Universidad Nacional Autónoma de México
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Maestría en Ciencia e Ingeniería de la Computación

Detección de expresión genética anómala, el caso de la vitamina D

T E S I S

que para optar por el grado de

Maestro en Ciencias de la Computación

Presenta:
Michelle Dubhé Mata Hernández

Tutor principal:
Dr. José Antonio Neme, IIMAS

Mérida, Yucátan, México. 2024

Agradecimientos

Quiero expresar mi más profundo agradecimiento a todas las personas e instituciones que hicieron posible la realización de esta tesis.

En primer lugar, agradezco a mi tutor, el Dr. José Antonio Neme, por su guía,sus consejos, paciencia y por su apoyo a lo largo de este proceso, gracias por sus atenciones y cuidados cuando más lo necesitaba. Su experiencia y compromiso fueron fundamentales para lograr este trabajo.

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca que me permitió concentrarme plenamente en mis estudios y desarrollo académico. A su vez extiendo mi gratitud al proyecto PAPIIT liderado por la Dra. Nidiyare Hevia, cuyo apoyo fue esencial para permitirme concentrarme en mis objetivos.

A mis padres y a mi hermano, por su amor incondicional, palabras de aliento y fe en mí, incluso en los momentos más difíciles. Su apoyo ha sido mi más grande bendición y mi inspiración constante.

Finalmente, a mi esposo, por su paciencia, comprensión y apoyo incondicional en cada paso de este camino. Su compañía ha sido mi más grande placer en la vida.

Gracias a todos por ser parte de esta travesía.

Índice general

Índice de figuras

Capítulo 1

Introducción

La vitamina D es un micronutriente fundamental que influye en múltiples procesos fisiológicos y metabólicos [1]. Su relevancia se destaca por su participación en diversos efectos polifacéticos en diferentes tejidos y funciones del cuerpo. El papel principal de la vitamina D es participar en la absorción del calcio, así como mantener la regulación de la homeostasis. Hoy en día se reconoce que la influencia de esta vitamina en el sistema humano es mayor, ya que participa en mecanismos de regulación de insulina, la función endotelial, control del ciclo celular, regulación de apoptosis, regulación del sistema inmunológico, entre otros [2].

Es fundamental comprender la importancia de la vitamina D en nuestro cuerpo. La deficiencia de esta vitamina puede causar diversas comorbilidades, cada una con un grado de riesgo diferente. Algunas enfermedades están directamente relacionadas con la falta de vitamina D, como la osteomalacia y el raquitismo. Otras, como la diabetes, el cáncer, las enfermedades auto-inmunes, las enfermedades cardiovasculares y la osteoporosis, están relacionadas de manera indirecta con esta deficiencia. [3] Según estimaciones en la actualidad aproximadamente mil millones de personas en el mundo padecen insuficiencia de vitamina D, presentándose más abruptamente en mujeres posmenopáusicas en un 50 %, por el contrario, solo el 40 % de los varones sufren de esta insuficiencia y sus efectos [4][5].

En este contexto, la detección de genes con expresión anómala, modulados por la vitamina D se presenta como una línea de investigación relevante. Los avances en algoritmos de detección de anomalías ofrecen una oportunidad sin precedentes para identificar patrones genéticos rela-

cionados con la deficiencia de vitamina D o proporcionar información sobre genes anómalamente expresados abriendo nuevas puertas para la prevención y tratamiento de estas enfermedades. En esta tesis se propone explorar estas posibilidades, utilizando técnicas de análisis de datos para detectar genes anómalos asociados con la vitamina D.

En el área de la genómica y la bioinformática, existen diversos métodos que facilitan la comprensión de la expresión génica y su relación con distintas condiciones fisiológicas y patológicas. Uno de los principales enfoques de investigación es la identificación de genes con expresión diferencial (DGE, por sus siglas en inglés). Tradicionalmente, los métodos estadísticos y bioinformáticos han sido las herramientas clave para analizar estos datos [6].

El propósito de los métodos estadísticos es identificar genes con expresión anómala. Esto se logra mediante pruebas estadísticas que cuantifican la expresión de genes a partir de análisis computacionales de lecturas de RNA-seq [7].Así, se determina qué genes presentan una diferencia estadísticamente significativa en su expresión y se obtiene información detallada sobre los niveles de expresión y las diferencias entre pares de genes [8].

En la actualidad existen grandes volúmenes de datos genómicos y surge la necesidad de enfoques más flexibles. Este trabajo se centra en la aplicación de técnicas innovadoras: los algoritmos de detección de anomalías, para identificar genes con expresión anómala, en contraste con genes diferencialmente expresados en el contexto de la vitamina D.

Los algoritmos de detección de anomalías ofrecen una perspectiva diferente para el análisis de datos de expresión génica al identificar genes con patrones de expresión que se desvían significativamente de los patrones normales. Estos algoritmos analizan características específicas de los datos, como la distancia entre distribuciones o las densidades locales de los puntos, para detectar anomalías sin depender de supuestos previos sobre la distribución estadística de los datos [9]. Este enfoque no solo reduce sesgos asociados con métodos tradicionales, sino que también optimiza el uso de recursos computacionales, proporcionando resultados con una comprensión diferente sobre los detalles del comportamiento de los genes.

Es importante reconocer la relevancia de las anomalías en los datos genómicos. Aunque existen distintas definiciones de lo que se puede considerar como una anomalía, en términos generales, se refiere a datos que se desvían significativamente del comportamiento del resto. A menudo, las anomalías se confunden con ruido o errores en la captación de datos, por lo que es fundamental ser cuidadoso al identificarlas correctamente [10].

Establecer criterios a evaluar como métodos de comparación es fundamental; estos pueden ser alguna métrica o característica compartida, donde solamente el caso de la anomalía se verá afectada, reflejando un cambio significativo en la variabilidad de los datos [10][11]. A medida que los conjuntos de datos incrementan en tamaño y el espacio de características aumenta, los métodos convencionales a menudo se vuelven computacionalmente intensivos. El proceso de detección de anomalías presentado permite entrenar un clasificador que utiliza sus características, y predice si forma parte de la misma clase, ofreciendo una solución rápida para el desafío de analizar grandes volúmenes de datos genómicos.

En este proyecto nos enfocaremos en el uso de los algoritmos K-means (K-medias) y Local Outlier Factor (LOF) para detectar anomalías en la expresión genética en el caso de la vitamina D. Estos algoritmos fueron seleccionados debido a su capacidad para identificar patrones no evidentes en datos complejos sin depender de supuestos estadísticos estrictos. LOF es utilizado para la detección de anomalías de un conjunto de datos evaluando la densidad del punto de interés con la densidad de sus vecinos más cercanos [12]. Por otro lado, el algoritmo K-medias permite analizar y agrupar los datos en diferentes grupos o cúmulos, donde cada cúmulo representa un conjunto de elementos similares [13]. Su elección responde a la necesidad de métodos versátiles que puedan identificar diferentes características de los datos y proporcionar una visión más amplia de los comportamientos anómalos en los genes analizados.

En resumen, la vitamina D juega un papel importante en numerosos procesos fisiológicos más allá de su conocida función en la regulación del calcio. Su deficiencia está asociada tanto a enfermedades directas como a una serie de condiciones crónicas, incluidas enfermedades autoinmunes,como por ejemplo el cáncer o la diabetes, además de afectar directamente a la salud mental. Ante la creciente insuficiencia de vitamina D a

nivel global, especialmente en mujeres posmenopáusicas, este estudio se centra en el análisis genético para detectar genes relacionados con la vitamina D mediante algoritmos de detección de anomalías, como K-medias y LOF. Aunque estos algoritmos son utilizados en otros campos, su aplicación en el análisis de expresión genética es poco común, lo que los convierte en una alternativa innovadora dentro de este contexto. Esto permite generar nuevas perspectivas para identificar patrones genéticos anómalos y desarrollar estrategias más efectivas para el tratamiento y la prevención de enfermedades vinculadas a la deficiencia de vitamina D.

1.1. Objetivo

El objetivo principal de este trabajo es aplicar algoritmos de detección de anomalías existentes para identificar genes con expresión anómala en el contexto de la vitamina D. Este estudio se enfoca en detectar genes que presentan comportamientos distintos entre el grupo de control y el grupo de tratamiento. Para ello, se analizarán datos de expresión génica obtenidos de tres repeticiones o experimentos independientes en cada caso, lo que permitirá identificar qué genes muestran una alteración en su expresión en respuesta al tratamiento aplicado.

1.2. Hipótesis

Se plantea que mediante el uso de técnicas de detección de anomalías es factible identificar genes cuyo comportamiento varía entre dos condiciones específicas, como, por ejemplo, entre un grupo de tratamiento y un grupo de control. Estas técnicas computacionales permitirán identificar patrones genéticos distintivos que pueden no ser evidentes mediante métodos tradicionales de análisis genómico.

Capítulo 2

Fundamentos biológicos de la expresión genética

2.1. Información genética básica

En esta sección se busca facilitar la comprensión básica de términos utilizados en genética, proporcionando una introducción al tema. La genética es la disciplina que estudia la herencia y nos ayuda a entender cómo nuestros genes influyen en distintos aspectos biológicos. Los genes tienen la capacidad de influir en las características bioquímicas y fisiológicas de los individuos, como la estatura, el color de cabello, piel y ojos. Además, influyen en la predisposición a desarrollar ciertas enfermedades, así como en capacidades físicas y mentales [14].

Los rasgos genéticos pueden influir en la salud y el bienestar de una persona. Mientras que algunas alteraciones genéticas o genes heredados pueden no tener un impacto significativo, otros pueden tener efectos menores o, en algunos casos, afectar de manera más notable la calidad o duración de la vida. [15][16].

El proceso comienza con el ADN (ácido desoxirribonucleico), la molécula que contiene la información genética hereditaria en los seres vivos [17]. El ADN está compuesto por dos cadenas que se enrollan formando una doble hélice. A su vez, cada cadena está formada por unidades más pequeñas llamadas nucleótidos. Cada nucleótido consta de tres componentes: una base nitrogenada (Adenina, Timina, Citosina o Guanina), un azúcar

llamado desoxirribosa y un grupo fosfato [18].

La secuencia de los nucleótidos en el ADN es lo que determina la información genética, ya que distintas combinaciones de estos nucleótidos codifican diferentes genes. Los genes son esenciales en la transmisión de la información genética, ya que son responsables de heredar rasgos de una generación a la siguiente [18].

Un gen es una sección específica del ADN que contiene la secuencia de nucleótidos que se traducirán, en diferentes pasos, en aminoácidos y de ahí en proteínas para regular diversas actividades moleculares dentro de la célula [19]. Cada gen está compuesto por una secuencia de nucleótidos que se organiza de manera específica y cuenta con regiones definidas que indican inicio y término. Estas delimitaciones permiten que los procesos de transcripción y traducción, que convierten la información genética en proteínas funcionales, se lleven a cabo correctamente [20] (ver Fig. 2.1).

El proceso de formación de un gen comienza con la transcripción, donde la secuencia de ADN se copia en una molécula de ARN (ácido ribonucleico). Posteriormente, el ARN es traducido en proteínas en los ribosomas de la célula [21]. Las proteínas, que son moléculas complejas, cumplen funciones esenciales en el cuerpo humano como puede ser: la estructura celular, el transporte de nutrientes, la comunicación entre células, la defensa inmunológica y la regulación de procesos bioquímicos, lo que convierte a las proteínas en componentes necesarios para el funcionamiento del organismo [22].

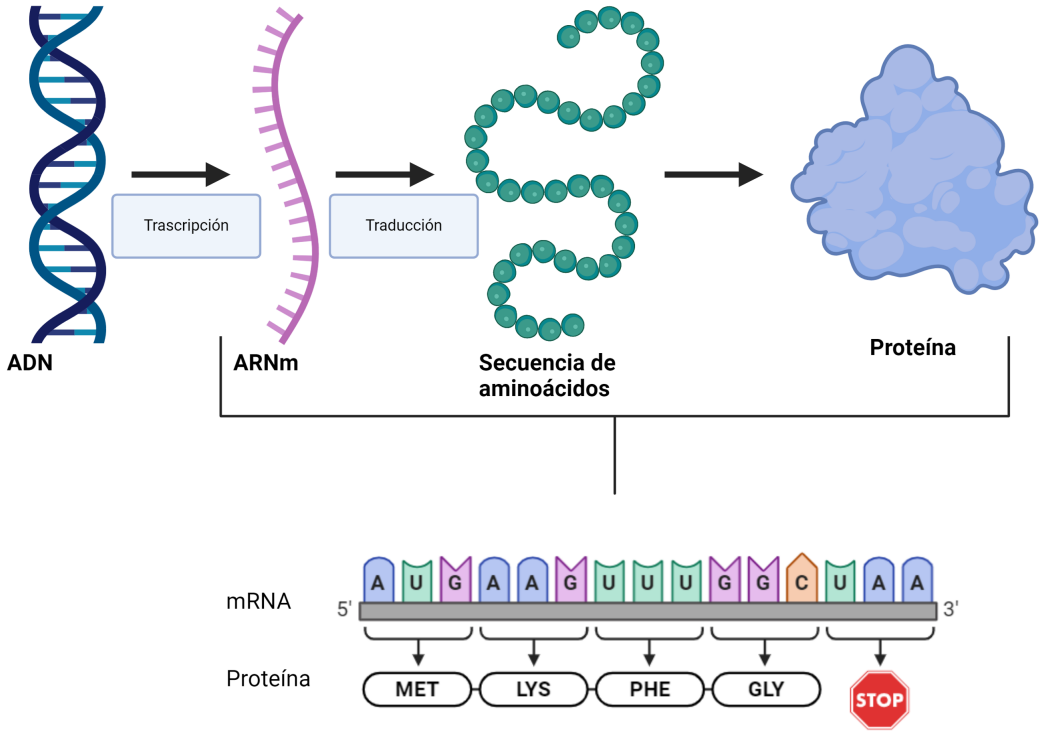


Figura 2.1: Inicio del proceso de expresión génica. La transcripción, permite al ADN ser utilizado como base para generar ARN mensajero (ARNm), este ARNm luego será traducido en una secuencia de aminoácidos, que se agrupan para formar una proteína funcional. En la parte inferior, se muestra un ejemplo de cómo la secuencia de codones en el ARNm determina el orden de los aminoácidos en la proteína, comenzando con metionina (MET) y terminando con un codón de parada (STOP), que indica el fin de la síntesis de proteínas [22][23].

