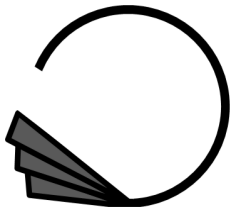
Instituto de Cibernética, Matemática y Física
Departamento de Física Teórica

TESIS DE MAESTRÍA

EXPLORACIÓN DE LOS ATRACTORES HOMEOSTÁTICO, ALZHEIMER Y GLIOBLASTOMA

Autor: Joan Andrés Nieves Cuadrado

Tutor: Dr. Augusto González García, *ICIMAF*



La Habana, 2024

Aquí va la dedicatoria o una frase cool

Resumen

Los datos disponibles de la materia blanca de cerebro permiten localizar los atractores normal (homeostático), Glioblastoma y Alzheimer en el espacio de expresión genética e identificar caminos relacionados con transiciones como la carcinogénesis o la aparición del Alzheimer. También se aprecia una trayectoria predefinida para el envejecimiento, lo cual es consistente con la hipótesis del envejecimiento programado. Adicionalmente, suposiciones razonables sobre la fortaleza relativa de los atractores permite dibujar un panorama esquemático del *fitnest*: diagrama de Wright. Estos sencillos diagramas reproducen relaciones conocidas entre el envejecimiento, el Glioblastoma y el Alzheimer, y plantea cuestiones interesantes como la posible conexión entre el envejecimiento programado y el Glioblastoma en este tejido. Prevemos que múltiples diagramas similares en otros tejidos podrían ser útiles en el entendimiento de la biología de enfermedades o trastornos aparentemente no relacionados, y para descubrir pistas inesperadas para su tratamiento.

Abstract

Available data for white matter of the brain allows to locate the normal (homeostatic), Glioblastoma and Alzheimer's disease attractors in gene expression space and to identify paths related to transitions like carcinogenesis or Alzheimer's disease onset. A predefined path for aging is also apparent, which is consistent with the hypothesis of programmatic aging. In addition, reasonable assumptions about the relative strengths of attractors allow to draw a schematic landscape of fitness: a Wright's diagram. These simple diagrams reproduce known relations between aging, Glioblastoma and Alzheimer's disease, and rise interesting questions like the possible connection between programmatic aging and Glioblastoma in this tissue. We anticipate that similar multiple diagrams in other tissues could be useful in the understanding of the biology of apparently unrelated diseases or disorders, and in the discovery of unexpected clues for their treatment.

Índice general

Introducción

El *fitness* celular refiere a la capacidad de una célula de sobrevivir y proliferar en un ambiente determinado. Abarca varios factores como la capacidad de la célula para adaptarse al ambiente, resistir al estrés y mantener la homeostasis. Este concepto es particularmente importante en la biología del cáncer, donde los niveles de de *fitness* celular pueden determinar su supervivencia y dominancia dentro de un tejido.

Un paradigma conocido en la biología molecular expresa que los máximos locales de *fitness*, en el espacio de expresión genética, están relacionados con estados biológicos accesibles. Un diagrama de Wright es una representación gráfica reducida del panorama genético, donde los picos y valles representan diferentes genotipos y sus niveles relativos de *fitness* [1], ver figura 1.

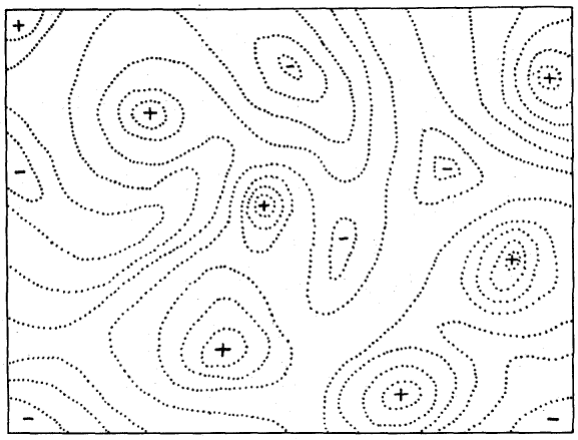


Figura 1: Representación esquemática de un diagrama de Wright. Las zonas de mayor *fitness* (+) están asociadas con estados biológicos accesibles. Las de menor *fitness* (-) son solo zonas de transito. Las lineas discontinuas delimitan los distintos niveles de *fitness*. Tomada de la referencia [1]

Esta representación ha sido aplicada a la descripción del destino celular a lo largo de una

linea de diferenciación [2]. Sin embargo, hasta donde sabemos, no hay gráficos basados en datos reales para un tejido dado que represente, al menos de forma parcial, un diagrama de Wright con más de dos máximos. En el presente trabajo, mostramos un diagrama para la materia blanca del cerebro en el cual el estado normal (N) se representa junto con el atractor de glioblastoma (GB) y el máximo relacionado con la enfermedad del Alzheimer (EA).

La EA se caracteriza por la pérdida progresiva de células neuronales, mientras que el GB es un tipo de cáncer cerebral que implica la proliferación descontrolada de células. Algunas investigaciones muestran que pacientes con la EA podrían tener un menor riesgo de desarrollar GB. Por ejemplo, en la referencia [3] se muestra que ambas enfermedades presentan un aumento en el estrés oxidativo, en la EA, esto hace que las células neuronales sean más vulnerables a la muerte, mientras que en el GB, las células cancerosas se vuelven más resistentes. Además, se plantea que la degeneración de células que secretan acetilcolina en la EA podría tener un efecto protector contra el cáncer, ya que esta sustancia puede estimular el crecimiento de células cancerosas.

Los estudios epidemiológicos señalan que aquellas personas que padecen una de estas enfermedades tiene un riesgo menor de desarrollas la otra. Concretamente, los pacientes con cáncer tienen un riesgo alrededor de 31 y 35 % menor de desarrollar la EA en comparación con aquellos sin antecedente de cáncer. Mientras que las personas que padecen la EA tienen un riesgo entre 41 y 61 % menor de desarrollar cáncer [4–6].

Esta idea de la EA y el GB como alternativas opuestas también está apoyada por un gran número experimentos de biología molecular. Por ejemplo, en la referencia [7] los autores encuentran que las vías de señalización ERK/MAPK están aumentadas en GB y disminuidas en la EA. Estas vías conectan señales extracelulares, como factores de crecimiento y citocinas, a respuestas intracelulares que regulan la proliferación celular, diferenciación y supervivencia. También muestran que las vías de señalización de la angiopoyetina están aumentadas en la EA y disminuidas en GB, estas relacionadas con la regulación de la angiogénesis y la estabilidad vascular.

