Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Departamento de Química Biológica

Uso de la inteligencia artificial para la clasificación de transcriptomas de sangre entera

Tesis presentada para optar al título de Doctora de la Universidad de Buenos Aires en el área de Ciencias Biológicas

MARÍA NELBA PÉREZ

Director: SANTIAGO MIRIUKA

Consejero de estudio: IGNACIO SÁNCHEZ

Lugar de trabajo: Instituto de Neurociencias-CONICET, FLENI

Buenos Aires, 31 de marzo de 2025

Resumen

El desarrollo de redes neuronales ha transformado nuestras actividades diarias, especialmente en biomedicina, donde ha mejorado la comprensión de enfermedades y el diagnóstico. Recientemente, se han logrado avances en el estudio de los ARNs y las técnicas de secuenciación masiva. El transcriptoma, que incluye todos los ARNs de una célula o tejido, varía según condiciones y tiempos. Aunque la sangre es un tejido accesible, puede no ser representativa. El análisis del transcriptoma podría reflejar mejor los estados de salud y enfermedad, pero su complejidad dificulta la interpretación individual de los ARNs.

Esta tesis tiene como objetivo principal desarrollar una plataforma que clasifique transcriptomas de sangre entera mediante aprendizaje profundo. Se generó una base de datos de transcriptomas de sujetos sanos y con patologías como insuficiencia cardíaca y aterosclerosis. Se realizaron estudios clínicos para recolectar datos y muestras. Se entrenó una red neuronal residual de 50 capas, logrando una precisión del 93 % en la clasificación de insuficiencia cardíaca y del 73.3 % en aterosclerosis. Los resultados sugieren que el análisis del transcriptoma mediante redes neuronales profundas puede tener aplicaciones clínicas significativas.

Palabras Clave: Transcriptoma; Redes Neuronales; ARN; Aterosclerosis; Sangre

Abstract

Use of artificial intelligence for the classification of whole blood transcriptomes

The development of neural networks has transformed our daily activities, especially in biomedicine, where it has improved the understanding of diseases and diagnosis. Recently, significant advances have been made in the study of RNAs and mass sequencing techniques. The transcriptome, which includes all the RNAs in a cell or tissue, varies according to conditions and times. Although blood is an accessible tissue, it may not be representative. Analyzing the transcriptome could better reflect health and disease states, but its complexity makes the individual interpretation of RNAs challenging.

The main objective of this thesis is to develop a platform that classifies transcriptomes from whole blood using deep learning. A database of transcriptomes from healthy subjects and those with conditions such as heart failure and atherosclerosis was generated. Clinical studies were conducted to collect data and samples. A 50-layer residual neural network was trained, achieving a classification accuracy of 93 % for heart failure and 73.3 % for atherosclerosis. The results suggest that analyzing the transcriptome using deep neural networks may have significant clinical applications.

Keywords: Transcriptom; Neural Network; RNA; Atherosclerosis; Blood

Índice general

Índice de figuras

Índice de Tablas

Introducción

3.1. Transcriptoma

Toda la información de un individuo está guardada en el ADN de su genoma. Esta información es trasladada a moléculas de ARN a través del mecanismo de transcripción. Los transcriptos así generados son traducidos en algunos casos a proteínas a través de los ARN mensajeros (ARNm), o bien pueden ejercer diferentes funciones como ARN no codificante (ARNnc). El concepto de transcriptoma engloba el conjunto completo de transcriptos que se encuentra en una célula o conjunto de células en un determinado momento (Peymani 2022). Entonces, involucra a todos los genes expresados en un contexto biológico determinado. Comparado con el genoma, que se mantiene mayormente estable entre todas las células y durante la vida de un individuo, el transcriptoma es inherentemente dinámico. Obtener la expresión de los genes de un organismo en diferentes condiciones, tejidos o puntos en el tiempo, revela cómo estos genes están regulados por la presión del entorno y las variantes individuales, convirtiéndose en un instrumento para la comprensión de enfermedades (Lowe 2017).