Cada gen contiene información específica que permite la síntesis de una proteína, este proceso se lleva a cabo en el citoplasma de la célula mediante la traducción. Por cada conjunto de tres nucleótidos en el ADN, denominados codones, se codifica para un aminoácido en específico [23]. Comúnmente existen 20 aminoácidos que componen las proteínas en los seres humanos, y se encuentran divididos dependiendo de su clasificación química. Existen 11 aminoácidos polares subdivididos en 3 clases: (A) aminoácidos básicos con carga positiva, (B) aminoácidos ácidos con carga negativa y (C) aminoácidos polares sin carga. Además, existe una cuarta clase (D) que contiene 9 aminoácidos neutros no polares [24]. Los aminoácidos se irán uniendo en cadena para la formación de una proteína. El conjunto de aminoácidos en un orden específico es denominado polipéptido, por lo

tanto las proteínas son cadenas polipeptídicas. Cuando se realiza la unión y formación de las cadenas, estas darán la forma a la proteína resultante que determina su función dentro de la célula [25].

Al realizar la síntesis de proteínas el ADN se transcribe en una molécula de ARN mensajero (ARNm) transportando información genética desde el núcleo hasta los ribosomas del citoplasma [24]. En los ribosomas, los aminoácidos se unen para formar una cadena larga siguiendo las instrucciones del ARNm, utilizando el código genético de los codones para poder crear una cadena de aminoácidos con una secuencia específica (proteína) [24]. Una vez formada esta cadena se adopta una estructura tridimensional única que determina la función de la proteína, como transportar nutrientes, catalizar reacciones químicas, o actuar como componentes estructurales. Además, existen proteínas reguladoras que, en conjunto con los productos del ARN, permiten regular la expresión de varios tipos de genes. En este proceso, la información contenida en un gen se utiliza para crear moléculas de ARN que codifican para la síntesis de proteínas. Este mecanismo de control determina dónde, cuántas, y cuando se producen las moléculas de ARN y proteínas, influyendo así en la regulación de la actividad génica [26].

La expresión génica ocurre cuando un gen se activa o desactiva en el ADN, lo que permite que se utilice para obtener la proteína correspondiente. Se debe tomar en cuenta que, no todos los genes se activan en el mismo tiempo o lugar. Si un gen no se transcribe en una célula, este no podrá producir proteínas. Por otro lado, si un gen si se transcribe, es probable que se utilice para producir proteínas. La cantidad de transcripción de un gen depende por ejemplo de factores como, el ADN que se compacta alrededor de proteínas de soporte, ya que puede afectar la capacidad de un gen para transcribirse [26, 27].

Los factores de transcripción son responsables de regular la actividad génica al determinar qué genes están activos en cada célula del cuerpo humano (ver Fig. 2.2). Estos factores se enlazan a secuencias específicas del ADN, lo que influye en la capacidad de la ARN polimerasa para unirse al promotor, facilitando o bloqueando así la transcripción génica [28, 26].

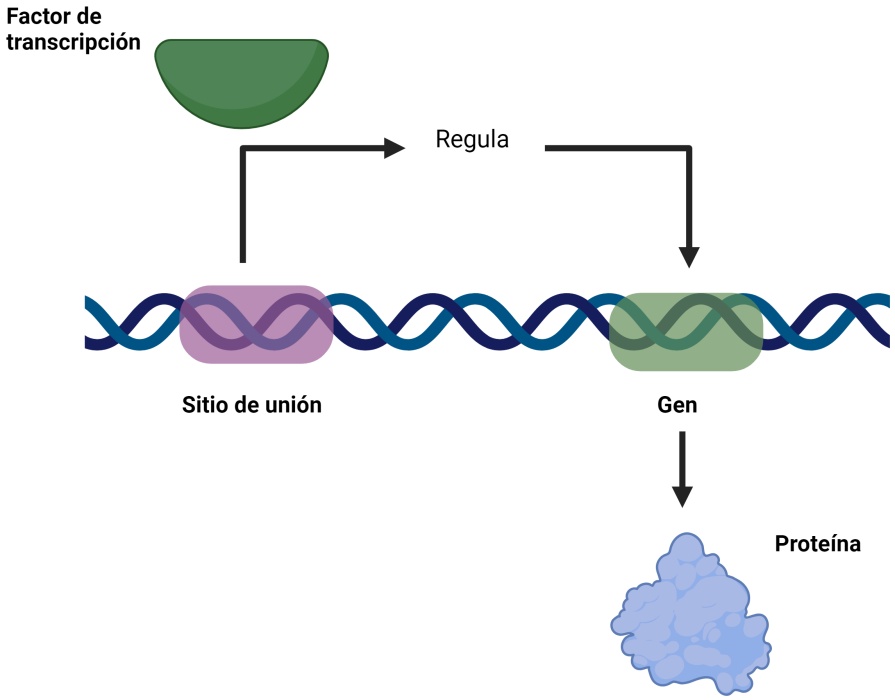


Figura 2.2: Representación de la interacción entre un factor de transcripción y un sitio de unión en el ADN. El factor de transcripción regula la expresión génica, lo que conduce a la síntesis de proteínas.

Al considerar la complejidad de los procesos biológicos que regulan la expresión génica y la síntesis de proteínas, es esencial también prestar atención a otros componentes fundamentales para la salud humana, como son las vitaminas. Aunque las vitaminas no participan directamente en la regulación génica, son cruciales para el mantenimiento de funciones corporales óptimas y la prevención de enfermedades.

Las vitaminas son compuestos orgánicos esenciales que el cuerpo necesita en cantidades pequeñas para llevar a cabo una variedad de funciones metabólicas y fisiológicas. Si bien muchas vitaminas se obtienen a través de la dieta, algunas de ellas, como la vitamina D, pueden ser sintetizadas directamente por la capacidad de nuestro organismo. Por ejemplo, la vitamina D se produce cuando la piel se expone a la luz solar, lo que desencadena una serie de reacciones químicas que culminan en la formación de esta vitamina en el cuerpo [29].

A pesar de la capacidad de sintetizar ciertas vitaminas internamente,

generalmente no se producen en cantidades suficientes para cubrir las necesidades diarias del cuerpo. Por lo tanto, es importante complementar la ingesta dietética con una variedad de alimentos ricos en vitaminas, como frutas, verduras, granos enteros y productos lácteos. En casos donde la dieta no proporciona suficientes vitaminas o en situaciones específicas que requieren dosis adicionales, se pueden recurrir a los suplementos vitamínicos [30].

La vitamina D desempeña un papel importante en diversas funciones fisiológicas y metabólicas. Se ha demostrado que influye en el sistema inmunológico, la salud cardiovascular, la función muscular y el bienestar mental [31]. Sin embargo, su deficiencia puede tener un impacto significativo en el organismo, afectando negativamente varias de estas funciones esenciales.

Para que la vitamina D pueda ser sintetizada por el cuerpo humano se requiere la exposición a luz solar, en este proceso la radiación ultravioleta (UVB) interacciona con el colesterol presente en la piel, desencadenando la conversión del 7-dehidrocolesterol en colecalciferol, conocido como vitamina D3. Este proceso de síntesis es esencial para mantener niveles óptimos de vitamina D en el organismo [32].

Esta transformación implica la apertura del anillo B del 7-dehidrocolesterol, formando el precolecalciferol, que rápidamente se convierte en colecalciferol. Una vez sintetizada, la vitamina D3 se libera al espacio extra celular y viaja al hígado unida a una proteína transportadora específica [33].

La forma activa de la vitamina D, el calcitriol, se une al receptor de la vitamina D (VDR), que a su vez forma un complejo con el receptor retinoide X (RXR) [34]. Este complejo se une a elementos de respuesta a la vitamina D (VDRE) en el ADN, en el sitio de unión DR3. Posteriormente, recluta co-activadores y el complejo de transcripción de la ARN polimerasa II (RNA POL II) para iniciar la transcripción de genes involucrados en la regulación del ciclo celular, apoptosis, inflamación y funciones del sistema inmunológico (ver Fig. 2.3[32].

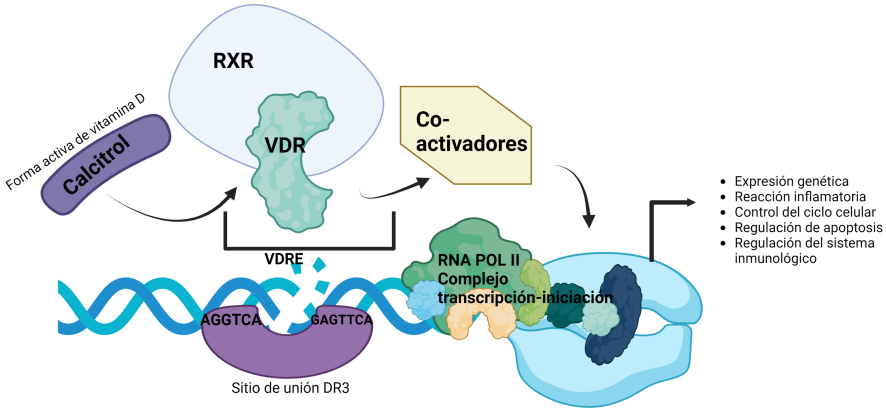


Figura 2.3: Representación esquemática de la activación del receptor de vitamina D (VDR) y su papel en la regulación de la expresión génica. Imagen basada en Carlberg, Carsten [32].

Una vez que el calcitriol, se une al receptor VDR se lleva a cabo sus efectos biológicos. La interacción entre la vitamina D y el VDR es fundamental para comprender cómo la vitamina D influye en una amplia gama de procesos fisiológicos y metabólicos en el cuerpo humano[34]. Cuando se realiza esta unión, se regula la actividad de ciertos genes que nos ayudan a absorber calcio y fósforo en el intestino, fortalece nuestros huesos y regula el equilibrio de calcio en la sangre y en los riñones [35][31].

La interacción entre el VDR y la vitamina D, así como las implicaciones biológicas resultantes son fundamentales para mantener la salud humana [36]. El gen VDR nos permite traducir las señales de la vitamina D y regula la expresión genética, lo que nos permite conocer como nuestro organismo está respondiendo ante la presencia de vitamina D [37].

La producción de la forma activa de la vitamina D, depende de varias hormonas y factores en el cuerpo. La hormona paratiroidea (PTH), junto con otros factores, estimula su producción en el riñón. Cuando hay niveles bajos de calcio en la sangre, el cuerpo produce más vitamina D activa para ayudar a mantener ese equilibrio. Pero si hay demasiado calcio o fósforo, la producción de vitamina D activa se reduce [37].

En esta tesis, se abordará la aplicación de técnicas computacionales para detectar genes cuyo comportamiento difiere significativamente entre dos condiciones diferentes. Este trabajo se enfoca en el caso de la vitamina D y utiliza criterios y métodos alternativos a los comúnmente aplicados en estudios genómicos, como los algoritmos de detección de anomalías, que permiten identificar patrones inusuales en los datos de expresión genética.

2.2. Biología molecular, expresión de genes e identificación de genes relevantes

En esta sección abordaremos aspectos clave de la biología molecular, centrándonos en la expresión de genes y la identificación de genes relevantes en diversos contextos biológicos. Comenzaremos profundizando en el concepto de gen, luego exploraremos el papel de los factores de transcripción y otros elementos relevantes en la regulación génica.

Como se mencionó en el anterior capítulo los genes son la base sobre la que se constituyen todos los organismos vivos, están formados por secuencias de ADN compuestas por cuatro bases nitrogenadas que crean esta molécula de doble hélice. La forma de la secuencia de las bases (adenina, timina, citosina y guanina) es lo que codifica la información genética [26].

2.2.1. Los genes codifican proteínas

Cada gen funciona como una guía con las instrucciones necesarias para la producción de moléculas que permiten desempeñar el funcionamiento de las células, principalmente proteínas. Las funciones que desempeñan las proteínas pueden ir desde la construcción de estructuras celulares, aceleración de reacciones químicas o la regulación de procesos [38].

El proceso del transporte de información de un gen a una proteína consta de dos etapas principales: transcripción y traducción. En la transcripción, el ADN de un gen es copiado a una molécula de ARN mensajero (ARNm), después en la traducción este ARNm permitirá el ensamble de la cadena de aminoácidos para formar una proteína [38].

Los factores de transcripción son proteínas que regulan la expresión génica. No todos los genes de una célula pueden estar activos al mismo tiempo, su activación o represión depende de las necesidades de la célula y de señales internas o externas. Los factores de transcripción actúan como interruptores, encendiendo o apagando la expresión de genes específicos según sea necesario [26].

El funcionamiento de los factores de transcripción se basa en la unión a secuencias específicas del ADN, conocidas como promotores o elementos reguladores, éstos están ubicados dentro del gen que regulan [39]. Al realizarse la unión a estas secuencias, los factores pueden atraer o inhibir a las proteínas y otros componentes celulares que son necesarios para comenzar la transcripción del gen en ARN mensajero [26].

La regulación del momento, lugar y nivel de expresión de un gen depende de los factores de transcripción. Estos elementos son esenciales para el desarrollo, la especialización celular y la adaptación de las células a su entorno, y juegan un papel fundamental en la biología y la salud [39].

Para entender cómo un tratamiento específico influye en la actividad de los factores de transcripción y, en consecuencia, en la expresión génica, es necesario realizar experimentos controlados. Es fundamental que estos experimentos incluyan condiciones adecuadamente controladas y un número suficiente de réplicas para medir con exactitud los cambios en la expresión génica inducidos por el tratamiento, diferenciando claramente entre los efectos directos y las variaciones naturales [40].

2.2.2. Diseños experimentales

En el proceso del diseño experimental en investigaciones científicas es fundamental la obtención de resultados confiables y reproducibles, ya que se busca medir el efecto de un tratamiento [41]. Un diseño bien estructurado debe considerar la comparación entre dos condiciones principales: control y tratamiento [42].

En la condición del grupo de control no se aplica ninguna clase de tratamiento o intervención. Este grupo sirve como referencia para determinar

los cambios que ocurren únicamente debido al tratamiento en cuestión. Por otro lado en la condición de tratamiento, al grupo se aplica la intervención que se desea estudiar o el tratamiento indicado, cualquier cambio observado en este grupo, en comparación con el grupo de control, se puede atribuir al efecto del tratamiento [42].

Para garantizar que los resultados obtenidos son válidos y no se deben al azar o por variabilidad natural, es importante realizar varias réplicas de cada condición. Las réplicas son experimentos repetidos bajo las mismas condiciones y permiten capturar la variabilidad biológica o experimental [41]. Cuando se realiza el análisis de los datos de múltiples réplicas, es posible realizar inferencias estadísticas más robustas y confiables, lo que permite mayor confianza en las conclusiones obtenidas [41].

Además, para el diseño experimental se debe tomar en cuenta factores diversos como pueden ser: el tamaño de la muestra, el control de variables externas y el uso de métodos apropiados para el análisis y la interpretación de los resultados. Un experimento bien realizado permite no sólo identificar si un tratamiento tiene un efecto, sino también cuantificar la magnitud de ese efecto y evaluar su relevancia biológica [43].

Capítulo 3

Técnicas de identificación de genes relevantes

3.1. Expresión diferencial de Genes y RNA-seq

La identificación de genes relevantes es importante dentro de la investigación biológica y médica, ya que permite comprender los mecanismos moleculares asociados a diversas condiciones fisiológicas y patológicas. En este contexto,la expresión diferencial de genes es fundamental ya que permite identificar genes cuya expresión varía significativamente entre diferentes condiciones , como la respuesta a un tratamiento aplicado y permite comprender los procesos biológicos que se están analizando.