El estudio realizados por C. Lanni y colaboradores en la referencia [8] destaca varios actores moleculares, especialmente PIN1 y p53, que están involucrados en interacciones moleculares complejas asociadas con la correlación inversa entre estas enfermedades. El aumento en la expresión de PIN1 esta relacionado con un retraso en la edad de inicio de la EA, mientras niveles bajos de expresión se asocian con un menor riesgo de desarrollar varios tipos de cáncer.

Muchas de estas investigaciones no solo muestran una oposición entre la EA y el GB, sino que también señalan el envejecimiento como un factor de riesgo común para ambas enfermedades [5–9]. Esta compleja relación entre ambas enfermedades cerebrales queda representada en nuestro diagrama de Wright. Además, se puede apreciar un camino o corredor hacia el envejecimiento normal, en línea con la teoría del envejecimiento programado [10, 11].

A nivel genético, hay genes que varían de la misma manera en los procesos de envejecimiento, progresión de AD y cáncer, mientras que también hay genes que indican una situación disyuntiva entre la EA y el GB. Un ejemplo de este último es el gen codificador de proteínas MMP9, que juega un papel importante en la invasión tumoral [12, 13], pero también conocido como neuroprotector, controlando las interacciones entre axones y fibras beta-amiloide [14]. Las desviaciones del valor de expresión génica de su referencia en el tejido normal pueden indicar una progresión potencial a AD (subexpresión) o a GB (sobreexpresión).

Esta visión inusual puede ayudar a comprender las relaciones entre la EA y el GB, e identificar marcadores génicos útiles para ambos procesos. Como un bono adicional, la representación permite encontrar preguntas muy interesantes que se discutirán a continuación.

El trabajo presentado como tesis de diploma consta de una introducción, tres capítulos, las conclusiones y las recomendaciones. El contenido se distribuye de la siguiente forma:

- **Introducción:** Se hace una breve descripción del panorama genético y epidemiológico que engloban a la EA, el GB y el envejecimiento. Se ejemplifican las relaciones que existen entre estos atractores.
- **Capítulo 1:**
- **Capítulo 2:**
- **Capítulo 3:**
- **Conclusiones:**
- **Recomendaciones:**

Capítulo 1

Materiales y métodos

1.1. Reducción de la dimensionalidad

La reducción de la dimensionalidad es un paso crucial en el análisis de datos, especialmente cuando se trata de grandes conjuntos con muchas variables. Este facilita la visualización y comprensión de los datos, lo que permite una rápida identificación de patrones y tendencias importantes. Algunas técnicas comunes de reducción dimensional son PCA (*Principal Component Analisis*), t-SNE (*t-Distributed Stochastic Neighbor Embedding*) y UMAP (*Uniform Manifold Approximation and Projection*). Estas técnicas son herramientas poderosas que ayudan a simplificar la complejidad de los datos sin perder información crítica, lo que permite un análisis más eficiente y efectivo.

El t-SNE es un método no linear y estocástico. Su funcionamiento se puede separar en dos etapas. En la primera, se seleccionan los vecinos de cada punto. Para ello se utiliza una distribución gaussiana alrededor de él, donde los más cercanos tienen un probabilidad mayor de ser seleccionados que los lejanos. Este paso permite al modelo preservar las estructuras locales. Durante la segunda etapa, se asignan posiciones iniciales aleatorias en un espacio de menor dimensión (generalmente 2 o 3 dimensiones). Luego, se define una distribución de probabilidad similar para los puntos en el nuevo espacio y se minimiza la divergencia entre las dos distribuciones. Esta etapa ayuda a mantener la fidelidad de la representación en el espacio reducido. De esta forma, el algoritmo logra una transformación de los datos hacia una dimensión reducida que preserva la similitud entre los vecinos cercanos. Buscar algunas

referencias donde expliquen t-SNE

UMAP es una técnica muy similar a t-SNE. Una de las diferencias principales es que, durante la selección de los vecinos, se asume que los datos forman una variedad de menor dimensión que el espacio original. Esto le permite ser más eficiente en términos de tiempo de cómputo y más efectivo en la conservación de relaciones a gran escala. Otra característica de este método es que, para hacer la representación reducida, minimiza la entropía cruzada en lugar de la divergencia entre las distribuciones. Estas diferencias permiten a este algoritmo preservar mejor tanto la estructura local como la global de los datos, además de hacerlo capaz de trabajar con datos que no se ajustan necesariamente a una distribución normal. Buscar algunas referencias donde expliquen UMAP

Por otro lado, el PCA es una técnica lineal y determinista que transforma variables correlacionadas en un conjunto reducido de variables no correlacionadas, conocidas como componentes principales. Al igual que en los casos anteriores, su funcionamiento se puede dividir en dos etapas. En la primera etapa, se centran los datos en su media aritmética y se calcula la matriz de covarianza. Esta matriz captura la variabilidad conjunta entre múltiples variables aleatorias y permite comprender las relaciones entre ellas. Luego, durante la segunda etapa, se obtienen los autovalores, ordenados de mayor a menor, y sus correspondientes autovectores. Los autovectores representan las direcciones de máxima varianza en los datos, mientras que los autovalores indican la cantidad de varianza que se encuentra en cada una de estas direcciones. Al proyectar los datos originales sobre los primeros autovectores, se obtiene una representación de los datos en un sistema ortogonal que maximiza la conservación de la varianza, utilizando el menor número posible de componentes [15]. Buscar algunas referencias donde expliquen PCA

t-SNE y UMAP se han vuelto muy populares últimamente debido a su eficiencia y la gran capacidad de para visualizar datos de alta dimensión en un espacio de 2 o 3 dimensiones [insert cites here]. Pero su naturaleza no linear y estocástica hace que sea complejo una interpretación cuantitativa de los resultados. Por otro lado, PCA es una técnica lineal y determinista que junto a su relativa sencillez permite realizar varias interpretaciones de sus resultados [insert cites here].

Sin embargo, la aplicación directa del PCA puede presentar algunas complicaciones. Una de las principales dificultades es la construcción de la matriz de covarianza, ya que el número de elementos que contiene es igual al cuadrado de la dimensión original de los datos. Esto hace que sea imposible almacenarla en la memoria RAM de la mayoría de los equipos de

cómputo personales. Por ejemplo, si el número de componentes de los datos es 5×10^4, la matriz de covarianza tendría $2{,}5 \times 10^9$ elementos. Suponiendo que cada elemento ocupe 8 bytes, el tamaño total de la matriz sería aproximadamente 19 GB. Si se hace uso de que esta es una matriz simétrica, se podría guardar solo los elementos de la triangular superior (o inferior), permitiendo reducir el almacenamiento necesario casi a la mitad. Sin embargo, aún así se requeriría mucho espacio y este escala rápidamente con el aumento de la dimensión de los datos. Por lo tanto, en general, es necesario recurrir al almacenamiento en disco, que tiene una velocidad de lectura y escritura menor que la memoria RAM.