La medición de la expresión de un gen en particular se utiliza para entender la relación entre el genoma y el entorno. Esto se realiza por medio de la reacción en cadena de la polimerasa con transcripción reversa cuantitativa (*qRT-PCR, del inglés quantitative Reverse Transcription Polymerase Chain Reaction*). La *qRT-PCR* es considerada una técnica *gold standard* para conocer los niveles de transcriptos por ser rápida, reproducible, sensible y precisa. Sin embargo, su utilidad para la comprensión de la expresión de muchos genes es limitada (Casamassimi 2017). Para ello, se han desarrollado técnicas de medición de la expresión génica en masa, como los microarreglos (*microarrays*) (Schena 1995) y la secuenciación de ARN (*RNAseq*) (Wang 2009).

Los microarreglos se basan en la hibridación del ADN copia (ADNc) de ARN marcado con fluorescencia con un oligonucleótido sonda unido ordenadamente

a una superficie sólida. La presencia o ausencia del gen expresado se basa en la detección de la fluorescencia. Es una técnica de bajo costo y sencilla, pero con algunas limitaciones. Por ejemplo, la necesidad del conocimiento previo del genoma para la preparación de las sondas no permite la identificación de genes desconocidos. Otra desventaja es el limitado rango para cuantificar la expresión génica, ya que la medición se realiza cuantificando la intensidad de la señal emitida por la fluorescencia, lo que puede llevar a una saturación en genes altamente expresados con lecturas de expresión inadecuadas, o la no detección de genes muy ligeramente expresados.

Con los avances en la secuenciación de próxima generación (*NGS*) se ha logrado la secuenciación del ARN para una cuantificación de la expresión génica más precisa. Si bien inicialmente su costo era muy elevado, actualmente ha disminuido al punto de poder realizar análisis de poblaciones grandes, lo que llevó a desplazar casi por completo a los microarreglos. Dentro de las ventajas que ofrece esta tecnología se pueden nombrar la posibilidad de detectar y cuantificar genes poco expresados, descubrir nuevos eventos de *splicing*, analizar la expresión alelo-específica y detectar fusión de transcriptos (Casamassimi 2017). Además, el rango dinámico de expresión es mucho mayor.

Las tecnologías de NGS se basan en la ejecución simultánea de millones de reacciones de secuenciación en paralelo, en volúmenes muy pequeños, de longitud de lectura relativamente corta (200 pb) y en la generación de gigabases (GB) de datos por experimento. Los pasos comunes en la mayoría de los enfoques NGS son:

1. Obtención de fragmentos de ADN copia retrotranscriptos.

2. Adición de adaptadores específicos a ambos extremos de los fragmentos, los cuales ligarán el fragmento al punto de polimerización y medición.

3. Amplificación clonal mediante el anclaje del fragmento de ADN, a través de sus adaptadores, a una superficie sólida como microesferas o placa de secuenciación (e.g. Emulsión PCR o bridge PCR).

4. Secuenciación (por síntesis o ligación) de los fragmentos y detección simultánea de las bases de forma masiva y paralela.

5. Adquisición de datos crudos -*raw data*- (por ejemplo, captura de fluorescencia en imágenes o detección de iones).

6. Conversión de los datos crudos en bases de nucleótidos (o sea las secuencias y/o lecturas).

7. Conteo bioinformático de las lecturas y normalización para obtener la información de la expresión génica (Mardis 2008, Metzker 2010, Maekawa 2014).

Las lecturas (o *reads*) finales de entre 50 a 150 bases obtenidas de una muestra a partir de NGS, en el caso de genomas conocidos -como el humano-, se mapean y alinean contra un genoma de referencia (Lee-Liu 2012). La sensibilidad y la exactitud de un experimento de *RNAseq* depende del número de lecturas obtenidas para cada muestra. Un número alto de lecturas asegura una buena cobertura de la expresión del genoma y permite encontrar transcriptos poco abundantes (Lowe 2017).

La sensibilidad del experimento se puede incrementar enriqueciendo los ARNs de interés o eliminando ARNs muy abundantes. Estos últimos pueden ser separados mediante sondas que unen sus colas poliadeniladas o los RNAs pequeños pueden ser purificados por tamaño en una electroforesis en gel. Por ejemplo, para remover el abundante y no informativo ARN ribosomal (ARNr) existen sondas taxón-específicas. Para muestras de sangre también se utilizan sondas específicas para la remoción de los RNAs de las globinas. Los transcriptos ribosomales y de la hemoglobina componen más del 90 % de los encontrados en la sangre y, por lo tanto, retirarlos antes de la secuenciación permite una mejor utilización de los recursos.