Entre las técnicas más utilizadas para este propósito se encuentran la secuenciación de ARN (RNA-seq), que proporciona una visión detallada y global de la actividad génica [8].

La expresión diferencial de genes busca identificar genes que muestren cambios significativos en sus niveles de expresión entre dos o más condiciones, lo que permite detectar genes implicados en procesos biológicos específicos, como la respuesta a enfermedades, efectos de algún medicamento o la adaptación a algún cambio específico [44].

El proceso inicia con la cuantificación de los niveles de expresión génica. Después se pueden utilizar técnicas como la reacción en cadena de la polimerasa en tiempo real con transcripción reversa (RT-PCR en tiempo

real), esta técnica de laboratorio permite detectar los cambios en un gen o cromosoma, o identificar la activación de ciertos genes [45].

Aunque técnicas convencionales como la RT-PCR en tiempo real permiten analizar cambios en la expresión génica, su alcance es limitado, ya que requieren conocimiento previo de los genes de interés y análisis simultáneos. En esta tesis, se emplean algoritmos de detección de anomalías para explorar patrones genéticos inusuales en la expresión génica, con el fin de identificar genes que podrían estar implicados en la respuesta a la vitamina D de manera más compleja.

3.1.1. Análisis RNA-seq

Con el avance de la tecnología se desarrollaron nuevas técnicas, como el RNA-seq, que permitió realizar los análisis de expresión diferencial de manera eficiente. El RNA-seq permite cuantificar la expresión de todos los genes que se encuentran presentes en las muestras y no existe la necesidad de un previo conocimiento de genes relevantes, este nuevo enfoque amplia las posibilidades dentro del área de investigación genética [44].

El análisis se realiza de forma completa al conjunto total de moléculas de ARN presentes en una célula o tejido. Además, utilizando RNA-seq, no solo permite la identificación y cuantificación del ARN mensajero, si no que también de otros tipos de ARN no codificadores, como el ARN ribosómico (ARNr), el ARN de transferencia (ARNt) y microARN, entre otros [6].

El proceso ARN-seq comienza extrayendo el ARN total de la muestra para después convertirse en ADN complementario mediante la enzima transcriptasa inversa. Cuando este ADN complementario se fragmenta permite su adaptación para la secuenciación en una plataforma de secuenciación [46]. Los datos generados por esta técnica consisten en millones de lecturas cortas que corresponden a fragmentos del ARN original, una vez obtenidas estas lecturas se alinean con un genoma de referencia para reconstruir el transcriptoma [8].

El RNA-seq posee una capacidad que facilita la identificación y cuan-

tificación de genes que poseen baja abundancia de transcritos [6]. Esta es una ventaja que permite identificar genes que podrían haber pasado desapercibidos con técnicas más simples.

Una vez que se obtienen los resultados de RNA-seq, el siguiente paso es realizar el análisis de expresión diferencial. Este análisis permite identificar qué genes se expresan de manera diferente entre las condiciones experimentales y requiere varios pasos bioinformáticos. El proceso comienza con el preprocesamiento de los datos, en el cual es fundamental asegurar la calidad de las lecturas. Para esto, se utilizan criterios como la eliminación de lecturas de baja calidad (por ejemplo, aquellas con secuencias demasiado cortas), y se lleva a cabo un alineamiento de las lecturas con el genoma de referencia [47].

Se utilizan herramientas como FastQC para evaluar la calidad de las lecturas [48] y herramientas de alineación como STAR para mapear las lecturas con alta precisión [49]. Estas etapas son necesarias para asegurar que los datos sean representativos, gracias a estos pasas se puede proceder en el análisis de expresión diferencial sin fallas en la calidad de las lecturas.

Después del preprocesamiento se debe realizar la cuantificación de la expresión génica donde las lecturas alineadas se cuentan para cada gen y se genera un perfil de expresión génica. Este proceso puede utilizar herramientas que permitan asignar las lecturas a genes específicos basándose en la ubicación que estos tienen dentro del genoma [50].

La cuantificación de la expresión génica es un paso importante dentro de los análisis de RNA-seq, ya que permite comparar la abundancia de transcritos existentes entre diferentes condiciones o muestras [50]. Una vez que se resuelve la parte de la cuantificación se necesita comenzar con la normalización de los datos, este paso es de suma importancia para corregir las variaciones técnicas entre muestras, un ejemplo puede ser, diferencias en la secuenciación [51].

Para normalizar los datos de expresión, se utilizan varias métricas, siendo TPM (*Transcripts Per Million*) [51], RPKM (*Reads Per Kilobase of transcript per Million mapped reads*) [52], y FPKM (*Fragments Per Kilobase of transcript per Million mapped reads*) [53] las más comunes.

TPM es una medida que normaliza tanto la longitud del gen como el número total de lecturas en una muestra, lo que facilita la comparación de la expresión génica entre diferentes genes y muestras [51]. A diferencia de RPKM y FPKM, TPM ajusta primero la longitud del gen y luego calcula la abundancia relativa, asegurando que la suma de total de las TPM en una muestra siempre sea un millón, lo que hace que TPM sea especialmente útil cuando se comparan niveles de expresión entre diferentes muestras o condiciones [54].

Primero se cuenta cuantas lecturas de RNA-seq se alinean a cada gen, luego estas cuentas se separan por la longitud del gen, esto sirve para corregir el hecho de que genes más largos usualmente posean una cantidad más alta de lecturas solo por el simple hecho de ser más largos. Finalmente, las cuentas de las lecturas se ajustan de tal forma que la suma total de todos los genes dentro de una muestra sea un millón. Este proceso permite realizar la comparación de la expresión de un gen entre diferentes muestras de una forma más práctica y sencilla [51].

El método RPKM se usa cuando las lecturas de RNA-seq provienen de una sola hebra de ADN. Este método también normaliza la longitud del gen y el número total de lecturas en la muestra, lo que facilita la comparación de la expresión génica. Sin embargo, RPKM no ajusta de la misma manera las diferencias en la longitud de los genes, lo que puede dificultar la comparación entre genes de diferentes tamaños [54].

Por otro lado, también existe el método FPKM, el proceso es muy similar a RPKM, solo que se usa en estudios de RNA-seq donde las lecturas provienen de ambas hebras de ADN. Al igual que en el método anterior, se realiza la cuenta de fragmentos y el ajuste de longitud de los genes, la diferencia yace que en lugar de utilizar las cuentas para que la suma total sea exactamente un millón como en TPM, en FPKM se realizará por la división del total de lecturas de la muestra [53]. Cuando se realizan comparaciones de expresión génica en diferentes muestras es más recomendable la utilización del método TPM[54].

Estas métricas son esenciales para interpretar correctamente los datos de RNA-seq y son comúnmente utilizadas en combinación con herramientas como *DESeq2* y *edgeR* para identificar genes diferencialmente expresados.

En el análisis realizado, se utilizaron datos de RNA-seq en formato TPM para cuantificar la abundancia de genes.

3.1.2. Pruebas estadísticas: DESeq2 y edgeR

Después de completar el proceso de normalización, se procede con las pruebas estadísticas. Estas se aplican para determinar si las diferencias en la expresión génica entre las condiciones observadas son estadísticamente significativas. Las herramientas más comunes en este contexto son DESeq2 y edgeR, que analizan los datos de expresión y realizan pruebas para identificar genes con cambios significativos en su expresión entre las diferentes condiciones [55].

DESeq2 es una herramienta utilizada en el área bioinformática que permite analizar datos de expresión génica obtenidos de RNA-seq. Con esta herramienta, se busca identificar genes que estén diferencialmente expresados entre dos o más condiciones como puede ser el caso de control y tratamiento. Esto funciona gracias al conteo de lecturas de RNA-seq, el ajuste por factores como tamaño de muestras y la dispersión para realizar las pruebas estadísticas determinadas que logran identificar si las diferencias observadas son significativas [56].

Otra herramienta altamente utilizada para el análisis de expresión diferencial es edgeR, que también se basa en la utilización de los datos de RNA-seq. Al igual que el método anterior DESeq2, edgeR permite comparar los niveles de expresión génica entre ambas condiciones identificando los genes que presenten en su expresión génica diferencias entre las condiciones. EdgeR utiliza un enfoque basado en la distribución binomial negativa que modela la variabilidad de los datos y realiza pruebas estadísticas para identificar que genes están regulados de manera diferente entre las condiciones observadas [57].

La distribución binomial negativa es una herramienta de la estadística que nos permite calcular la probabilidad de cuántos fracasos ocurrirán antes de obtener un número fijo de éxitos en un experimento con intentos repetidos [58]. A diferencia de la distribución binomial, utilizada para saber cuántos éxitos tendremos en un número fijo de intentos, la distribución

binomial negativa se usa cuando queremos alcanzar una cantidad predefinida de éxitos sin un número fijo de intentos.Esta particularidad permite su aplicación en diversos campos, como puede ser, la Ingeniería, Economía, o en este caso en el área de la Biología. La fórmula de la distribución binomial negativa nos permite obtener la probabilidad de obtener un número específico de fracasos antes de alcanzar los casos de éxitos deseados [59].

Esta distribución es útil en situaciones en la que los casos de éxitos no siguen una secuencia predecible, como en experimentos, en producción, estudios, análisis de riesgo, etc. También es de utilidad cuando no se conoce cuántos intentos necesitaremos para lograr un resultado deseado. Esta herramienta nos ayuda a entender mejor el comportamiento de los fracasos y éxitos en repetidos ensayos y lograr un análisis más preciso para las variables discretas [58].

Cuando se utiliza un enfoque basado en la distribución binomial negativa, es importante considerar que los datos de expresión génica pueden tener una variabilidad mayor de la que una distribución binomial simple podría manejar. Esto es común en los datos de RNA-seq, donde el número de lecturas alineadas a cada gen puede variar significativamente [60].

EdgeR modela esta variabilidad asumiendo que las cuentas de lecturas para cada gen siguen una distribución binomial negativa. Este enfoque permite tener en cuenta tanto las diferencias en la secuenciación como las variaciones en la expresión génica [60]. Usando esta técnica, edgeR realiza pruebas estadísticas, como la prueba de dispersión común o la prueba de la razón de verosimilitud, para determinar si las diferencias observadas en la expresión génica entre condiciones son significativas [57].

Capítulo 4

Algoritmos de detección de anomalías: K-medias y LOF

K-medias es un algoritmo de agrupamiento que permite organizar o agrupar los datos en un número específico de grupos o clusters basándose en la similitud que comparten entre ellos (ver Fig. 4.1) [61].Este algoritmo es utilizado por su simplicidad y eficacia en las predicciones de clasificación de grandes volúmenes de datos. El proceso comienza seleccionando aleatoriamente "k"puntos, que actuarán como los centros iniciales de los grupos. Luego, el algoritmo a cada dato le asignará un grupo con el centro más cercano, basado en la distancia euclidiana [62],basada en la siguiente fórmula:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{4.1}$$

Donde p y q son dos puntos en un espacio de n dimensiones, y p_i y q_i representan las coordenadas de los puntos p y q en la dimensión i.

Una vez que se realiza la asignación de puntos a los grupos, se actualizan los centros de los grupos obteniendo el promedio del total de los puntos que pertenezcan a cada grupo, esto permite que los centros reflejen mejor la distribución de los datos con cada iteración. El algoritmo repetirá los pasos de asignación y actualización hasta que los centros de cada grupo ya

no presenten variaciones, o en su caso sea mínima la diferencia y no afecte de forma significativa el resultado final del agrupamiento [63].

El objetivo de este algoritmo es minimizar la distancia existente de cada punto a su centro asignado, permitiendo así que los puntos que pertenecen a un mismo grupo sean lo más similares posibles, mientras que los demás grupos estén bien diferenciados entre sí [62].

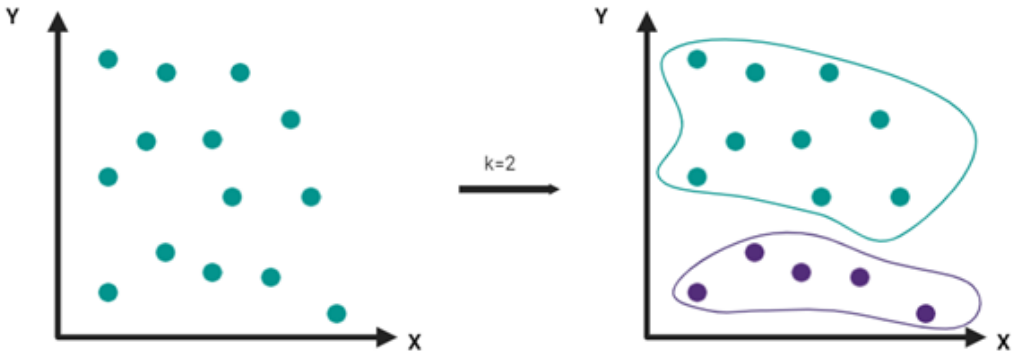


Figura 4.1: Ejemplo de detección de cúmulos utilizando k= 2 en K-medias.

El algoritmo LOF (Local Outlier Factor) es utilizado para detectar anomalías o valores atípicos en un conjunto de datos. A diferencia de otros métodos que solo identifican los puntos alejados del resto mediante distancias simples, LOF basa su enfoque en la densidad de los datos utilizados [64]. Esto le permite identificar si un punto es anómalo dependiendo de las diferencias que se presenten de acuerdo con su densidad local en comparación con sus vecinos cercanos [12]. Por lo tanto este algoritmo puede ser más efectivo en contextos donde los datos tienen una estructura local compleja, como en el análisis de datos genéticos .

El funcionamiento de LOF comienza evaluando cada punto y su entorno, es decir, los puntos más cercanos o puntos vecinos. Para ello, se define un valor de k, que representa el número de vecinos más cercanos que serán considerados para cada punto. El algoritmo después, obtiene la distancia entre un punto específico y sus vecinos más cercanos utilizando igualmente como métrica la distancia euclidiana [65].

Una vez identificados los vecinos de cada punto, LOF mide la densidad local. La densidad local de un punto se calcula analizando las distancias existentes entre vecinos. La idea es que los puntos en una región densamente poblada (donde los puntos vecinos presenten una distancia menor) será considerados normales, mientras que un punto en una región menos densa (donde los vecinos están más alejados) podría ser anómalo (ver Fig. 4.2)[12].

Aquí es donde el algoritmo LOF presenta un factor importante: el factor de anormalidad local o grado de anormalidad. Este factor compara la densidad local de un punto con la densidad local de sus vecinos [65]. Así es como clasificará dependiendo del resultado obtenido, siendo catalogados como puntos anómalos o puntos normales [66].