Otro problema grave que enfrenta el algoritmo estándar del PCA es el cálculo de los autovalores y autovectores. La mayoría de las implementaciones de los métodos directos usados para este cálculo no permiten que se apliquen a grandes matrices debido a las limitaciones de la memoria RAM. Una característica del PCA que resulta de gran ayuda en esta parte es que, en general, no es necesario calcular todos los autovalores, solo los más grandes y sus correspondientes autovectores.

Un algoritmo relativamente sencillo de implementar y que permite calcular solo los autovalores más grandes y sus correspondientes autovectores es el método de Lanczos. En el caso del análisis de datos de expresión genética, donde solo un subconjunto diferente de genes se expresa en cada tejido, la matriz de covarianza podría tener un número elevado de valores nulos. Esto es beneficioso para el método de Lanczos, ya que funciona mejor con matrices dispersas.

El método de Lanczos puede ser de gran utilidad en algunos problemas donde las implementaciones estándar pueden verse limitadas. Sin embargo, en la práctica se utilizan otras técnicas para realizar el PCA de forma indirecta. Una de las alternativas es la descomposición en valores singulares (SVD, por sus siglas en inglés). Por estas razones, a continuación mostraremos brevemente la implementación básica del método de Lanczos. Luego, en qué consiste el SVD y cómo realizar el PCA de forma indirecta a partir de este.

Algoritmo de Lanczos

El método de Lanczos es una técnica numérica utilizada para encontrar los autovalores y autovectores de una matriz grande y dispersa. Es especialmente útil en problemas de álgebra lineal donde la matriz es demasiado grande para ser manejada por métodos directos.

Poner aquí el algoritmo general del método de Lanczos

Una de las limitaciones de este algoritmo es su estabilidad numérica. Buscar citas

Descomposición en valores singulares

La SVD provee una descomposición numéricamente estable de matrices que puede ser usado en una gran variedad de propósitos. Como resultado de aplicar este algoritmo se obtiene una descomposición matricial única que existe para toda matriz de valores complejos $\mathbf{X} \in \mathbb{C}^{n\times m}$:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*, \tag{1.1}$$

donde $\mathbf{U} \in \mathbb{C}^{n\times n}$ y $\mathbf{V} \in \mathbb{C}^{m\times m}$ son matrices unitarias con columnas ortonormales, y $\mathbf{\Sigma} \in \mathbb{R}^{n\times m}$ una matriz con valores reales no negativos en la diagonal y ceros fuera de la diagonal. Aquí * denota la transpuesta conjugada.

Cuando $n \geq m$, la matriz $\mathbf{\Sigma}$ tiene como máximo m valores distintos de cero en la diagonal y puede ser escrita como $\mathbf{\Sigma} = \begin{bmatrix} \hat{\mathbf{\Sigma}} \\ 0 \end{bmatrix}$. Por lo tanto, es posible representar $\mathbf{X}$ de forma exacta usando la versión reducida de SVD:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \begin{bmatrix} \hat{\mathbf{U}} & \mathbf{U}^{\perp} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{\Sigma}} \\ 0 \end{bmatrix} \mathbf{V}^*, \tag{1.2}$$

las columnas de $\mathbf{U}^{\perp}$ abarcan un espacio vectorial que es complementario y ortogonal a $\hat{\mathbf{U}}$. Las columnas de $\mathbf{U}$ son llamadas vectores singulares izquierdos de $\mathbf{X}$ y forman una base del espacio de los vectores columnas de $\mathbf{X}$. Las columnas de $\mathbf{V}$ son los vectores singulares derechos y forman una base para los vectores filas de $\mathbf{X}$. Los elementos diagonales de $\hat{\mathbf{\Sigma}} \in \mathbb{C}^{m\times m}$, los llamados valores singulares, están ordenados de mayor a menor. El rango de $\mathbf{X}$ es igual a la cantidad de valores singulares distintos de cero.

Para ver la relación de esta técnica con el PCA, partimos de la matriz de covarianza. Esta se construye a partir de la siguiente expresión:

$$\sigma_{ij} = \frac{1}{N-1}\sum_{l=1}^{N}\left(x_i^{(l)} - \mu_i\right)\left(x_j^{(l)} - \mu_j\right), \tag{1.3}$$

donde i y j son la característica i-ésima y j-ésima del conjunto de datos estudiado, en

nuestro caso son el gen i y j, respectivamente. La variable l se recorre por todas las muestras. σ_{ij} es el elemento $(i,\ j)$ de la matriz de covarianza, es decir, es la covarianza entre el gen i y el j para los elementos no diagonales y la varianza del gen i para los elementos de la diagonal principal. El número total de muestras es N, $x_i^{(l)} \in \mathbb{R}$ es el valor de la componente i en la muestra l, y, por último, μ_i es la media de los valores de la componente i sobre todas las muestras.

Si los datos ya han sido previamente centrados, es decir, se cumple que $\mu_i = 0$ para todo i, entonces la ecuación (1.3) puede reducirse a:

$$\sigma_{ij} = \frac{1}{N-1}\sum_{l=1}^{N} x_i^{(l)} x_j^{(l)}, \tag{1.4}$$

Si definimos una matriz $\mathbf{X} \in \mathbb{R}^{n\times m}$, tal que el elemento $(i,\ j)$ es igual a $x_i^{(j)}/(N-1)$, la ecuación (1.4) queda representada en forma matricial como:

$$\mathbf{C} = \mathbf{X}^T\mathbf{X}. \tag{1.5}$$

Si usamos la SVD de $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ y la propiedad de ortonormalidad de $\mathbf{U}$ y $\mathbf{V}$, obtenemos:

$$\mathbf{C} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{1.6}$$

$$= \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T, \tag{1.7}$$

es decir, $\mathbf{\Sigma}$ y $\mathbf{V}$ son la solución del siguiente problema de autovalores:

$$\mathbf{C}\mathbf{V} = \mathbf{V}\mathbf{\Sigma}^2. \tag{1.8}$$

En otras palabras, cada valor singular de $\mathbf{X}$ distinto de cero es la raíz cuadrada positiva de los autovalores de su matriz de covarianza, y las columnas de $\mathbf{V}$ son los autovectores. En términos del PCA, las columnas de $\mathbf{V}$ son las componentes principales de $\mathbf{X}$ y los elementos de la diagonal de $\mathbf{\Sigma}^2$ representan la varianza de los datos en cada una de las componente.