En el análisis de secuenciación del ARN hay que tener especial consideración al origen de la muestra, ya que los diferentes tejidos o tipos celulares presentan grandes diferencias en los patrones de expresión y en los eventos de splicing alternativo. La accesibilidad y la invasividad en la toma de muestra es un obstáculo importante en la práctica clínica. Aunque la sangre es un tejido accesible con un mínimo de invasividad, puede no ser representativa de la patología en estudio (Peymani 2022). Otra decisión a tomar respecto al tejido es la posibilidad de analizar una fracción particular de la sangre. En muchos casos, se utiliza la separación de las células mononucleares de sangre periférica (*PBMC*), lo cuál es una técnica habitual para la extracción de ARN que sólo contiene linfocitos y monocitos, eliminando los basófilos, eosinófilos y neutrófilos. Esta separación pierde entre el 50 al 80 % de la heterogeneidad de la sangre periférica generando un sesgo en

el análisis (Koks 2021, Xing 2021). Por el contrario, la secuenciación de sangre entera captura el promedio del perfil de expresión de los diferentes tipos celulares presentes en la misma, incluyendo las células hematopoyéticas, sus vesículas extracelulares y el microbioma sanguíneo.

A lo largo de los últimos años se han desarrollado bases de datos, tales como GEO, ENCODE o GTEx, que compilan la información de expresión de genes que ha sido extraída mediante diferentes tecnologías (microarray, secuenciación o mixtas) y de diferentes organismos, tejidos y/o líneas celulares (Abouelwafa 2020, Lachman 2018). Esto ha derivado en una gran cantidad y variabilidad de datos almacenados imposibles de procesar y comprender por las técnicas habituales de laboratorio. El aumento sostenido en la capacidad de generación de datos ha sobrepasado con creces los recursos bioinformáticos tanto humanos como de equipamiento disponibles hasta la fecha, por lo que el análisis bioinformático es altamente apreciado. Por otro lado, debido a la tasa de error inherente a la tecnología (0.1-1 %, o algo mayor en equipos de tercera generación), a la corta longitud de las lecturas (de 100-200 pb) y a la mayor profundidad con la que se realiza la secuenciación, el análisis de los datos de NGS (en especial con datos de humanos) es computacionalmente costoso, intensivo y complejo. En este contexto, sin lugar a dudas, surge la necesidad de desarrollar herramientas bioinformáticas que permitan aplicar con éxito la secuenciación masiva al diagnóstico en la práctica clínica (Kanzi 2020). En la era de los biomarcadores y de la medicina personalizada, la transcriptómica ofrece la posibilidad de realizar la integración de miles de biomarcadores, el desafío entonces es el análisis e interpretación de los mismos.

3.1.1. Uso de la transcriptómica en biomedicina

La transcriptómica no sólo permite ampliar el conocimiento científico, sino que ya se trasladó a la clínica hace unos pocos años para su uso en el diagnóstico molecular de enfermedades. Se comenzó a utilizar como un complemento a la secuenciación del genoma o el exoma completo (WGS o WES) en enfermedades poco frecuentes en las que no se hallaban variantes patogénicas responsables. Surgió como una manera de integrar información funcional para encontrar perturbaciones transcripcionales causadas por cambios genéticos y aumentó la tasa de diagnóstico entre 8 % y un 36 % en diferentes cohortes analizadas (Yépes 2022). Las principales causas detectadas fueron eventos de splicing aberrante y desbalances alélicos,

principalmente expresión monoalélica; también ayudó a confirmar variantes sinónimas patogénicas (Cummings 2017, Kernohan 2017, Gonorazky 2019, Yépes 2022). Es importante recalcar que, aunque la sangre puede no ser representativa de muchas patologías poco frecuentes, existen antecedentes de identificación de enfermedades mendelianas no hematológicas mediante una transcriptómica de sangre entera (Kernohan 2017, Frésard 2019).