El valor del LOF varía dependiendo del resultado del análisis de densidades. Un valor cercano a 1 indica que el punto tiene una densidad similar a la de sus vecinos y, por lo tanto, es clasificado como normal. Sin embargo, cuando el valor del LOF es significativamente diferente, significa que el punto está en una región de menor densidad en comparación con sus vecinos, lo que indica una mayor probabilidad de que sea una anomalía y es etiquetado como -1 [12].

Este enfoque tiene grandes ventajas sobre otros métodos de detección de anomalías, ya que no solo se basa en la distancia entre los puntos, sino que también considera las relaciones entre los puntos cercanos [67]. Por ejemplo, en un conjunto de datos donde algunas áreas son densas y otras están más dispersas, un punto puede parecer normal en una zona dispersa pero resultar anómalo en una zona densa. LOF puede hacer esta diferenciación porque trabaja de forma local, adaptándose al entorno y características específicas de cada punto [65].

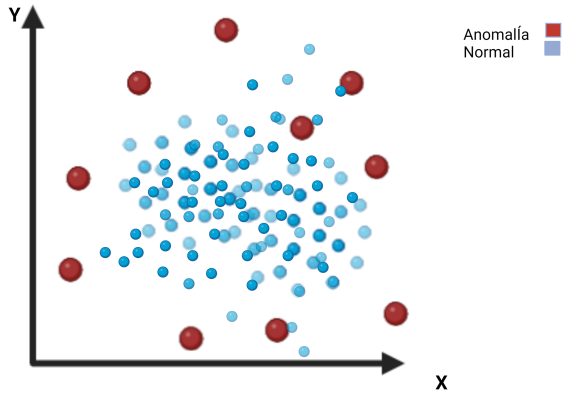


Figura 4.2: Ejemplo de detección de anomalías utilizando el algoritmo LOF basado en densidades.

En resumen, LOF permite detectar puntos que se desvían del patrón de densidad local, analizando a sus vecinos más cercanos para determinar si su comportamiento es anómalo. Es especialmente útil cuando los datos están distribuidos de forma irregular, ya que otros métodos más sencillos para detectar anomalías no logran captar la información local con la misma eficacia [12].

Capítulo 5

Metodología

En esta sección, se describe el proceso experimental y los procesos analíticos llevados a cabo para identificar genes con patrones de expresión anómalos en respuesta al tratamiento con vitamina D. El objetivo principal de este estudio fue detectar genes dónde sus expresiones presentarán variaciones significativas entre las condiciones de control y tratamiento, utilizando diversas técnicas como algoritmos de detección de anomalías.

El análisis comienza con el preprocesamiento de los datos de expresión génica obtenidos mediante secuenciación de ARN (RNA-seq) en formato TPM (transcritos por millón), seleccionando tres réplicas biológicas para cada condición. Estos datos proporcionan una medida de la cantidad de ARN mensajero (ARNm) presente en las células bajo diferentes condiciones experimentales.

Es importante que los datos se encuentren expresados en TPMs, por que normaliza la longitud del gen y el número de lecturas totales en cada muestra lo que facilita la comparación de los genes bajo distintas condiciones. Sin TPM los datos de RNA-seq podrían ser difíciles de interpretar debido a diferencias en la secuenciación o la longitud de los genes.

El enfoque implementado permite no solo comparar los genes en condiciones distintas como control y tratamiento, sino también utilizar métodos no tradicionales, como algoritmos de aprendizaje no supervisado, para identificar genes con comportamientos anómalos. A lo largo del proceso, se utilizaron métodos basados en distancias geométricas como técnicas de agrupamiento, lo que permitió visualizar y clasificar la expresión génica de

una nueva forma.

Posteriormente, estos datos fueron transformados en datos composicionales para facilitar su análisis. A partir de esta transformación, se generaron distancias entre los genes en ambos grupos, las cuales fueron utilizadas como entrada para los algoritmos de detección de anomalías, como LOF y K-medias.

El enfoque busca nuevas estrategias basadas en geometría y teoría de probabilidades, proporcionando una nueva perspectiva sobre la respuesta genética al tratamiento con vitamina D (25(OH)D3), precursor de esta vitamina. A continuación, se detallan los pasos seguidos en el análisis de este trabajo.

5.0.1. Datos de expresión génica

En esta tesis en particular, se trabajó con muestras de células mononucleares de sangre periférica (PBMC), una célula del sistema inmunológico, se utilizó la base de datos elaborada por Hanel A. et al [68]. Los datos de expresión génica se obtuvieron de 12 individuos tratados por 24 horas para un grupo de control y un grupo que recibió un tratamiento basado en la forma activa de la vitamina D, 25(OH)D3. Este tipo de experimento es necesario si se quiere investigar cómo los genes responden a un tratamiento específico, en este caso, el de vitamina D.

Se seleccionaron tres réplicas o muestras para cada condición experimental, tomadas de la expresión genética un individuo: tres para el grupo de control y tres para el grupo de tratamiento, estas replicadas fueron tomadas a lo largo de tres semanas exactas. Las réplicas son fundamentales en estudios de expresión génica porque permiten medir las diferencias que ocurren de manera natural entre muestras en un experimento, estas diferencias pueden venir de diversos factores, como puede ser la genética, entorno o cambios en el proceso experimental [69]. Las muestras de control representan la condición normal o base, mientras que las muestras de tratamiento reflejan cómo las células responden a la administración de vitamina D.

Tener tres réplicas por condición permite medir la consistencia de los

resultados y asegurar que los cambios observados en la expresión génica no se deban al azar o a simples errores a la hora de la captura de los datos. Esto proporciona una base más sólida para interpretar cómo los genes realmente responden al tratamiento. Al usar TPM como formato de expresión la comparación entre réplicas se realiza de manera uniforme, eliminando posibles confusiones, como el número total de lecturas por muestra. Así, se pueden hacer análisis más precisos sobre la respuesta de los genes al tratamiento con vitamina D.

Una vez que los datos de expresión génica se encuentran normalizados en TPM y ya se seleccionaron las réplicas exactas por condición, el siguiente paso es convertirlos en datos composicionales. Este proceso consiste en tomar las réplicas de cada condición, tanto del grupo de tratamiento como del control y, para cada gen, sumar la expresión de todas las réplicas. Posteriormente, cada réplica se divide entre esa suma total, lo que transforma los datos en proporciones (ver Fig 5.1), esto se realiza para los 12333 genes del individuo. Estas proporciones permiten que cada réplica de datos se interprete como una distribución de probabilidad.

	Tratamiento			Control		
	S1	S2	S3	S1	S2	S3
Gen 1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$C_{1,1}$	$C_{1,2}$	$C_{1,3}$
Gen 2	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$C_{2,1}$	$C_{2,2}$	$C_{2,3}$
Gen N	$X_{n,1}$	$X_{n,2}$	$X_{n,3}$	$C_{n,1}$	$C_{n,2}$	$C_{n,3}$

Ec. De conversión de los datos de expresión génica a datos composicionales

$$C'_{n,m} = \frac{C}{\sum_{m=1}^{m=3} C_{n,m}} \quad x'_{n,m} = \frac{x_{n,m}}{\sum_{m=1}^{m=3} x_{n,m}}$$

	Tratamiento			Control		
	S1	S2	S3	S1	S2	S3
Gen 1	$X'_{1,1}$	$X'_{1,2}$	$X'_{1,3}$	$C'_{1,1}$	$C'_{1,2}$	$C'_{1,3}$
Gen 2	$X'_{2,1}$	$X'_{2,2}$	$X'_{2,3}$	$C'_{2,1}$	$C'_{2,2}$	$C'_{2,3}$
Gen N	$X'_{n,1}$	$X'_{n,2}$	$X'_{n,3}$	$C'_{n,1}$	$C'_{n,2}$	$C'_{n,3}$

Figura 5.1: Proceso de generación de los datos composicionales a partir de los datos de expresión génica entre las condiciones de control y tratamiento. Donde el resultado está dado por la suma de las 3 réplicas de cada condición, y la división de una réplica en particular entre la sumatoria de cada condición respectivamente.

Trabajar con datos composicionales es particularmente útil en estudios de expresión génica porque lo que se mide ya no es una expresión absoluta de los genes, sino cómo se expresa un gen en comparación con otros genes en la misma muestra. Este enfoque asegura que los resultados sean consistentes y comparables.

Una vez que los datos se transforman en datos composicionales, cada gen puede visualizarse como un punto en el simplex de probabilidad o complejo simplicial, un espacio geométrico donde se representan todas las posibles combinaciones de distribución de probabilidad entre los genes (ver Fig. 5.2). Este enfoque geométrico es clave porque facilita la comparación de las expresiones génicas entre las condiciones de control y tratamiento, permitiendo ver de manera más clara cómo varían los patrones de expre-

sión.

El simplex de probabilidad es lo que nos permite analizar la expresión genética de manera intuitiva. Esta representación permite visualizar la distribución de las probabilidades de los genes en ambas condiciones. La posición de cada punto del simplex está determinada por las proporciones de expresión en las réplicas seleccionadas para cada condición, estas proporciones son proyectadas en el simplex en forma de triángulo en el caso de tres réplicas, donde cada vértice representa una de estas réplicas como se muestra en la figura 5.2.

Por ejemplo, si un gen presenta una expresión similar en las tres réplicas del control, pero cambia significativamente en el grupo de tratamiento en comparación con variaciones sutiles, este gen ocupará diferentes posiciones dentro del simplex, lo que indicará un comportamiento atípico. Estas diferencias son importantes para identificar genes anómalos o con una respuesta significativa al tratamiento.

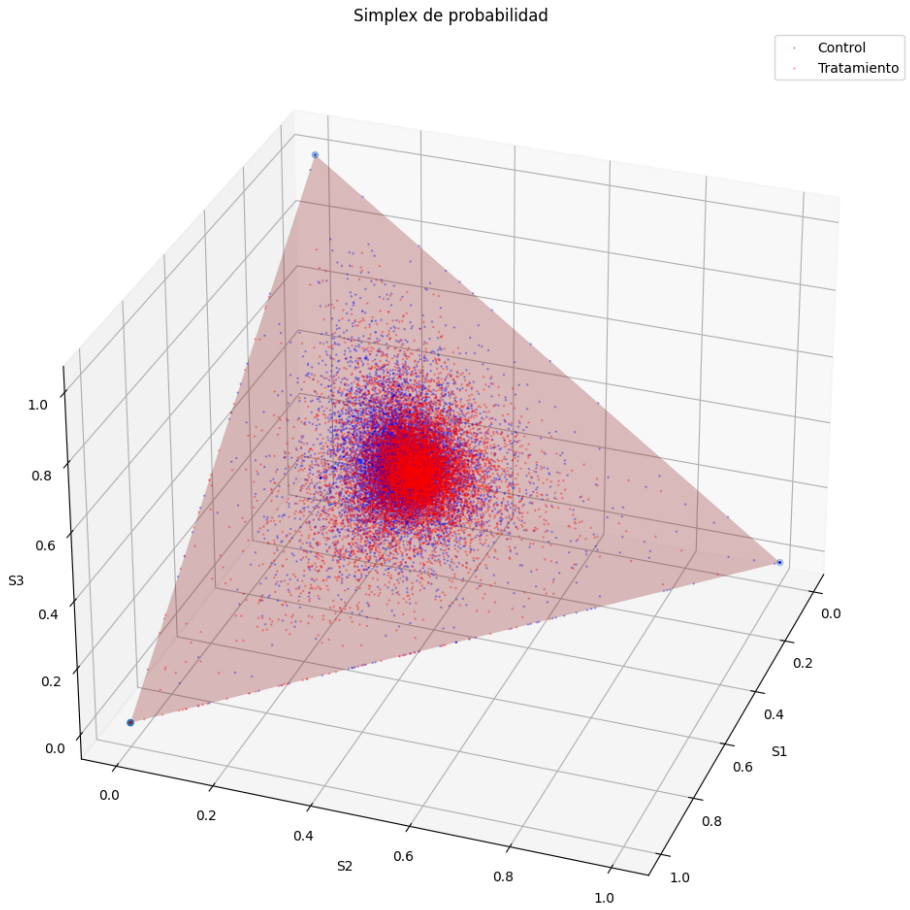


Figura 5.2: Comparación de los simplex de probabilidad para visualizar el comportamiento de los datos. Los datos de control están denotados por el color azul y los datos de tratamiento por el color rojo. La estructura del triángulo y la distribución de los datos nos proporciona información sobre la composición de los grupos y las distancias proyectadas.

5.0.2. Distancias proyectadas

En el siguiente paso del análisis, la distancia entre la posición de un gen en el grupo de control y su posición en el grupo de tratamiento se mide utilizando la distancia de Wasserstein, una métrica que nos ayuda a cuantificar las diferencias entre dos distribuciones de probabilidad. Estas distancias son fundamentales porque permite obtener una medida directa de cuán drásticamente cambia la expresión de un gen en respuesta a la vitamina D.

Finalmente, estas distancias sirven como entrada para los algoritmos de detección de anomalías, como LOF (Local Outlier Factor) y K-medias.

Estos algoritmos permiten identificar genes con comportamiento anómalo en función de sus patrones de expresión bajo el tratamiento con vitamina D.

La distancia de Wasserstein se eligió para este análisis por su capacidad única para comparar distribuciones de probabilidad de manera más precisa y efectiva que otras métricas. A diferencia de otras medidas, como la distancia euclidiana, la distancia de Wasserstein es especialmente útil cuando se trabaja con datos composicionales, ya que tiene en cuenta la forma en que los datos están distribuidos en su espacio geométrico, en este caso, el simplex de probabilidad, y no solo toma en cuenta que tan alejados se encuentran los datos unos de otros, considerando distribución y no solo la posición de éstos datos [70].

Existen ventajas destacables cuando se utiliza como métrica la distancia de Wasserstein, siendo una de las principales su capacidad para capturar no solo las diferencias entre los valores de expresión genética, sino también la manera de organización de los valores dentro de cada simplex de probabilidad [71]. Gracias a que los datos utilizados están expresados como distribuciones de probabilidad, al convertir las expresiones genéticas en proporciones, la distancia Wasserstein calcula el esfuerzo que se requiere para convertir una distribución en otra, es decir, tomará en cuenta cuantó y cómo se deben ajustar los valores para que las distribuciones a evaluar sean iguales. La distancia Wasserstein está dada por la siguiente formula [71].

$$W_p(P,Q) = \left(\inf_{J \in \mathcal{J}(P,Q)} \int |x-y|^p \, dJ(x,y) \right)^{1/p} \tag{5.1}$$

La distancia de Wasserstein es una medida que nos permite comparar dos distribuciones **P** y **Q**. Se puede entender como el costo mínimo de esfuerzo que se necesita para mover una distribución a la otra, este costo se calcula como la función de la cantidad de masa a mover y la distancia a desplazar de dicha masa (ver Eq.5.1). Para lograr esta transformación se busca todas las posibles formas de emparejamiento entre puntos de ambas distribuciones [72].

