t-SNE. En: *Journal of machine learning research* 9 (2008), Nr. Nov, p. 2579–2605

[9] McCulloch, Charles E. ; Neuhaus, John M.: Generalized linear mixed models. En: *Encyclopedia of biostatistics* 4 (2005)

[10] Oshlack, Alicia ; Robinson, Mark D. ; Young, Matthew D.: From RNA-seq reads to differential expression results. En: *Genome biology* 11 (2010), Nr. 12, p. 1–10

[11] Ozsolak, Fatih ; Milos, Patrice M.: RNA sequencing: advances, challenges and opportunities. En: *Nature reviews genetics* 12 (2011), Nr. 2, p. 87–98

[12] Rencher, Alvin C. ; Christensen, WF: Méthods of multivariate analysis. a john wiley & sons. En: *Inc. Publication* (2002), p. 727

[13] Saliba, Antoine-Emmanuel ; Westermann, Alexander J. ; Gorski, Stanislaw A. ; Vogel, Jörg: Single-cell RNA-seq: advances and future challenges. En: *Nucleic acids research* 42 (2014), Nr. 14, p. 8845–8860

[14] Tirosh, Itay ; Venteicher, Andrew S. ; Hebert, Christine ; Escalante, Leah E. ; Patel, Anoop P. ; Yizhak, Keren ; Fisher, Jonathan M. ; Rodman, Christopher ; Mount, Christopher ; Filbin, Mariella G. [u. a.]: Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. En: *Nature* 539 (2016), Nr. 7628, p. 309–313

[15] Wang, Zhong ; Gerstein, Mark ; Snyder, Michael: RNA-Seq: a revolutionary tool for transcriptomics. En: *Nature reviews genetics* 10 (2009), Nr. 1, p. 57–63

[16] Yan, Peng ; Zhou, Bin ; Ma, Yingdong ; Wang, Ani ; Hu, Xiaojun ; Luo, Youli ; Yuan, Yajun ; Wei, Yajun ; Pang, Pengfei ; Mao, Junjie: Tracking the important role of JUNB in hepatocellular carcinoma by single-cell sequencing analysis. En: *Oncology Letters*

[17] Yu, Pingjian ; Lin, Wei: Single-cell transcriptome study as big data. En: *Genomics, proteomics & bioinformatics* 14 (2016), Nr. 1, p. 21–30

[18] Zhu, Li-Ping ; Zhu, Li-Xing ; Feng, Zheng-Hui: Dimension reduction in regressions through cumulative slicing estimation. En: *Journal of the American Statistical Association* 105 (2010), Nr. 492, p. 1455–1466