En enfermedades frecuentes la transcriptómica se utiliza para encontrar isoformas diferencialmente expresadas que explican la diferencia molecular entre casos y controles o entre diferentes formas o etiologías de la patología (Kõks 2016). Sin embargo, la factibilidad de la traslación de la transcriptómica en masa a la clínica aún tiene muchos interrogantes, como la falta de métodos técnicos y computacionales de referencia. Por lo tanto, es necesario desarrollar una rutina estandarizada que responda al tipo de biblioteca, la técnica de mapeo, la normalización y la prueba de expresión diferencial que debe utilizarse para alcanzar reproducibilidad (Kuksin 2021). En el estudio del cáncer se puede observar un esfuerzo de la comunidad científica en este sentido. Se ha encontrado evidencia clínica relevante cuando se adiciona la transcriptómica a los estudios de secuenciación de exoma y de genoma completo a los pares de tejido normal y tumor. Por este motivo, la transcriptómica ha sido recomendada en la evaluación y el manejo de cáncer recidivante o refractario (Newman 2021). La tendencia, que se observa en la figura 1, de estudios clínicos en los cuales se integra la secuenciación del ARN en masa para el estudio del cáncer evidencia el interés creciente de la comunidad médica (Kuksin 2021). Actualmente, de los 72 estudios clínicos reportados en clinicaltrials.gov que utilizan esta técnica, 24 son estudios relacionados al cáncer.

Otro ejemplo de utilización de la expresión génica en cáncer es la comparación, ya no entre tejido sano y tumoral, sino entre diferentes puntos en el tiempo. En una biopsia líquida se analiza una muestra de sangre, orina u otro líquido corporal con el fin de buscar células cancerosas o trozos pequeños de ADN, ARN u otras moléculas que las células tumorales liberan en los líquidos corporales. La posibilidad de tomar varias muestras a lo largo del tiempo permite comprender los cambios genéticos o moleculares que tienen lugar en un tumor. Monitorear un tumor sólido con células circulantes en la sangre resulta de valor ya que los cambios en la expresión de determinadas vías metabólicas de las células circulantes de tumor (CTCs) están relacionadas a la respuesta terapéutica de los pacientes (Xu 2021). En el trabajo de Sharma y colaboradores desarrollaron un modelo para de-

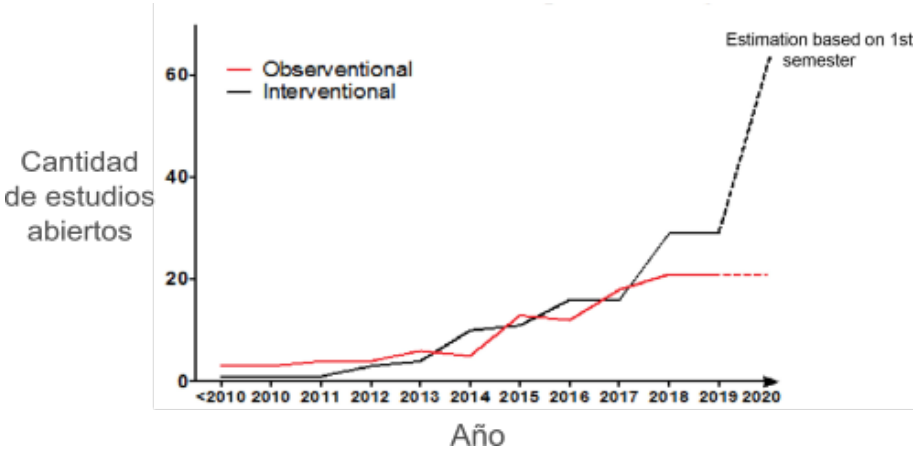


Figura 1: **Utilización de transcriptomas en estudios clínicos en cáncer.** Gráfico de la evolución del número de estudios clínicos de cáncer que utilizaron análisis de secuenciación de ARN en masa entre los años 2010 y 2019 según clinicaltrials.gov. La curva roja representa a los estudios observacionales y la negra a los estudios intervencionales. Tomado de Kuksin 2021.

tectar cáncer de mama en pacientes asintomáticos basados en un patrón de genes a partir de sangre periférica (Sharma 2005).

Finalmente, existen casos en que la transcriptómica se utiliza para diferenciar entre formas o etiologías dentro de una patología. Un ejemplo de ello es la Clasificación de Consenso Internacional (*ICC,* por sus siglas en inglés) para neoplasias mieloides y leucemias agudas. En esta circunstancia se integraron datos morfológicos, clínicos y genómicos para consensuar una clasificación moderna de estas patologías (Arber 2022).