Los emparejamientos están dados por $J(x,y)$ y la distancia existente entre dos puntos se elevará a la potencia p, esto permite proporcionar una medida del costo de mover un punto específico de la distribución **P** a un punto específico de la distribución **Q**. Luego se calcula el costo total de mover la masa entre ambas distribuciones mediante una integral, y finalmente nos quedamos con el menor valor posible de costo entre todas las posibles maneras de emparejar los puntos de las distribuciones (ver Fig. 5.3) [72].

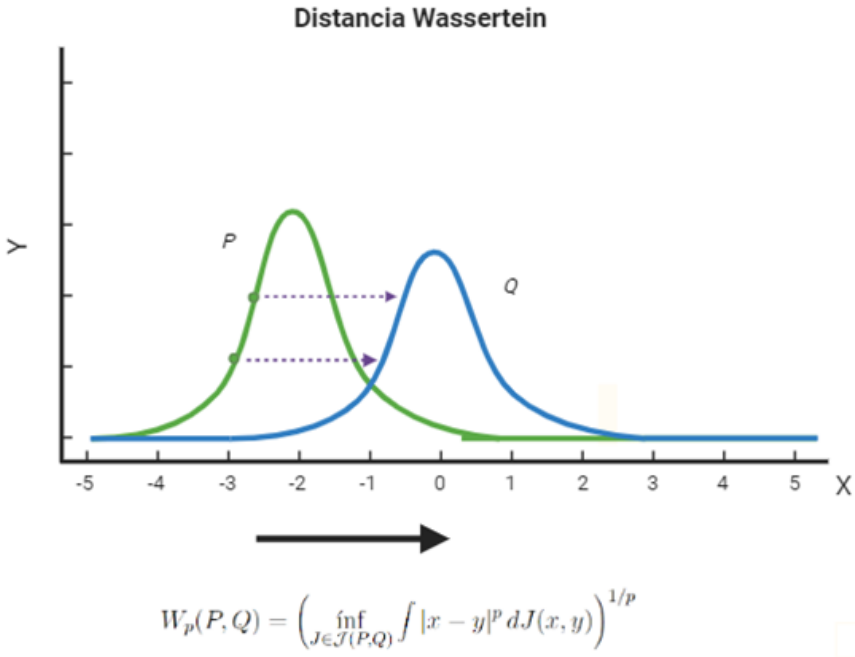


Figura 5.3: Representación gráfica de la distancia de Wasserstein entre dos distribuciones de probabilidad P y Q. La distancia entre los puntos indica la cantidad mínima de trabajo requerida para transformar una distribución en otra. Imagen basada en Nakazato e Ito (2021) [73].

En el análisis de los resultados obtenidos de las distancias Wasserstein entre los grupos de control y tratamiento encontramos que la mayor parte de las distancias se encuentran muy cercanas a 0, lo que indica que la diferencia que se percibe entre ambos grupos es pequeña o poco relevante. Las distancias cercanas a 0 podrían indicar que los datos tienden a ser bastante similares en muchas de las mediciones.

A medida que las distancias aumentan, la frecuencia de las observa-

ciones disminuye, esto indica que existen pocos casos con distancias más grandes entre los grupos, reflejando en la Fig. 5.4, donde se reduce la frecuencia de estas distancias significativas, y se encuentran con mayor frecuencia los datos de distancias pequeñas.

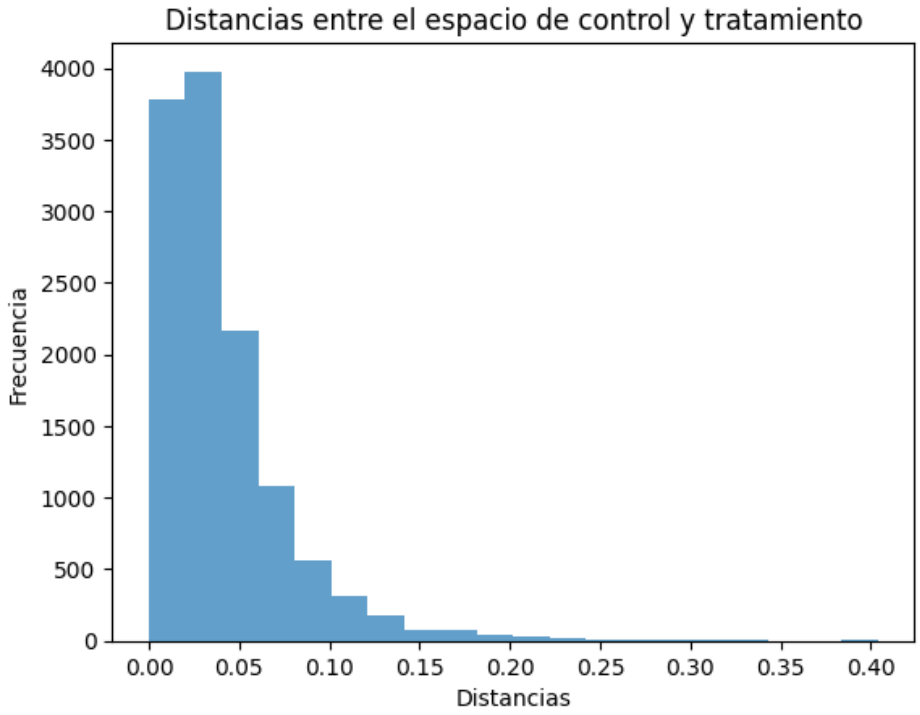


Figura 5.4: El histograma muestra las distancias existentes entre los grupos de control y tratamiento que son mayoritariamente pequeñas, lo que sugiere que las diferencias significativas entre los grupos son raras.

Detectar las variaciones entre las condiciones experimentales no es sencillo, las métricas tradicionales pueden tener un mayor margen de error, ya que los genes suelen mostrar cambios sutiles en sus niveles de expresión. Sin embargo, estos pequeños cambios pueden influir de manera significativa en la distribución general de los puntos, por lo que la distancia de Wasserstein es más precisa para detectarlos. Además, al considerar tanto la magnitud como la dirección del cambio, esta métrica permite captar de manera más completa la variación entre el grupo control y el grupo de tratamiento [70].

En esta tesis, la distancia de Wasserstein fue fundamental para alimentar los algoritmos de detección de anomalías. Estas distancias proporciona-

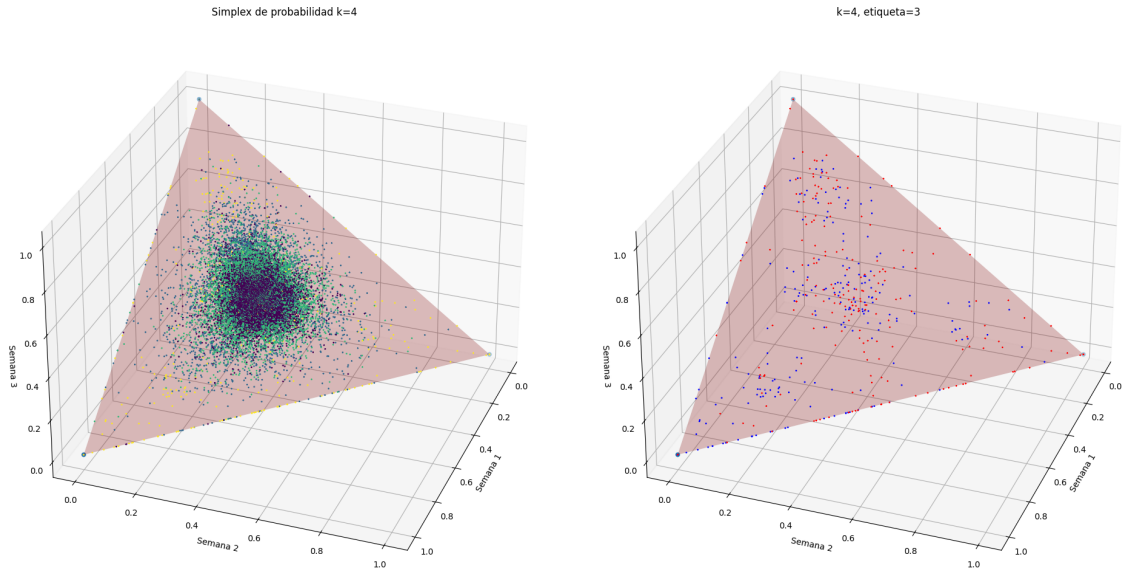
ron una medida cuantitativa de cómo varía la expresión de cada gen entre las dos condiciones. Cuanto mayor sea la distancia, mayor es la diferencia en el comportamiento de ese gen, lo que podría indicar una anomalía o un cambio significativo en su expresión.

5.0.3. Enfoques implementados

El primer enfoque aplicado fue el algoritmo de agrupamiento K-medias, que es una técnica de aprendizaje no supervisado utilizada para dividir un conjunto de datos en k grupos o agrupamientos, basándose en las similitudes encontradas entre los datos (Ver Fig. 4.1). En este caso, se utilizaron las distancias de Wasserstein como los puntos de entrada para el algoritmo, lo que permitió agrupar los genes según la importancia de los cambios en sus perfiles de expresión.

Se seleccionó un valor de k=4 basado en el método del codo, el cual sugirió que este número es el óptimo para el valor k. El método del codo permite identificar el punto donde añadir más agrupamientos ya no genera una variación significativa dentro de estos.

Esto resultó en cuatro diferentes agrupamientos de genes. Cada agrupamiento contenía genes con diferentes niveles de anomalía en su expresión. El grupo más pequeño, que contenía menos genes, fue el que correspondió a los genes con las mayores distancias de Wasserstein, lo que sugiere que estos genes presentan los cambios más drásticos en la expresión entre las condiciones de tratamiento y control (ver Fig 5.5. Estos genes fueron seleccionados para un análisis más detallado, ya que es posible que estén más directamente involucrados en la respuesta al tratamiento con vitamina D.



(a) (b)

Figura 5.5: (a) Representación del simplex de probabilidad que contiene todos los agrupamientos cuando k=4.(b) Representación del grupo más pequeño encontrado que contiene los genes menos comunes y su distancia Wasserstein es mayor.

Además de K-medias, se utilizó el algoritmo LOF (Local Outlier Factor) para complementar el análisis y expandir las opciones, esto permitió también detectar genes con comportamientos atípicos en sus perfiles de expresión de otra forma diferente. LOF es una técnica diseñada para identificar datos que muestran diferencias considerables en respecto a sus vecinos en un espacio de características, en este caso, las distancias de Wasserstein.

El algoritmo LOF se basa en la comparación de densidades locales para identificar puntos anómalos (outliers), que son aquellos que se encuentran en regiones de menor densidad en el espacio de los datos. En el eje X de la figura 5.6 se presentan las distancias de Wasserstein, que fueron las entradas para el análisis, mientras que en el eje Y se muestra el grado de anomalía generado por LOF, donde los valores más bajos corresponden a los datos considerados más anómalos (en naranja). Los puntos con mayor densidad se consideran normales (en azul).

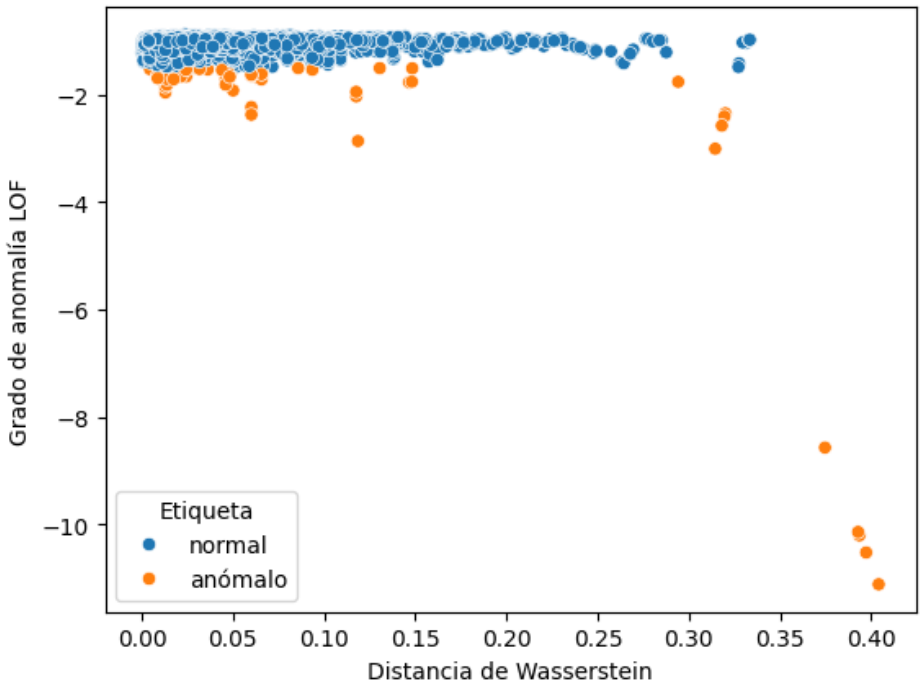


Figura 5.6: Gráfico que muestra la detección de anomalías utilizando el algoritmo Local Outlier Factor (LOF) con los datos utilizados en la tesis.

Existen varias ventajas al utilizar este algoritmo, como puede ser la facilidad con la que se adapta a los datos, permitiendo visualizar además del grado de anomalía de cada gen, realizar un conteo de la cantidad de genes anómalos existentes en comparación con genes con expresiones catalogadas como normales. En este análisis se aplicó el algoritmo LOF con diferentes valores de k, y se buscó quedarnos con el grupo de menor cantidad de genes anómalamente expresados. En la figura 5.7 se muestra una comparación entre la frecuencia de estos genes cuando el número de vecinos k es igual a 10.

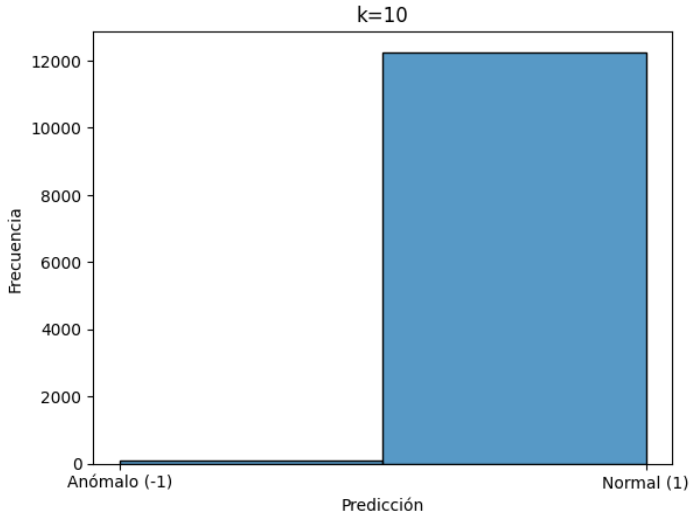


Figura 5.7: Representación de la distribución de los valores de anomalía asignados, donde 1 es valor típico, y -1 refleja un valor anómalo después de la implementación del algoritmo Local Outlier Factor(LOF) identificando los genes anómalos en base a su distancia Wasserstein.