Esta es la manera usual en la que los algoritmo actuales realizan el PCA, por ejemplo, la librería *scikit-learn* de Python [16]. En general, nosotros preferimos usar directamente SVD para hacer el PCA, ya que brinda mayor flexibilidad y control. Por ejemplo, si en lugar usar

una matriz de media cero, los datos se centran en el valor medio de un subconjunto, las componentes principales van a indicar la dirección en la cual hay más dispersión respecto a este subjconjunto.

Una ventaja significativa de este método para realizar el PCA, es que se obtiene las componentes principales directamente, sin tener que calcular la matriz de covarianza. Esto permite ahorrar tiempo y recursos computacionales, algo que toma mayor importancia en la medida que aumenta la cantidad de datos y el números de variables a procesar. Un procedimiento similar se siguió en las referencias [17, 18].

1.2. Datos de expresión genética

Utilizamos datos de expresión genética obtenidos de dos experimentos distintos. El primero de ellos contiene muestras patológicamente normales (o "sanas") y con glioblastoma, y fueron tomadas del Atlas del Genoma del Cáncer (TCGA, https://www.cancer.gov/tcga) [19, 20]. Estas son tomadas durante procedimientos quirúrgicos. Los tumores se pueden localizar en diferente zonas cerebrales pero, como es común en el Glioblastoma, estos son tumores tomados de la sustancia blanca del cerebro [21].

Hay 5 muestras normales de pacientes con edades en el rango entre 49 y 74 años, mientras que el intervalo de edades de las 169 muestras de glioblastoma es entre 21 y 89 años. Las pacientes femeninas representan aproximadamente dos tercios de la cohorte.

El segundo grupo de datos proviene del estudio longitudinal del Instituto Allen sobre el envejecimiento y la demencia (https://aging.brain-map.org/) [22]. Las muestras son tomadas *post mortem*. El grupo de control está conformado por 47 muestras, mientras que en el otro hay 28 muestras. El intervalo de edad para todas estas muestras es entre 77 y 101 años. El diagnóstico de la enfermedad del Alzheimer está respaldado por pruebas cognitivas y otras pruebas clínicas. Alrededor del 40 % de la cohorte son mujeres.

En ambos experimentos los valores de expresión genética se encuentran en unidades FPKM (*fragments per kilobase of transcript per million fragments mapped*), que es una unidad común en este tipo de estudios. En términos simples, dicha unidad de medida significa: la tasa de fragmentos por base multiplicada por un número muy grande (10^9). El cálculo de FPKM para el gen i se realiza por medio de la siguiente fórmula [23]:

$$
\begin{aligned}
FPKM_i &= \frac{q_i}{(l_i/10^3)(\sum_j q_j/10^6)}, \\
&= \frac{q_i}{l_i \sum_j q_j} * 10^9,
\end{aligned}
\tag{1.9}
$$

donde q_i es la cantidad de fragmentos contados, l_i es la longitud del gen, y $\sum_j q_j$ corresponde al número total de fragmentos.

Capítulo 2

Diagrama de tres atractores

2.1. El diagrama de N + GB + EA

Nuestro punto de partida es el diagrama de los datos de expresión genética del análisis de componentes principales para la materia blanca del cerebro, mostrado en la Fig. 2.1. Como se puede notar en la figura, las dos primeras componentes principales capturan más del 80 % de la varianza del sistema. Por lo tanto, es una representación bidimensional adecuada de la distribución real de los puntos en el espacio de expresión genética.

En la figura se pueden apreciar 4 grupos de muestras. Las muestras marcadas como N y GB corresponden, a especímenes patológicamente normales y tumorales en los datos del TCGA para el Glioblastoma [19]. Los centros de las nubes de muestras de N y GB en el espacio de expresión genética definen, respectivamente, los atractores Normal (homeostático) y Glioblastoma de Kauffman [17, 24]. De hecho, la acumulación de puntos en una determinada región de este espacio indica que esta es un atractor de la red de regulación Genética que gobierna la dinámica del sistema.

Por otro lado, los grupos etiquetados como EA y O corresponden a las muestras de la materia blanca del cerebro de la enfermedad del Alzheimer y del grupo de control (*old*) en el estudio del Instituto Allen [22].

La Fig. 2.2 es una reconstrucción de la figura 3 de la referencia [18]. En esta se muestra los resultados del PCA para los datos de expresión genética de la materia blanca del cerebro del Instituto Allen. La primera componente principal (PC1), la cual contiene el 24,7 % de la

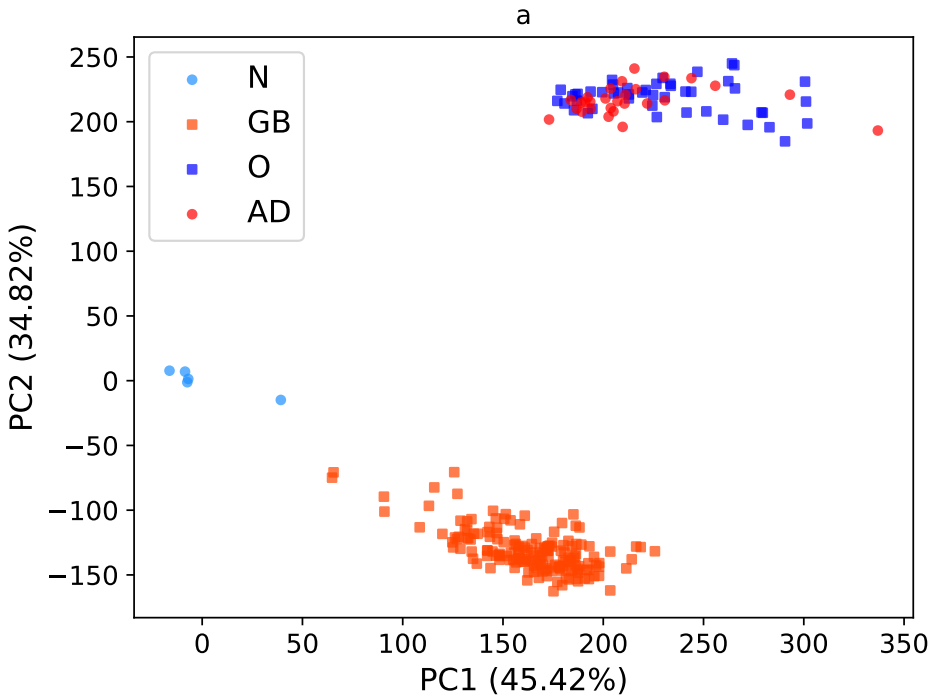


Figura 2.1: Análisis de componentes principales para los datos de expresión genética.

varianza total, discrimina entre las muestra de O y EA. La posición del centro de la nube de O en este eje es $\langle x_1 \rangle = 0$, y para la EA es $\langle x_1 \rangle = 40{,}97$. Sin embargo, los radios de las nubes de las muestras de O y EA son más grandes que la distancia entre los centros, que son 80,69 y 72,64 respectivamente.