3.1.2. Antecedentes del uso de transcriptómica en enfermedad cardiovascular

El uso del transcriptoma en enfermedades cardiovasculares se encuentra en etapas iniciales. Como ejemplo, Asakura y Kitakaze recapitularon estudios de transcriptómica realizados mediante microarreglos en pacientes con insuficiencia cardíaca reportados entre 2000 y 2009. Comparando los perfiles de expresión de corazones sanos y con insuficiencia cardíaca en etapa terminal lograron determinar 107 genes diferencialmente expresados, muchos involucrados en la disfunción de la mitocondria y la fosforilación oxidativa (Asakura 2009). Más tarde, con el advenimiento de la secuenciación del ARN, más sensible y confiable, Ramirez Flores y otros aunaron el conocimiento de estudios realizados entre 2005 y 2019 generando una base de datos llamada ReHeat (Reference for the HEArt failure

Transcriptome). En este repositorio agruparon los resultados de 16 estudios públicos de transcriptómicas sumando 263 muestras ventriculares sanas y 653 con insuficiencia cardíaca en etapa terminal, combinando microarreglos y secuenciación de próxima generación. Dentro de los resultados se destacan la subregulación del factor de necrosis tumoral alfa (TNFα), NF-κB y el receptor de andrógeno (Ramirez Flores 2021). Vale destacar que estos estudios analizan el transcriptoma del tejido miocárdico, el cual no es de fácil acceso.

A diferencia del cáncer, donde la biopsia líquida es utilizada hace unos años, los antecedentes de estudios del transcriptoma en enfermedad cardiovascular parten, generalmente, de tejido que se ha logrado obtener de forma invasiva, con pocos antecedentes de transcriptómicas de sangre entera. Un ejemplo lo brindan Rosenberg y colaboradores al validar un clasificador de enfermedad coronaria obstructiva utilizando un panel de expresión de 23 genes en pacientes no diabéticos. Para ello extrajeron sangre periférica de pacientes que iban a recibir una angioplastía coronaria y obtuvieron los valores de expresión génica mediante RT-PCR. Los resultados arrojaron una modesta, pero estadísticamente significativa, mejora en la clasificación de los pacientes al compararla con las imágenes no invasivas y los factores clínicos usualmente utilizados en la práctica clínica (Rosenberg 2010). Estos estudios llevaron al desarrollo de un score sexo y edad específico (ASGES) (Voora 2017) que culminaron en el lanzamiento de un producto al mercado disponible para el uso clínico (Corus® CAD) para predecir la presencia de enfermedad coronaria obstructiva en aquellos pacientes con síntomas sugestivos (Gul 2019).

3.2. Enfermedad cardiovascular y factores de riesgo

En 1991 Braunwald y Dzau (Dzau 1991, 2006) introdujeron el concepto de enfermedad cardiovascular como un continuo, una cadena de eventos (Figura 2) iniciada desde factores de riesgo que van influenciando su evolución en mayor o menor medida, por diversos mecanismos a través de muchas vías metabólicas para desencadenar en insuficiencia cardíaca y muerte cardiovascular. Cabe destacar que la presencia de alguno de estos factores dista mucho de significar la presencia de enfermedad cardiovascular de manera taxativa. Son por lo tanto factores que sólo predisponen el desarrollo de la enfermedad cardiovascular aterosclerótica.