Este enfoque permitió que utilizando los algoritmos de detección de anomalías, K-medias y LOF se pudieran detectar por medio de su análisis los genes con cambios significativos en su expresión. Estos genes serán identificados como genes cuyo comportamiento era atípico en relación con el resto de los datos.

En resumen, el objetivo fue identificar genes anómalos utilizando algoritmos de detección de anomalías. Los datos trabajados representaban la expresión génica que se obtuvo mediante secuenciación de ARN en formato TPM, lo cual normaliza las lecturas y facilita comparaciones. A partir de estos datos, se obtuvieron las distancias proyectadas dentro del simplex de probabilidad de las distribuciones de los datos, utilizando como métrica la distancia de Wasserstein, que mide las diferencias entre distribuciones de probabilidad.

Los algoritmos de K-medias y LOF se usaron para identificar genes con patrones de expresión anómalos. K-medias permitió el agrupamiento de los genes en función de sus distancias de Wasserstein, mientras que LOF detectó genes que mostraban diferencias en su comportamiento respecto a sus vecinos y densidades. En ambos casos, se identificaron genes con cambios significativos en respuesta al tratamiento con vitamina D, destacándose

aquellos con mayores distancias de Wasserstein como posibles candidatos para un análisis más detallado.

En conclusión, el enfoque geométrico y probabilístico, junto con las herramientas de detección de anomalías, permitió identificar genes con cambios relevantes en su expresión bajo tratamiento, resaltando su potencial implicación en la respuesta genética a la vitamina D.

Capítulo 6

Resultados

En este estudio, se aplicaron los algoritmos K-medias y LOF para identificar genes con patrones de expresión anómalos en dos condiciones diferentes: control y tratamiento. El análisis se realizó utilizando datos de expresión génica obtenidos de experimentos realizados en base al tratamiento 25(OH)D3, precursor de la vitamina D. La célula de donde los datos son obtenidos es PBMC (célula mononuclear de sangre periférica) una célula del sistema inmunológico, todo este conjunto de datos fue obtenido mediante RNA-seq.

El algoritmo K-medias se utilizó para agrupar los genes en función de sus perfiles de expresión en ambas condiciones. Se seleccionó el número óptimo de cúmulos utilizando el método del codo, el cual sugiere el número ideal de cúmulos donde la disminución en la inercia dentro del grupo comienza a estabilizarse [74]. Basado en este análisis, se seleccionaron un total de 4 distintos valores para k para asegurar una partición adecuada de los datos.

Cada cúmulo fue analizado en detalle para determinar las características de los genes agrupados en él. Se observó que los cúmulos más grandes, compuesto por un mayor porcentaje de los genes, mostraron niveles de expresión consistentes entre las dos condiciones, mientras que otros cúmulos reflejaron variaciones significativas. En particular, los genes del cúmulo con un valor de k=4 mostraron una diferencia notable en la expresión entre control y tratamiento, sugiriendo un posible efecto del tratamiento sobre estos genes.

Al agrupar los datos, se encontró un cúmulo que presentaba un total de 244 genes con un comportamiento especialmente interesante. Este grupo más pequeño, visualizado en un simplex de probabilidad, se destacó por la notable variabilidad en la expresión génica entre ambas condiciones, lo que sugiere una respuesta anómala al tratamiento con vitamina D.

Relaciones encontradas	**Base de datos**	**Genes encontrados con K-medias (244)**
Factores de transcripción	FT de consenso ENCODE y ChEA de Chip-X, Perturbaciones de FT seguidas de expresión	• AR 21572438 ChIP-Seq LNCaP • TP53 20018659 ChIP-ChIP R1E • VDR ChEA • SMAD4 CHEA
Rutas metabólicas	WikiPathway2023 Humano, BioPlanet 2019, MSigDB Hallmark 2020	• Ruta Metabólica de la vitamina DWP2877 • Receptores nucleares en el metabolismo lipídico y su toxicidad WP299 • Regulación de NFAT factores de trascripción • p53 Ruta metabólica
Ontología	GO Procesos biologicos 2023, GO Funciones moleculares 2023, GO Componentes celulares 2023	• Regulación negativa de la fase de ejecución de (GO:1900118) • Proceso metabolico de la vitamina D (GO:0042359) • Complejo de cadena respiratoria III
Enfermedades	Genes relacionados con COVID-19 2021, Catalogo GWAS 2023, DisGeNET	• Infección por SARS-CoV-2 en el pulmón humano según GSE150316 • Artritis • Raquitismo resistente a la vitamina D

Figura 6.1: Relaciones encontradas con la vitamina D para los 244 genes detectados por k-medias,

Se realizó un análisis utilizando la plataforma Enrichr, una herramienta bioinformática que permite realizar análisis a partir de una lista de genes dada [75, 76, 77]. Enrichr ofrece información detallada sobre rutas metabólicas, enfermedades asociadas, factores de transcripción, tipos celulares y otros aspectos biológicos relevantes. Al analizar los genes anómalos identificados, se reveló que varios de ellos estaban involucrados en vías de señalización y factores de transcripción estrechamente relacionados con la vitamina D. Por ejemplo, en los conjuntos de datos FT de consenso ENCODE y ChEA de Chip-X (ENCODE and ChEA Consensus TFs from Chip-X), Perturbaciones de FT seguidas de expresión (TF Perturbations Followed by Expression), y Factores de Transcripción TRRUST 2019 (TRRUST Transcription Factors 2019), se encontraron conexiones con el

receptor de vitamina D (VDR), lo que sugiere una implicación directa de estos genes en la respuesta biológica a la vitamina D de acuerdo con su ontología.

En el análisis de vías de señalización, se identificó la vía metabólica del receptor de la vitamina D (Vitamin D Receptor Pathway) en los resultados de WikiPathway 2024 Humano (WikiPathway 2023 Human). Además, se encontraron relaciones con otras vías metabólicas conocidas por su relación con la regulación de procesos inmunológicos y metabólicos influenciados por la vitamina D, como pueden ser: la inducción de apoptosis y la regulación de factores de transcripción. Esto respalda la hipótesis de que el tratamiento influye de manera significativa en la expresión de genes clave en la respuesta a la vitamina D.

Continuando con el análisis, además de la conexión directa con el factor de transcripción VDR (Ver Fig. 6.1), podemos encontrar el receptor de andrógenos (AR) que posee una interacción indirecta con VDR, ya que ambos son receptores nucleares que regulan procesos relacionados con el metabolismo, la unión de proteínas y la diferenciación celular [78]. Al igual que el gen VDR, el gen TP53 un gen clave en la regulación del ciclo celular y la apoptosis por lo que se ha demostrado que responde con al menos dos regiones promotoras de unión a VDR [79]. De igual forma podemos encontrar el gen SMAD4 que juega un papel en la regulación de la diferenciación celular, especialmente en el contexto del cáncer y la fibrosis [80].

El algoritmo LOF fue utilizado para analizar las distancias Wasserstein obtenidas, con el objetivo de identificar genes con un comportamiento atípico en sus perfiles de expresión. Este análisis permitió detectar 83 genes cuya distancia entre las condiciones de tratamiento y control fue anormalmente alta en comparación con otros genes, sugiriendo una respuesta atípica al tratamiento con vitamina D.

Se calcularon los valores de LOF para cada gen, considerando tanto las condiciones de control como las de tratamiento. Un umbral de LOF mayor a 1.5 se definió para clasificar un gen como *outlier*. Este valor es comúnmente utilizado como un criterio estándar para detectar anomalías en datos multidimensionales, ya que indica una densidad significativamente menor en comparación con sus puntos vecinos. Los genes identificados

como *outliers* mostraron distancias significativamente diferentes entre las condiciones de control y tratamiento, lo que sugiere una variación inusual en la expresión génica que podría estar relacionada con los efectos de la vitamina D.

Los genes clasificados como *outliers* fueron analizados para identificar posibles funciones o vías biológicas relevantes en las que están involucrados. En particular, se observaron patrones interesantes: muchos de estos genes están asociados con vías metabólicas específicas o están relacionados con la respuesta inmune, lo que refuerza la idea de que el tratamiento con vitamina D podría estar afectando procesos metabólicos relacionados.

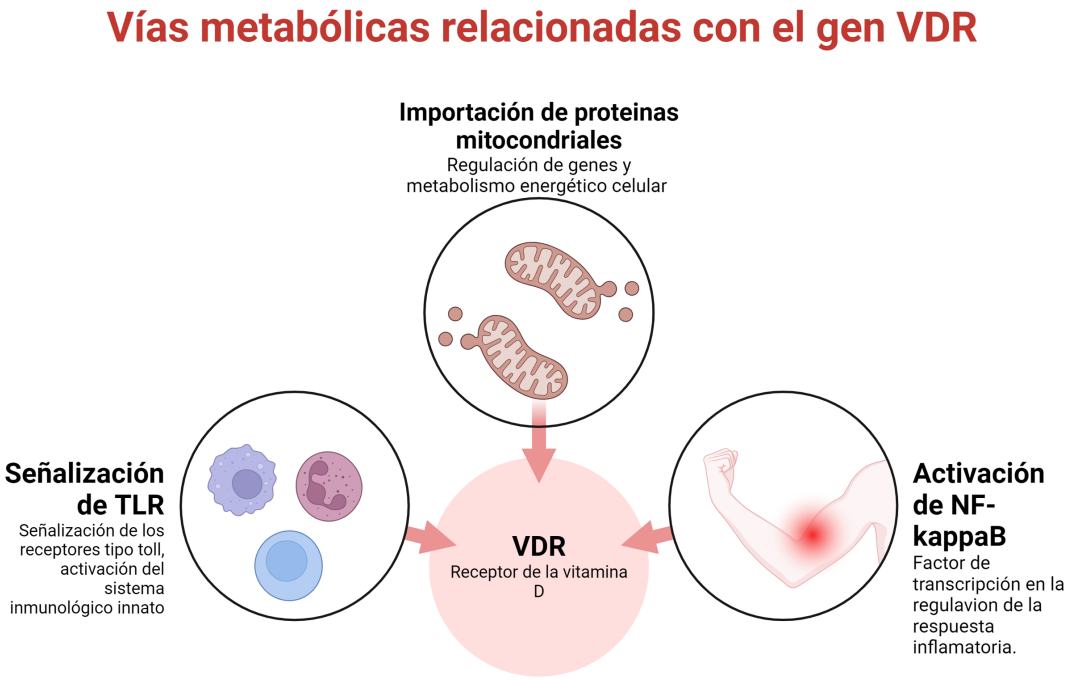


Figura 6.2: Ejemplo de relaciones entre el receptor de vitamina D (VDR) y las vías metabólicas importantes como, activación de la respuesta inflamatoria e importación de proteínas mitocondriales. Estas conexiones destacan la influencia del VDR en diversas funciones y procesos biológicos relacionados con la vitamina D.

Los 83 genes fueron sometidos al mismo análisis de enriquecimiento funcional utilizando la plataforma antes mencionada. Este análisis permitió identificar que algunos de los genes están asociados con factores de transcripción importantes, como VDR y CDH1 (Proteína de adhesión dependiente de calcio). Estos factores desempeñan roles clave en la regulación de la expresión génica en respuesta a la vitamina D, así como en el desarrollo de las conexiones existentes entre estos factores de transcripción para la proliferación de tumores [81].

Relaciones encontradas	**Base de datos**	**Genes encontrados con LOF (83)**
Factores de transcripción	ChEA, FT de consenso ENCODE y ChEA de Chip-X, Factores de transcripción TRRUST	• VDR 33458620 ChIP-Seq Epitelial primario de próstata humana • VDR 23849224 ChIP-Seq CD4+ Humano • AR CHEA, CDH1, FOS humano
Rutas metabólicas	Reactome 2022, WikiPathways 2024 Human, BioPlanet	• Importación de proteínas mitocondriales • Ruta de resultados adversos de COVID-19 WP4891 • Función no clásica de la vitamina D WP5133
Ontología	GO Proceso Biologico 2023, GO Funcion Molecular 2023, Fenotipo de mamífero MGI nivel 4 2024	• Señalización integrada de respuesta al estrés (GO:0140467) • Unión al ARN (GO:0003723) • Disminución del número de células monocíticas MP:0000223
Enfermedades	Conjuntos de genes relacionados con la COVID-19 2021, Orphanet ampliado 2021	• Genes regulados a la baja por el SARS-CoV-2 en células humanas hiPSC-CMs a las 24 h de GSE150392 • Síndrome de melanoma y tumor del sistema nervioso ORPHA:252206

Figura 6.3: Relaciones encontradas con la vitamina D para los genes 83 detectados por LOF

También se detectaron genes implicados en la respuesta inflamatoria y vías relacionadas con enfermedades metabólicas y autoinmunes, además de genes relacionados con la importación de proteínas mitocondriales y la respuesta al estrés celular (Ver Fig. 6.3). Además, se puede señalar directamente dentro de las rutas metabólicas a la función no clásica de la vitamina D.

Recordando que la vitamina D forma parte de los receptores nucleares que interactúan con secuencias específicas del ADN, induciendo represión o

activación de la transcripción. Se puede dividir las funciones no clásicas en 3 tipos: formando parte de la secreción hormonal, regulación de la respuesta inmune y la proliferación y diferenciación celular [82].

Es importante destacar que también, se identificaron asociaciones con conjuntos de genes relacionados con COVID-19 y otros síntomas del sistema nervioso, lo que sugiere que la vitamina D está involucrada en la respuesta inmune durante infecciones virales además de encontrarse asociaciones con condiciones como tumores y cáncer de piel, lo cual podría ser indicativo de un papel protector o modulador de la vitamina D en estas patologías.

Al comparar los resultados obtenidos con LOF y K-medias, se encontraron coincidencias entre algunos genes detectados como *outliers* por LOF y ciertos cúmulos identificados por K-medias, lo que sugiere que la detección de estos genes con variabilidad en la expresión génica refuerza su carácter anómalo. Estos hallazgos destacan la utilidad de ambos algoritmos para identificar genes potencialmente relevantes en estudios de expresión génica, especialmente en este contexto de tratamientos con vitamina D.

Aunque no se pretende afirmar que LOF y K-medias sean superiores a los métodos estadísticos tradicionales, estos algoritmos ofrecen una alternativa valiosa para explorar patrones complejos en datos genómicos. La aplicación de estos métodos han permitido identificar genes que podrían pasar desapercibidos con enfoques convencionales, subrayando su importancia en la investigación de los efectos de la vitamina D.

Además de los métodos K-medias y LOF, se utilizó una prueba *T-student* para detectar genes anómalos, lo cual es un enfoque estadístico común en estudios de expresión génica. Este método también permite identificar genes con diferencias estadísticamente significativas, a través de esta técnica, se detectaron aproximadamente 900 genes como anómalos.