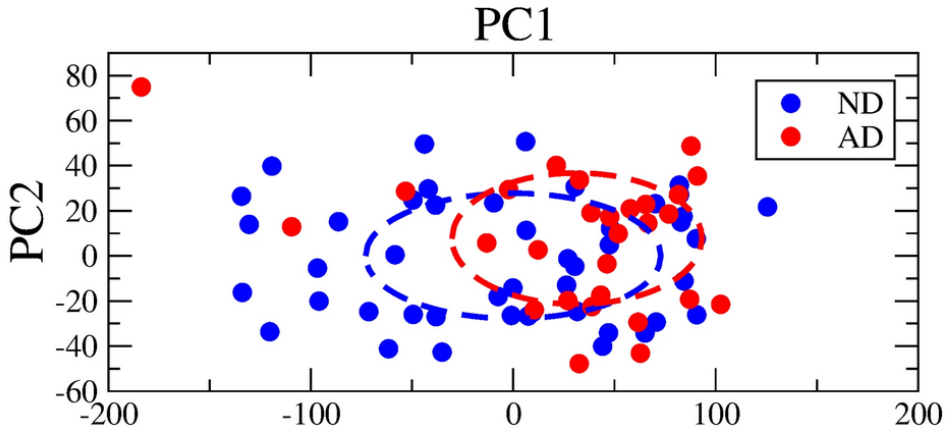


Figura 2.2: Figura tomada de la referencia [18].

Es bien conocido el papel de la edad en la EA, especialmente en ancianos [25]. Por lo tanto, podemos usar la edad como una variable de tiempo para seguir la transición. A pesar del número relativamente pequeño de muestras, se realizó un análisis de regresión lineal de la posición media de $\langle x_1 \rangle$ en función de la edad en las muestras de O, Fig. 2.3a, muestra que $\langle x_1 \rangle = -287{,}12 + 3{,}24 \cdot edad$. En las muestras de la EA, sin embargo, no se encontró correlación entre $\langle x_1 \rangle$ y la edad observada. Por lo tanto, la posición de la zona EA es aproximadamente fija, y la nube de muestras de O muestra una deriva hacia el mínimo de EA a medida que aumenta la edad.

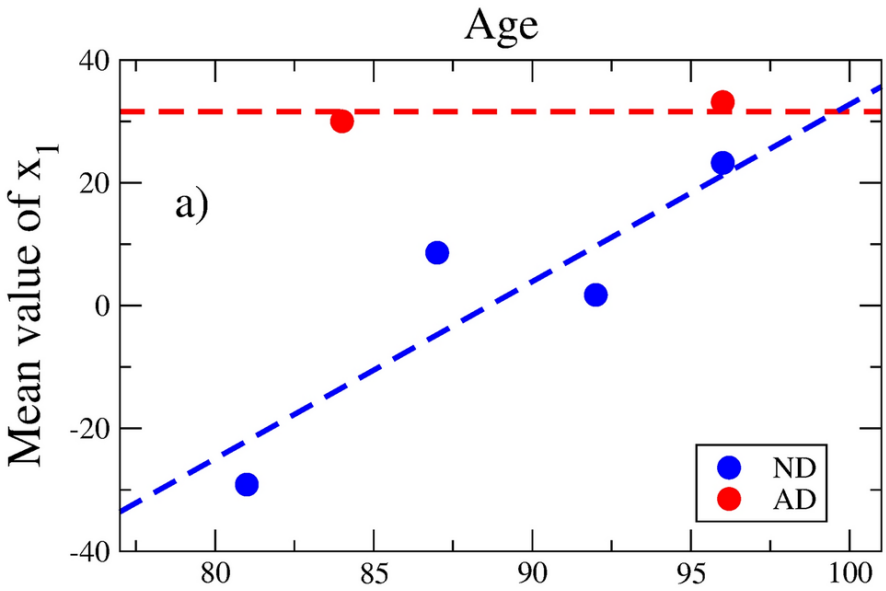


Figura 2.3: Figura tomada de la referencia [18]

Una mejor ilustración de este hecho viene representada en la Fig. 2.4, donde se compara la densidad de probabilidad de las muestras de O y de la EA. Se definen cuatro intervalos de edades, que contienen aproximadamente la misma cantidad de muestras de O: [77, 84], [84, 90], [90, 95], [95, 100+]. La probabilidad total de las muestras de la EA es mostrada en los cuatro paneles. Es aparente un desplazamiento de las muestras de O hacia la zona de la EA.

Esta propiedad sugiere que el centro de la nube de muestras de la EA define un atractor en el espacio de expresión genética. Las muestras de O parecen ser atrapadas por el atractor de la EA en el proceso del envejecimiento.

Así, en nuestra aproximación, obtenemos un panorama en el espacio de expresión genética