La prueba T de *Student* es un método estadístico muy común donde se obtiene la media de dos grupos y al realizar la comparación se puede identificar si las diferencias encontradas son significativas o no [83]. En el contexto del análisis de expresión génica, se utiliza para verificar si la expresión de un gen es diferente entre dos condiciones, como en este caso

entre el grupo de control y un grupo de tratamiento.

Al estar basada en estadística, la prueba toma la diferencia que existe entre las medias de los grupos y divide esa diferencia por la variabilidad de los datos dentro de cada grupo. La variabilidad está representada por la desviación estándar, que mide qué tan alejados están los valores de la media [84]. Al hacer este cálculo, se obtiene un valor t, que nos indica qué tan grande es la diferencia entre los grupos. Si el valor t es alto, significa que la diferencia entre los grupos es significativa.

Sin embargo, al comparar los genes identificados por las tres técnicas K-medias, LOF y *T-student*, se encontró que solo tres genes fueron consistentes en estas metodologías. Esta coincidencia sugiere que estos tres genes representan los casos más robustos de anomalía en la expresión génica bajo las condiciones analizadas, ya que fueron detectados de manera consistente por métodos basados en diferentes principios matemáticos y estadísticos (Figura 6.4). La identificación de un número reducido de genes comunes refuerza la necesidad de utilizar múltiples enfoques para detectar anomalías con mayor confianza.

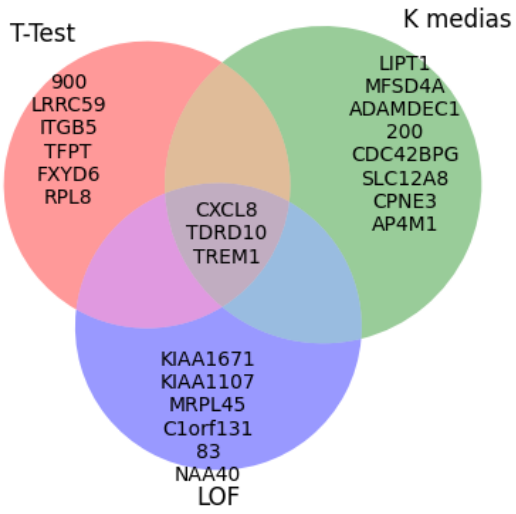


Figura 6.4: Diagrama de Venn que refleja la cantidad de genes anómalamente expresados en ambas condiciones mediante los tres métodos distintos, en el centro se encuentran los genes que comparten estas técnicas. Los genes compartidos están relacionados con la regulación de la expresión génica .

Además de las técnicas previamente mencionadas, se llevaron a cabo dos análisis adicionales utilizando las implementaciones de pruebas estadísticas DESeq2 y edgeR, ambas ampliamente utilizados para la identificación de genes con expresión diferencial en estudios de RNA-seq.

DESeq2 y edgeR son las herramientas bioinformáticas más utilizadas en la actualidad para analizar datos de expresión génica diferencial en estudios de RNA-seq, estas herramientas logran identificar genes con cambios en su expresión bajo diferentes condiciones, como es este contexto, entre tratamientos y controles.

Ambas herramientas se utilizan en el contexto donde se necesita comparar las diferencias en la expresión génica entre grupos, y son fundamentales porque permiten identificar genes clave que podrían estar asociados a procesos biológicos importantes en respuesta a algún tratamiento dado.

El análisis de expresión diferencial realizado utilizando edgeR identificó un total de 1223 genes con diferencias significativas entre las condiciones de tratamiento y control. Este método, basado en la distribución binomial negativa, es particularmente eficaz para manejar variabilidad biológica en experimentos de RNA-seq, permitiendo una detección precisa de genes con alteraciones en su expresión [57]. Esta técnica tiene una gran capacidad para manejar los datos de RNA-seq de manera flexible y robusta ya que se adapta a las características estadísticas específicas de los datos.

Tras la identificación de estos genes, se llevó a cabo un análisis de enriquecimiento funcional utilizando nuevamente la plataforma Enrichr. Este análisis permitió no solo identificar los genes diferenciales, sino también explorar las funciones biológicas, rutas metabólicas y factores de transcripción asociados a estos genes. Enrichr, a través de su vasta base de datos de anotaciones, proporciona una visión integral de los posibles mecanismos involucrados.

Los resultados obtenidos reflejan que los factores de transcripción mas destacados son el gen VDR, receptor de la vitamina D, GATA1 y Runx1, los cuales están involucrados en la regulación de procesos inmunitarios o respuestas inflamatorias y en procesos relacionados con la formación y desarrollo de células cancerígenas en el cuerpo. La implicación de estos

factores sugiere que el tratamiento podría estar modulando vías biológicas relacionadas a estos factores de transcripción [85].

Relaciones encontradas	Base de datos	Genes encontrados con edgeR (1123)
Factores de transcripción	ChEA, FT de consenso ENCODE y ChEA de Chip-X	• VDR, GATA1, RUNX1 CHEA • SP1 humano • VDR 24787735 ChIP-Seq THP-1 Humano
Rutas metabólicas	Reactome 2022, WikiPathways 2024 Human, BioPlanet	• Transporte de electrones respiratorios RHSA-611105 • Vía del receptor de vitamina D WP2877 • Regulación de la apoptosis por FSH, Coagulación
Ontología	GO Proceso Biologico 2023, GO Funcion Molecular 2023, GO Componentes Celulares 2023	• Respiración celular (GO:0045333) • Regulación negativa de la vía de señalización apoptótica (GO:2001234) • Complejo IV de la cadena respiratoria (GO:00452777)
Enfermedades	Conjuntos de genes relacionados con la COVID-19 2021, ClinVAR 2019	• Pacientes con COVID-19 en PBMC en aumento • Trombocitopenia inmunitaria ORPHA:3002

Figura 6.5: Relaciones encontradas con la vitamina D para los 1123 genes detectados por edgeR.

Además de los factores de transcripción, el análisis en Enrichr destacó la participación de los genes diferenciales en otras vías biológicas, se identificaron alteraciones significativas en las rutas implicadas en la respuesta de la vitamina D y perturbaciones genéticas asociadas con los genes involucrados en la regulación del ciclo celular sugiere que el tratamiento podría estar afectando la proliferación celular (ver Fig. 6.5).

Por otro lado el análisis de expresión diferencial utilizando DESeq2 reveló un total de 79 genes con diferencias significativas en la expresión entre las condiciones de tratamiento y control. DESeq2 también utiliza su base en la distribución binomial negativa para modelar los conteos de RNA-seq, lo que le permite un mejor manejo de la variabilidad y dispersión de los datos de manera efectiva, particularmente en experimentos con un número reducido de réplicas [56].

Los 79 genes encontrados fueron sometidos igualmente al análisis dentro de las plataformas de Enrichr para comprender mejor sus funciones biológicas, al igual que sus rutas metabólicas y factores de transcripción asociados. Dentro de los factores de transcripción claves asociados a los genes detectados por DESeq2, sugiere que el tratamiento podría estar afectando las vías biológicas que regulan procesos relacionados con la diferenciación celular y la respuesta inmune. Además dentro del análisis se mostró una asociación con genes implicados en la regulación de respuestas inflamatorias.

Relaciones encontradas	**Base de datos**	**Genes encontrados con Deseq2 (79)**
Factores de transcripción	ChEA, FT de consenso ENCODE y ChEA de Chip-X	• VDR, GATA1 CHEA • SP1 humano • VDR 24787735 ChIP-Seq THP-1 Humano
Rutas metabólicas	Reactome 2022, WikiPathways 2024 Human, BioPlanet	• Desgranulación de neutrófilos R-HSA-6798695 • Sistema inmunológico R-HSA-168256 • Vía del receptor de vitamina D WP2877
Ontología	GO Proceso Biologico 2023, GO Funcion Molecular 2023, GO Componentes Celulares 2023	• Activación de neutrófilos (GO:0042119) • Regulación de la vía de señalización apoptótica extrínseca (GO:2001236) • Matriz extracelular que contiene colágeno (GO:0062023)
Enfermedades	Conjuntos de genes relacionados con la COVID-19 2021, ClinVAR 2019	• Genes más importantes para la enfermedad leve por COVID-19 en células NK humanas de GSE165461 • Osteólisis carpo-tarsiana multicéntrica con o sin nefropatía ORPHA:2774 • Enfermedad granulomatosa crónica

Figura 6.6: Relaciones encontradas con la vitamina D para los 79 genes detectados por Deseq2.

El análisis de las rutas biológicas también proporcionó información sobre las posibles funciones de los genes identificados como la desgranulación de neutrófilos, proceso crucial en la respuesta inmune innata que permite la defensa contra infecciones bacterianas. La implicación de estas rutas sugiere que el tratamiento podría estar afectando la capacidad de respuesta del organismo ante estímulos inflamatorios. Además, se ven involucradas rutas que intervienen en la apoptosis celular y la respuesta a estrés oxidativo lo que podría indicar alteraciones en los procesos relacionados con la respuesta a daño celular.

Al comparar los resultados obtenidos con DESeq2 y edgeR, se observan ciertas diferencias y similitudes en los genes identificados y las vías involucradas. Mientras que edgeR detectó un mayor número de genes, DESeq2 proporcionó una lista más restringida, pero con una mayor especificidad en ciertas rutas biológicas clave. Estos métodos analizados permiten obtener múltiples enfoques, señalando una visión más completa de los cambios en la expresión génica bajo diferentes condiciones experimentales. La última

imagen presentada (ver Fig. 6.7), muestra una representación de intersecciones de genes identificados como anómalos por los cinco métodos utilizados en este estudio: K-medias, LOF, T-student, edgeR y DESeq2. En la parte inferior izquierda se muestran los tamaños totales de los conjuntos de genes detectados por cada método: edgeR identificó 1223 genes, T-student 737, K-means 244, LOF 83 y DESeq2 79.

En la parte inferior derecha, las combinaciones de métodos están representadas por puntos conectados por líneas. Cada combinación indica cuántos genes fueron identificados exclusivamente por esos métodos. Finalmente, en la parte superior, las barras negras muestran el tamaño de la intersección para cada combinación de métodos, proporcionando una visión clara del número de genes únicos o compartidos entre los métodos, estos genes no son encontrados entre los otros conjuntos.

Existen genes que fueron identificados por K-medias y LOF en comparación con los otros métodos, compartiendo entre todos un total de 2 genes. Los genes compartidos entre los 5 métodos son los que tienen mayor carácter anómalo en ambas condiciones. Este análisis destaca las diferencias y similitudes entre los enfoques estadísticos (como DESeq2 y edgeR) y los geométricos (como LOF y K-means).

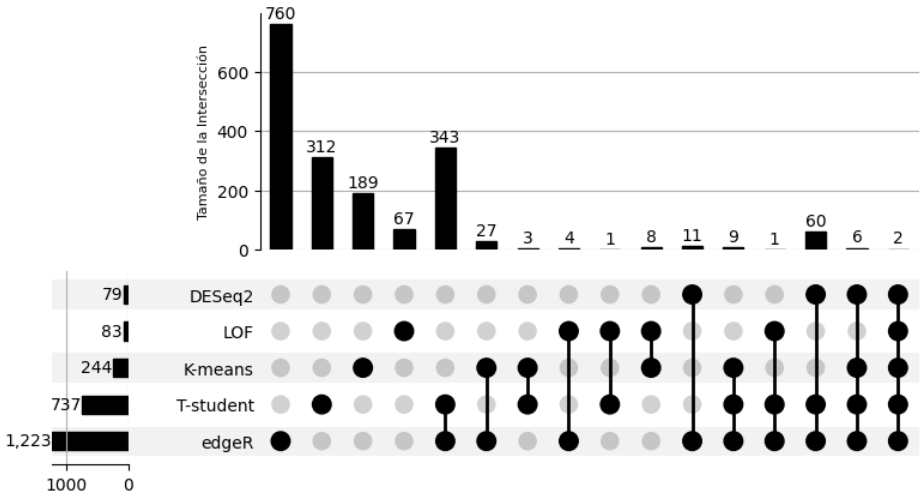


Figura 6.7: Representación de las intersecciones de genes detectados por cinco métodos: LOF, K-medias, T-student, edgeR y DESeq2. Las barras superiores indican el tamaño de cada intersección, mientras que las líneas y puntos en la matriz inferior representan las combinaciones de métodos que detectaron los mismos genes. Por ejemplo, 760 genes fueron identificados exclusivamente por edgeR y DESeq2, y 2 genes fueron detectados por los cinco métodos.

En esta tesis se emplearon los algoritmos K-medias y LOF, junto con pruebas estadísticas como *T-student*, edgeR y DESeq2, para identificar genes anómalos en respuesta al tratamiento con vitamina D en células PBMC. K-medias permitió agrupar los genes según sus perfiles de expresión utilizando como base las distancias proyectadas, destacando un total de 244 genes con un comportamiento diferente entre las condiciones de control y tratamiento. El algoritmo LOF, basado igualmente en la distancia de Wasserstein, identificó 83 genes como *outliers*. Los genes identificados fueron analizados en plataformas como Enrichr, revelando asociaciones con factores de transcripción como el receptor de vitamina D (VDR) y vías metabólicas relacionadas con la respuesta inmune y la inflamación.

Aunque se encontraron coincidencias entre los métodos utilizados, inicialmente solo tres genes fueron comunes en K-medias, LOF y *T-student*, sugiriendo que estos genes son los que presentan en términos de anomalía una expresión diferencial más representativa. Adicionalmente, edgeR identificó 1223 genes, mientras que DESeq2 detectó 79 genes con diferencias significativas. Los resultados del análisis funcional sugieren que la vitamina D influye en diversas vías biológicas, incluidas la señalización inmune y el daño celular.

Si bien edgeR y DESeq2 se basan en métodos estadísticos más tradicionales y ampliamente utilizados en estudios de RNA-seq, K-medias y LOF ofrecen una perspectiva diferente al abordar la identificación de genes anómalos desde un enfoque geométrico y basado en distancias, lo que justifica la relevancia de su aplicación en este análisis. Aunque los métodos tradicionales identificaron un mayor número de genes, los resultados sugieren que la utilización de estos nuevos enfoques es relevante para una comprensión más profunda de la expresión génica en el contexto del tratamiento con vitamina D.

Capítulo 7

Conclusión

Este estudio ha permitido identificar genes con patrones de expresión anómalos en respuesta a tratamientos relacionados con la vitamina D, utilizando enfoques como K-medias y LOF, complementando con comparativas con pruebas estadísticas tradicionales como la *T-student*. Los resultados obtenidos proporcionan una visión clara de cómo estos algoritmos, a pesar de no ser superiores a los métodos estadísticos tradicionales, ofrecen una alternativa efectiva para detectar genes que podrían pasar desapercibidos con métodos convencionales. Además, el análisis realizado a través de la plataforma Enrichr ha revelado vías metabólicas y factores de transcripción que se encuentran relacionados con la vitamina D, permitiendo conocer la relevancia de estos hallazgos para futuras investigaciones.

En pasos futuros, este trabajo puede ampliarse de varias maneras. Un posible siguiente paso sería la extensión del análisis a series de tiempo, lo que permitiría observar cómo varía la expresión génica en diferentes momentos tras la administración del tratamiento. Esta aproximación ayudaría a identificar genes que no solo tienen una respuesta anómala, sino que además presentan variaciones específicas a lo largo del tiempo, proporcionando una visión más profunda del efecto del tratamiento utilizado en relación con la vitamina D.

Otra vía interesante sería la creación de una librería en python que integre los métodos de detección de genes anómalos utilizados en este estudio, incluyendo todo el análisis existente y herramientas utilizadas. Una librería de código abierto no solo facilitaría la reproducibilidad de los resultados, sino que también pondría a disposición de la comunidad científica

herramientas eficientes para el análisis de expresión génica. Al incluir algoritmos como K-medias y LOF en lugar de otros enfoques estadísticos tradicionales, se podría crear una herramienta flexible que permita a los investigadores adaptar sus análisis según las características de sus datos.

Aparte de estas ideas, una tercera posible extensión sería combinar estas técnicas con enfoques más avanzados de aprendizaje automático o aprendizaje profundo, para implementar mejoras en la identificación de genes anómalos a gran escala. El uso de redes neuronales, por ejemplo, podría mejorar la detección de patrones complejos en datos genómicos, y la combinación de estos modelos con algoritmos no supervisados como K-medias y LOF podría generar resultados aún más robustos.

Esta tesis proporciona una sólida base para futuras investigaciones. Las posibles extensiones propuestas, ya sea hacia series de tiempo, el desarrollo de herramientas en python, o el uso de técnicas más avanzadas de aprendizaje de máquina, tienen el potencial de aportar y ampliar aún más el conocimiento en el campo de la biología computacional y la expresión génica.

Bibliografía

[1] Carsten Carlberg, Marianna Raczyk y Natalia Zawrotna. "Vitamin D: A master example of nutrigenomics". En: *Redox Biology* (2023).

[2] Zuluaga Espinosa. "Vitamina D: nuevos paradigmas." En: *Medicina y Laboratorio. Vol. 17* (2011).

[3] Holick MF. "Vitamin D: evolutionary, physiological and health perspectives". En: *Curr Drug Targets* (2011).

[4] María Calatayud. "Prevalencia de concentraciones deficientes e insuficientes de vitamina D en una población joven y sana". En: *Endocrinología y Nutrición* (2009).

[5] Lips P. "Vitamin D deficiency and secondary hyperparathyroidism in the elderly: consequences for bone loss and fractures and therapeutic implications." En: *Endocr Rev.* (2001).

[6] Wang Z., Gerstein M y Snyder M. "RNA-Seq: a revolutionary tool for transcriptomics." En: *Nat Rev Genet 10* (2009).

[7] Lus M. Muñiz-Rivera. "Análisis sobre métodos de pruebas de hipótesis múltiple en la identificación de genes diferencialmente expresados". Tesis doct. College of Arts y Sciences - Sciences, 2009.

[8] Madrigal P Conesa A. "A survey of best practices for RNA-seq data analysis. Genome Biol 17". En: *Genome Biol 17* (2016).

[9] Amit Thakkar Durgesh Samariya. "comprehensive survey of anomaly detection algorithms." En: *Annals of Data Science* (2023).

[10] S. Mota y A. Neme N. Hevia. "LAS ANOMALÍAS:¿ QUÉ SON?,¿ DÓNDE SURGEN?,¿ CÓMO DETECTARLAS?." En: *Tecnología e Innovación en Educación Superior.* (2021).

[11] Ayman Taha y Hadi. "Detection Methods for Categorical Data: A Review." En: *ACM Computing Surveys.* (2019).

[12] Markus M. Breunig et al. "LOF: identifying density-based local outliers". En: *Association for Computing Machinery* (2000).

[13] Islam Ahmed M. Seraj R. "The k-means algorithm: A comprehensive survey and performance evaluation." En: *Electronics, 9(8)* (2020).

[14] A. J. Griffiths. *An Introduction to Genetic Analysis.* WH Freeman y Company., 2005.

[15] Melina Claussnitzer et al. "A brief history of human disease genetics". En: *Nature* (2020).

[16] Barzilai N. Martin G. M. Bergman A. "Genetic determinants of human health span and life span: progress and new opportunities." En: *PLoS genetics* (2007).

[17] Trejo María J. Gil, Alejandra Laureano Viveros y Sofía González Salinas. "Historia y estructura del adn." En: *Boletín Científico de la Escuela Superior Tepeji del Río 5.10* (2018).

[18] Pierce B. A. *Genetics: A Conceptual Approach (6th ed.)* Macmillan., 2016.

[19] MedlinePlus Genetics. *What is a gene?* 2024. URL: `https://medlineplus.gov/genetics/understanding/basics/gene/`.

[20] Y Horacio Merchant. Jiménez Luis Felipe. *Biología celular y molecular.* México: Pearson educación, 2003.

[21] Cramer Patrick. "Organization and regulation of gene transcription." En: *Nature 573* (2019).

[22] William Xavier Cascante Mosquera. "Transcripción ADN". En: *Biología - FIMCBOR* (2009).

[23] Becker Wayne M. *The world of the cell.* Pearson New York, 2005.

[24] Strachan T. y Read A. P. *Human Molecular Genetics (5th ed.)* Garland Science., 2018.

[25] Young Vernon R. *Proteínas y aminoácidos.* Conocimientos actuales de nutrición, 2003.

[26] Harvey F Lodish. *Molecular cell biology.* Macmillan, 2008.

[27] Kouzarides T. "Chromatin modifications and their function." En: *Cell* (2007).

[28] Castrillo José Luis. "Factores de transcripción específicos de tejido." En: *Investig Ciencia* (1995).

[29] Holick MF. Wacker M. "Sunlight and Vitamin D: A global perspective for health". En: *Dermatoendocrinol* (2013).

[30] Ph.D. Michael F. Holick M.D. "Vitamin D Deficiency". En: *The New England Journal of Medicine* (2007).

[31] Dahiara. Vanegas Losada. *Fisiopatología ósea: papel de la vitamina D en salud y enfermedad.* Revisión de la literatura. Universidad Nacional de Colombia. 2017.

[32] Carsten Carlberg. "Vitamin D and Its Target Genes". En: *Nutrients* 14 (2022).

[33] Bikle DD. "Vitamin D metabolism, mechanism of action, and clinical applications". En: *Chem Biol* (2014).

[34] Prashant Sakharkar, Subrata Deb y Don Vu. "Vitamin D receptor (VDR) gene polymorphism: implications on non-bone diseases". En: *J Basic Clin Pharm* (2017).

[35] A. W Norman. "From vitamin D to hormone D: Fundamentals of the vitamin D endocrine system essential for good health." En: *The American Journal of Clinical Nutrition* (2008).

[36] Sirajudeen S., Shah I. y Al Menhali A. "A Narrative Role of Vitamin D and Its Receptor: With Current Evidence on the Gastric Tissues." En: *International journal of molecular sciences, 20(15)* (2019).

[37] Pike J. W. et al. "The vitamin D receptor: contemporary genomic approaches reveal new basic and translational insights." En: *The Journal of clinical investigation* (2017).

[38] Lewis Benjamin. *Genes IX*. MacGrawHill, 2008.

[39] Latchman D. S. "Transcription factors: an overview". En: *The international journal of biochemistry and cell biology* (1997).

[40] Alon Uri. "Network motifs: theory and experimental approaches." En: *Nature reviews. Genetics, 8(6)* (2007).

[41] Vaux David. L., Fidler F. y Cumming G. "Replicates and repeats–what is the difference and is it significant? A brief discussion of statistics and experimental design." En: *EMBO reports* (2012).

[42] Motulsky Harvey. *Intuitive biostatistics: a nonmathematical guide to statistical thinking.* Oxford University Press, 2014.

[43] Halsey L., Curran-Everett D. y Vowler S. "The fickle P value generates irreproducible results". En: *Nat Methods* (2015).

[44] Jacqueline Michelle Jara Moscoso y Kerly Samantha Saquipay Nieves. "Evaluación y comparación de técnicas bioinformáticas para el análisis de expresión diferencial en NGS". Tesis doct. Universidad Politécnica Salesiana, 2023.

[45] Bustin S A et al. "The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments." En: *Clin Chem.* (2009).

[46] Milos PM. Ozsolak Fatih. "RNA sequencing: advances, challenges and opportunities." En: *Nat Rev Genet.* (2011).

[47] Peñaloza Oscar M. R. y Portugal Patricia M. "Análisis bioinformático de Arn-Seq con una perspectiva para Bolivia." En: *Revista Boliviana de Química* (2017).

[48] Verónica Jiménez Jacinto. "Manejo de Datos de NGS". En: *Unidad Universitaria de Secuenciación Masiva de ADN, UNAM.* 2014.

[49] Dobin A et al. "STAR: ultrafast universal RNA-seq aligner." En: *Bioinformatics.* (2013).

[50] Sedano Johana C. S. y Carrascal C. E. L. "RNA-seq: herramienta transcriptómica útil para el estudio de interacciones planta-patógeno." En: *Fitosanidad* (2012).

[51] Dewey C.N. Li Bo. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." En: *BMC Bioinformatics 12,323* (2011).

[52] Mortazavi Ali et al. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." En: *Nature methods, 5(7), 621-628.* (2008).

[53] Trapnell Cole et al. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." En: *Nature biotechnology, 28(5)* (2010).

[54] Wagner GP, Kin K y Lynch VJ. "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples." En: *Theory Biosci.* (2012).

[55] Koch C. M.and Chiu S. F. et al. "A Beginner's Guide to Analysis of RNA Sequencing Data." En: *American journal of respiratory cell and molecular biology, 59(2)* (2018).

[56] Love M.I.and Huber W. y Anders S. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." En: *Genome Biol 15, 550* (2014).

[57] Robinson MD, McCarthy DJ y Smyth GK. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." En: *Bioinformatics.* (2010).

[58] L Simon y IOB ECTIVE. "An introduction to the negative binomial distribution and its applications". En: *Proceedings of the Casualty Actuarial Society.* Vol. 49. 91 Part 1. 1962.

[59] Navarro A et al. "La distribución binomial negativa frente a la de Poisson en el análisis de fenómenos recurrentes". En: *Gaceta Sanitaria* 15.5 (2001), págs. 447-452.

[60] Anders S.and Huber W. "Differential expression analysis for sequence count data." En: *Genome Biol 11* (2010).

[61] A.K. Jain. "Data Clustering: 50 Years Beyond K-means". En: *Machine Learning and Knowledge Discovery in Databases.* Ed. por W. Daelemans, B. Goethals y K. Morik. Vol. 5211. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2008.

[62] J. MacQueen. "Some methods for classification and analysis of multivariate observations". En: *FIFTH BERKELEY SYMPOSIUM.* 1967.

[63] Igor Melnykov y Volodymyr Melnykov. "On K-means algorithm with the use of Mahalanobis distances". En: *Statistics and Probability Letters* 84 (2014), págs. 88-95. ISSN: 0167-7152. DOI: `https://doi.org/10.1016/j.spl.2013.09.026`.

[64] Wei Wang, Ji Zhang y Hai H. Wang. "Grid-ODF: Detecting Outliers Effectively and Efficiently in Large Multi-dimensional Databases". En: *International Conference on Computational Intelligence and Security.* 2005.

[65] Jin Wen, Tung Anthony K. H. y Han Jiawei. "Mining top-n local outliers in large databases". En: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Association for Computing Machinery, 2001. ISBN: 158113391X. DOI: `10.1145/502512.502554`.

[66] Krishna Gopal Sharma, Yashpal Singh y Atul Kumar Srivastava. "Variance on outlier factor". En: *International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)* (2017).

[67] Wen Jin, Anthony Kum Hoe Tung y Jiawei Han. "Mining top-n local outliers in large databases". En: *Knowledge Discovery and Data Mining*. 2001.

[68] Hanel A et al. "Common and personal target genes of the micronutrient vitamin D in primary immune cells from human peripheral blood." En: *Scientific Reports* ().

[69] Doerge RW. Auer PL. "Statistical design and analysis of RNA sequencing data." En: *Genetics* (2010).

[70] Sung-Hyuk Cha. "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions". En: *INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES* (2007).

[71] Larry Wasserman. *Optimal Transport and Wasserstein Distance*. 36-708 Statistical Methods for Machine Learning. 2019.

[72] Victor M. Panaretos y Yoav Zeme. "Statistical Aspects of Wasserstein Distances". En: *Annual Review of Statistics and Its Application* (2018).

[73] Muka Nakazato y Sosuke Ito. "Geometrical aspects of entropy production in stochastic thermodynamics based on Wasserstein distance". En: *Phys. Rev. Res.* (2021).

[74] Zaragoza Galiana y Aitor. "Clustering y Analítica de clientes de SEMIC mediante Machine Learning". Tesis doct. Universitat de Lleida, 2021.

[75] Evan Y Chen et al. "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool". En: *BMC Bioinformatics* 14.1 (2013), pág. 128.

[76] Maxim V Kuleshov et al. "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update". En: *Nucleic Acids Research* 44 (2016), W90-W97. DOI: gkw377.

[77] Zheng Xie et al. "Gene set knowledge discovery with Enrichr". En: *Current Protocols* 1.3 (2021), e90. DOI: 10.1002/cpz1.90.

[78] Ting HJ et al. "Androgen-receptor coregulators mediate the suppressive effect of androgen signals on vitamin D receptor activity." En: *Endocrine.* (2005).

[79] Saramäki A et al. “Regulation of the human p21(waf1/cip1) gene promoter via multiple binding sites for p53 and the vitamin D3 receptor.” En: *Nucleic Acids Res.* (2006).

[80] Yu RT Ding N et al. “A vitamin D receptor/SMAD genomic circuit gates hepatic fibrotic response.” En: *Cell* (2003).

[81] Peña C et al. “E-cadherin and vitamin D receptor regulation by SNAIL and ZEB1 in colon cancer: clinicopathological correlations.” En: *Hum Mol Genet* (2005).

[82] Calle Pascual, Alfonso L. y María J. “La vitamina D y sus efectos "no clásicos"”. En: *Revista Española de Salud Pública* 86 (2012).

[83] Manuel Molina. “Paso a paso. Prueba de la t de Student para muestras independientes.” En: *Revista Electrónica AnestesiaR* 14 (2022).

[84] Sánchez Turcios y Reinaldo Alberto. “t-Student: Usos y abusos”. En: *Revista mexicana de cardiología* 26 (2015).

[85] Göbel F et al. “Reciprocal role of GATA-1 and vitamin D receptor in human myeloid dendritic cell differentiation”. En: *Blood.* (2009).