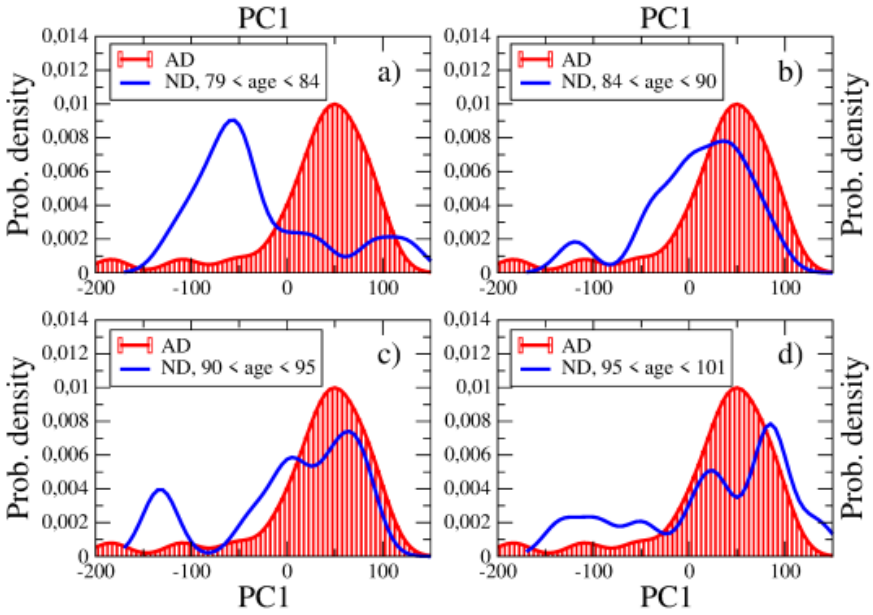


Figura 2.4: Densidad de probabilidad de las muestras de O y de la EA a lo largo del eje de PC1. Cada panel es para un intervalo de edad para las muestras de O. La probabilidad de la EA, la cual es aproximadamente independiente de la edad, es mostrada en los cuatro paneles. Figura tomada de la referencia [18]

de tres atractores: N, GB y EA, y un conjunto de muestras de O que se desplaza hacia la EA. Las posiciones relativas y las principales transiciones entre los atractores se resumen en la Fig. 2.5. Asumimos que estas transiciones están determinadas por la biología subyacente a los procesos en los tejidos. La transición de N a la EA se denomina "EA anticipada" para enfatizar que también existe una vía hacia la EA a través del envejecimiento: "EA tardía". La figura también indica una vía para el GB y para el envejecimiento.

2.2. Panorama del fitness

Existe información cualitativa que puede introducirse en nuestra descripción. Esta se relaciona con una variable de *fitness*, de modo que dibujamos una especie de diagrama de Wright [1]. En la Fig. 2.6 se muestra un diagrama esquemático que contiene un gráfico de contorno hipotético del *fitness*. Los atractores N y GB son máximos de *fitness* y deberían estar separados por una barrera de bajo *fitness* [18]. El GB debería ser el máximo más alto

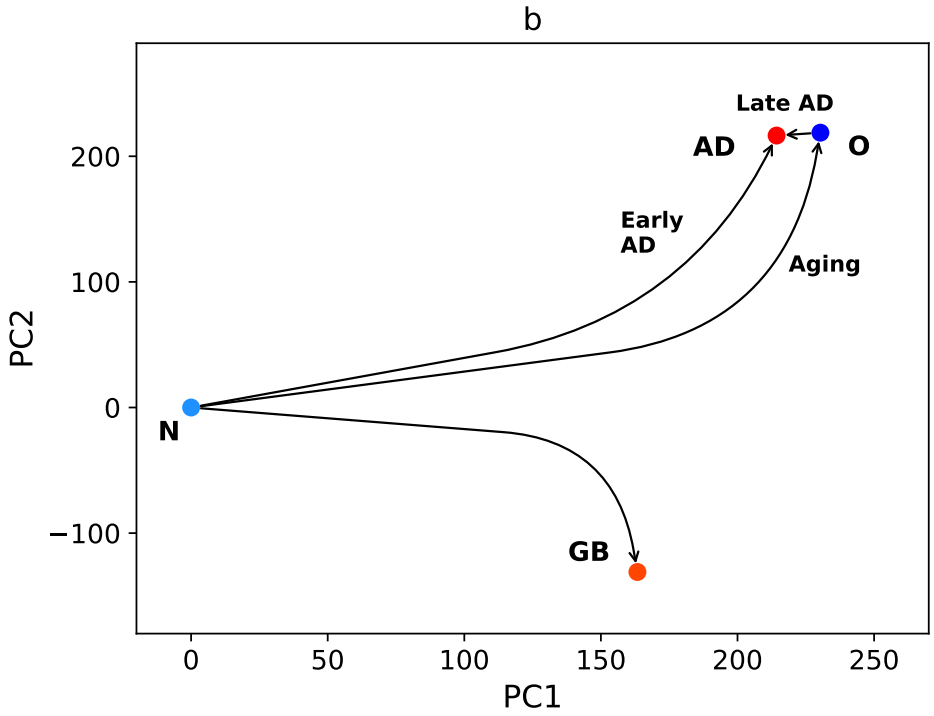


Figura 2.5: Posiciones relativas y principales transiciones entre los atractores.

de los tres actores representados [18, 26]. Por otro lado, la transición de O a la EA es casi continua, con un número relativamente pequeño de genes expresados diferencialmente [18]. Esto significa que existe una barrera muy pequeña, o incluso una ruta sin barrera, que conecta a O y a la EA. Esperamos una barrera de bajo *fitness* que impida las transiciones directas de O a la EA, y un máximo de la EA pequeño, ya que este atractor se encuentra en la región de bajo *fitness*, lejos de N.

Todos estos hechos se representan en la Fig. 2.6. El esquema se construye a partir de una suma de gaussianas centradas en los atractores, con desviaciones estándar proporcionales a los valores reales observados en la Fig. 2.1 y con alturas que respetan cualitativamente la fuerza relativa de los atractores.

Resaltemos el significado de un diagrama de Wright en el tejido cerebral. En otros tejidos, la evolución somática se relaciona principalmente con la replicación de células madre. Sin embargo, en su estado normal, el cerebro es un tejido de replicación muy lenta [27]. Los cambios en pequeñas regiones cerebrales, es decir, los desplazamientos en el diagrama, son

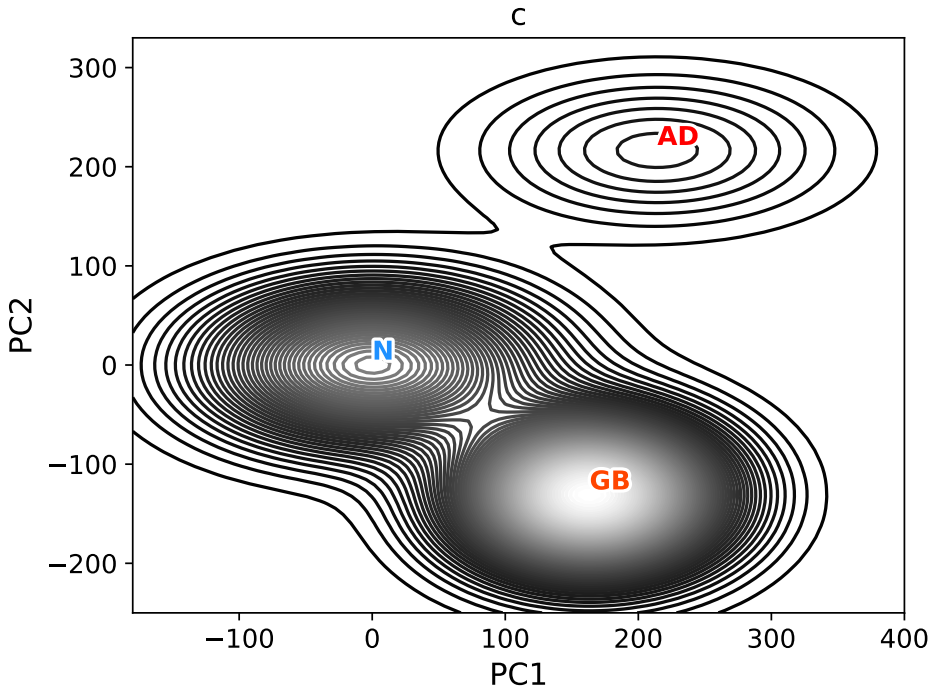


Figura 2.6: Diagrama de Wright que muestra un gráfico de contorno hipotético del *fitness*. El máximo absoluto corresponde con el estado de GB. El atractor de la EA se representa como un ligero máximo local.

básicamente daños acumulados, es decir, envejecimiento [28]. Sin embargo, una vez que se produce la transición al estado GB, se produce un enorme aumento de la tasa de replicación de las células tumorales. Observemos, además, que los cambios relacionados con el envejecimiento son muy evidentes en la sustancia blanca [29].

2.3. Limitaciones

En este trabajo se utilizaron datos de expresión genética, en formato FPKM, de las referencias [19, 22]. Estos fueron obtenidos usando diferentes plataformas. Nosotros tomamos aproximadamente 30 000 genes que están perfectamente identificados en ambas plataformas y realizamos un sencillo análisis de componentes principales [15], como se definió en la sección 1.1. Para definir los valores de la expresión diferencial logarítmica y calcular la matriz de covarianza utilizada para el PCA, se utilizó como referencia común la media geométrica en el

conjunto de muestras N.

Debido al uso de estos datos, proveniente de dos experimentos distintos, para realizar un solo calculo de PCA surgen problemas tanto técnicos como conceptuales. Por ejemplo, la referencia N corresponde precisamente al estado normal del cerebro, sino que son un conjunto de muestras patológicamente normales que fueron tomadas de individuos con GB. Además, dos de los pacientes tienen más de 70 años. Desde el punto de vista computacional, por otro lado, se podrían utilizar correcciones por lotes [30, 31], que corrigen parcialmente los sesgos asociados a cada grupo de muestras, pero también pueden introducir problemas incontrolados.

En lugar de introducir procedimientos muy avanzados, preferimos extraer los datos directamente de las fuentes y utilizar la técnica de PCA más sencilla. No creemos que ninguna corrección altere sustancialmente el análisis cualitativo derivado del diagrama de tres atractores que se muestras en la Fig. 2.1.

La situación ideal podría ser repetir el experimento dentro de un único marco tecnológico, incluyendo datos de personas jóvenes sanas, que se utilizarían para establecer la referencia de los cálculos de la expresión genética diferencial, incluyendo datos de pacientes con GB y la EA, y datos de pacientes sanos en diferentes rangos de edad. Este es un experimento complejo, pero podría ser particularmente factible en un modelo de ratones [32], por ejemplo. Consideramos nuestro diagrama de la Fig. 2.1 como una aproximación cualitativa de este experimento ideal.

Capítulo 3

Resultados

Sobre la base de nuestros diagramas, podemos formular las siguientes observaciones o afirmaciones, que son los principales resultados del trabajo.

1. **Existe una dirección en el espacio de expresión genética, que a grandes rasgos se puede identificar con el eje PC1, asociada al envejecimiento y a un aumento del riesgo de padecer GB y la EA.**

De hecho, el desplazamiento en esta dirección implica escalar parcialmente las barreras de baja aptitud que separan N de los estados GB y EA, y por lo tanto aumentar el riesgo tanto para GB como para la EA.

Vale la pena observar los principales genes involucrados en este proceso. Para ello, observamos el vector unitario a lo largo del eje PC1. Los genes se clasifican según su contribución al vector unitario. El procedimiento es similar al algoritmo de Page Rank [29]. Lo usamos en nuestro trabajo anterior [19].

Conclusiones

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

Recomendaciones

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdiet sed, pellentesque nec, nisl. Vestibulum imperdiet neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero. Suspendisse in elit quis nisl aliquam dapibus. Pellentesque auctor sapien. Sed egestas sapien nec lectus. Pellentesque vel dui vel neque bibendum viverra. Aliquam porttitor nisl nec pede. Proin mattis libero vel turpis. Donec rutrum mauris et libero. Proin euismod porta felis. Nam lobortis, metus quis elementum commodo, nunc lectus elementum mauris, eget vulputate ligula tellus eu neque. Vivamus eu dolor.

Nulla in ipsum. Praesent eros nulla, congue vitae, euismod ut, commodo a, wisi. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Aenean nonummy magna non leo. Sed felis erat, ullamcorper in, dictum non, ultricies ut, lectus. Proin vel arcu a odio lobortis euismod. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Proin ut est. Aliquam odio. Pellentesque massa turpis, cursus eu, euismod nec, tempor congue, nulla. Duis viverra gravida mauris. Cras tincidunt. Curabitur eros ligula, varius ut, pulvinar in, cursus faucibus, augue.

Bibliografía

[1] S. Wright, "The roles of mutation, inbreeding, crossbreeding and selection in evolution," in *Proceedings of the sixth international congress of Genetics*, vol. 1, pp. 356–366, 1932.

[2] M. J. Casey, P. S. Stumpf, and B. D. MacArthur, "Theory of cell fate," *WIREs Systems Biology and Medicine*, vol. 12, p. e1471, Dec. 2019.

[3] S.-M. Ou, Y.-J. Lee, Y.-W. Hu, C.-J. Liu, T.-J. Chen, J.-L. Fuh, and S.-J. Wang, "Does Alzheimer's Disease Protect against Cancers? A Nationwide Population-Based Study," *Neuroepidemiology*, vol. 40, pp. 42–49, Oct. 2012.

[4] C. M. Roe, A. L. Fitzpatrick, C. Xiong, W. Sieh, L. Kuller, J. P. Miller, M. M. Williams, R. Kopan, M. I. Behrens, and J. C. Morris, "Cancer linked to Alzheimer disease but not vascular dementia," *Neurology*, vol. 74, pp. 106–112, Jan. 2010.

[5] J. A. Driver, A. Beiser, R. Au, B. E. Kreger, G. L. Splansky, T. Kurth, D. P. Kiel, K. P. Lu, S. Seshadri, and P. A. Wolf, "Inverse association between cancer and Alzheimer's disease: results from the Framingham Heart Study," *BMJ*, vol. 344, pp. e1442–e1442, Mar. 2012.

[6] M. Musicco, F. Adorni, S. Di Santo, F. Prinelli, C. Pettenati, C. Caltagirone, K. Palmer, and A. Russo, "Inverse occurrence of cancer and Alzheimer disease: A population-based incidence study," *Neurology*, vol. 81, pp. 322–328, July 2013.

[7] T. Liu, D. Ren, X. Zhu, Z. Yin, G. Jin, Z. Zhao, D. Robinson, X. Li, K. Wong, K. Cui, H. Zhao, and S. T. C. Wong, "Transcriptional signaling pathways inversely regulated in Alzheimer's disease and glioblastoma multiform," *Scientific Reports*, vol. 3, Dec. 2013.

[8] C. Lanni, M. Masi, M. Racchi, and S. Govoni, "Cancer and Alzheimer's disease inverse relationship: an age-associated diverging derailment of shared pathways," *Molecular Psychiatry*, vol. 26, pp. 280–295, May 2020.

[9] J. A. Driver and K. Ping Lu, "Pin1: A New Genetic Link between Alzheimers Disease, Cancer and Aging," *Current Aging Sciencee*, vol. 3, pp. 158–165, Dec. 2010.

[10] J. P. Magalhães, "Programmatic features of aging originating in development: aging mechanisms beyond molecular damage?," *The FASEB Journal*, vol. 26, pp. 4821–4826, Sept. 2012.

[11] D. Gems, "The hyperfunction theory: An emerging paradigm for the biology of aging," *Ageing Research Reviews*, vol. 74, p. 101557, Feb. 2022.

[12] G. Choe, J. K. Park, L. Jouben-Steele, T. J. Kremen, L. M. Liau, H. V. Vinters, T. F. Cloughesy, and P. S. Mischel, "Active matrix metalloproteinase 9 expression is associated with primary glioblastoma subtype1," *Clinical Cancer Research*, vol. 8, pp. 2894–2901, 09 2002.

[13] Q. Xue, L. Cao, X.-Y. Chen, J. Zhao, L. Gao, S.-Z. Li, and Z. Fei, "High expression of mmp9 in glioma affects cell proliferation and is associated with patient survival rates," *Oncology Letters*, vol. 13, pp. 1325–1330, Jan. 2017.

[14] A. Kaminari, N. Giannakas, A. Tzinia, and E. C. Tsilibary, "Overexpression of matrix metalloproteinase-9 (mmp-9) rescues insulin-mediated impairment in the 5xfad model of alzheimer's disease," *Scientific Reports*, vol. 7, Apr. 2017.

[15] J. Lever, M. Krzywinski, and N. Altman, "Principal component analysis," *Nature Methods*, vol. 14, pp. 641–642, 7 2017.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] A. Gonzalez, D. A. Leon, Y. Perera, and R. Perez, "On the gene expression landscape of cancer," *PLOS ONE*, vol. 18, p. e0277786, Feb. 2023.

[18] A. Gonzalez, J. Nieves, D. A. Leon, M. L. Bringas Vega, and P. V. Sosa, "Gene expression rearrangements denoting changes in the biological state," *Scientific Reports*, vol. 11, Apr. 2021.

[19] C. W. Brennan *et al.*, "The Somatic Genomic Landscape of Glioblastoma," *Cell*, vol. 155, pp. 462–477, Oct. 2013.

[20] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemp Oncol (Pozn)*, vol. 1A, pp. 68–77, 2015.

[21] B. Ellingson, A. Lai, R. Harris, J. Selfridge, W. Yong, K. Das, W. Pope, P. Nghiemphu, H. Vinters, and L. a. Liau, "Probabilistic radiographic atlas of glioblastoma phenotypes," *American Journal of neuroradiology*, vol. 34, no. 3, pp. 533–540, 2013.

[22] J. A. Miller *et al.*, "Neuropathological and transcriptomic characteristics of the aged brain," *eLife*, vol. 6, Nov. 2017.

[23] Y. Zhao, M.-C. Li, M. M. Konaté, L. Chen, B. Das, C. Karlovich, P. M. Williams, Y. A. Evrard, J. H. Doroshow, and L. M. McShane, "TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository," *Journal of Translational Medicine*, vol. 19, jun 2021.

[24] S. Huang, I. Ernberg, and S. Kauffman, "Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective," *Seminars in Cell & Developmental Biology*, vol. 20, pp. 869–876, Sept. 2009.

[25] "2019 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 15, pp. 321–387, Mar. 2019.

[26] A. Gonzalez, F. Quintela, D. A. Leon, M. L. Bringas-Vega, and P. A. Valdes-Sosa, "Estimating the number of available states for normal and tumor tissues in gene expression space," *Biophysical Reports*, vol. 2, no. 2, 2022.

[27] K. L. Spalding, R. D. Bhardwaj, B. A. Buchholz, H. Druid, and J. Frisén, "Retrospective birth dating of cells in humans," *Cell*, vol. 122, no. 1, pp. 133–143, 2005.

[28] B. Schumacher, J. Pothof, J. Vijg, and J. H. Hoeijmakers, "The central role of dna damage in the ageing process," *Nature*, vol. 592, no. 7856, pp. 695–703, 2021.

[29] C. R. Guttmann, F. A. Jolesz, R. Kikinis, R. J. Killiany, M. B. Moss, T. Sandor, and M. S. Albert, "White matter changes with normal aging," *Neurology*, vol. 50, no. 4, pp. 972–978, 1998.

[30] L. Haghverdi, A. T. L. Lun, M. D. Morgan, and J. C. Marioni, "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors," *Nature biotechnology*, vol. 36, no. 5, pp. 421–427, 2018.

[31] Y. Zhang, G. Parmigiani, and W. E. Johnson, "Combat-seq: batch effect adjustment for rna-seq count data," *NAR genomics and bioinformatics*, vol. 2, no. 3, p. lqaa078, 2020.

[32] O. Hahn, A. G. Foltz, M. Atkins, B. Kedir, P. Moran-Losada, I. H. Guldner, C. Munson, F. Kern, R. Pálovics, N. Lu, *et al.*, "Atlas of the aging mouse brain reveals white matter as vulnerable foci," *Cell*, vol. 186, no. 19, pp. 4117–4133, 2023.