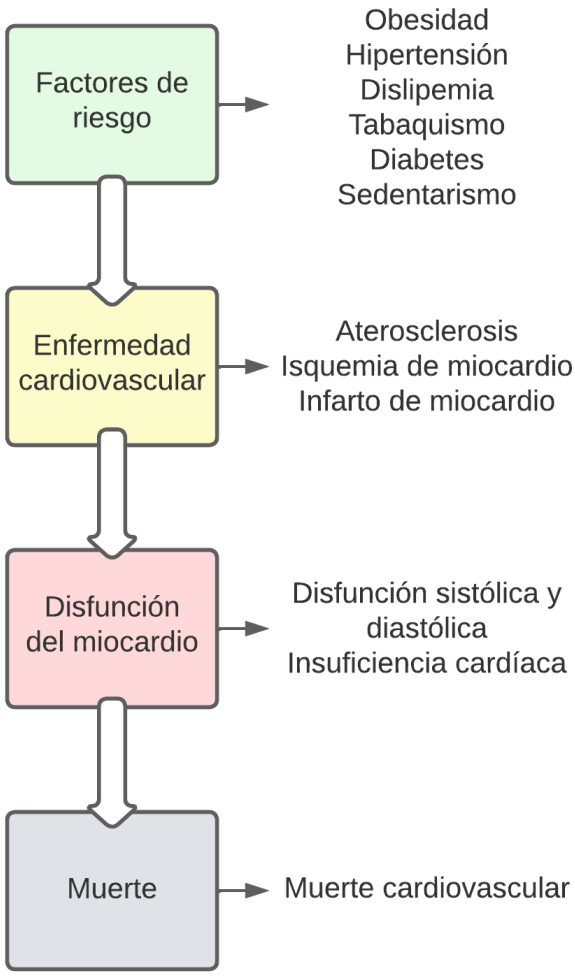


Figura 2: **El continuo de la enfermedad cardiovascular.** Imagen inspirada en Park 2021.

3.2.1. Obesidad

La presencia de obesidad en una persona es un factor de riesgo cardiovascular. La obesidad es una condición que se caracteriza por la acumulación excesiva de grasa corporal y se define por un indice de masa corporal mayor a 30. La obesidad se desencadena al ingerir constantemente en la dieta una cantidad de calorías mayor a la necesaria para satisfacer las necesidades metabólicas. En las últimas décadas se observa una mayor prevalencia de obesidad en la población humana atribuida a un ambiente obesogénico que ofrece rápido acceso a comida y bebida altas en calorías y una disminución en el ejercicio físico diario. Esta condición aumenta el riesgo de desarrollar diabetes y problemas cardiovasculares, entre otras enfermedades (Weihrauch-Blüher 2019, Aras 2021, Faulkner 2021, Jokela 2023). En algunos casos la obesidad puede ser debida a trastornos genéticos monogénicos. El gen MC4R es el más comúnmente implicado, encontrándose disminuído su función en menos del 5% de los obesos pertenecientes a diferentes poblaciones humanas. Este gen codifica al receptor de melanocortina-4. Los niños afectados experimentan un marcado incremento del apetito y se vuelven obesos por hiperfagia. Se han encontrado variantes poco frecuentes en una decena de genes que generan obesidad monogénica (Choquet 2011, OMIM entrada 601665). No obstante, en la mayoría de los casos, la obesidad es multifactorial, siendo el resultado de complejas interacciones entre varios genes y el ambiente, influenciados por la microbiota intestinal y la epigenética (Walley 2009, Herrera 2011, Thompson 2012, Ang 2023).

Como se mencionó, el índice de masa corporal (IMC, definido como el peso expresado en kg dividido por la altura en metros al cuadrado) se utiliza como un indicador general para saber si una persona tiene un peso saludable para su estatura. A pesar de sus muchas limitaciones es ampliamente utilizado para cuantificar masa corporal y clasificar a las personas en obesas, con sobrepeso, peso normal o de bajo peso (Jensen 2013). En la tabla 1 se detalla la clasificación de masa corporal para hombres y mujeres de 20 años de edad o más (Garrow 1985).

Aunque es un indicador comúnmente utilizado por su practicidad, se deben tener en cuenta sus limitaciones (Romero-Corral 2008). Al no tener en cuenta la composición y la distribución de los tejidos -como los músculos, la grasa y los huesos- deben considerarse otras medidas adicionales para determinar un peso corporal saludable. En los adultos, el IMC se ve influenciado por la edad, el sexo,

Clasificación de masa corporal	**Rango de IMC (Kg/m^2)**
Delgadez extrema	<16
Delgadez moderada	16 - 17
Delgadez leve	17 - 18.5
Normal	18.5 - 25
Sobrepeso	25 - 30
Obesidad clase I	30 - 35
Obesidad clase II	35 - 40
Obesidad clase III	>40

Tabla 1: Clasificación para el índice de masa corporal (IMC) para personas de 20 años o más.

la etnia, la actividad física y la masa muscular, entre otros factores. Por lo tanto, según los Centros para el Control y Prevención de Enfermedades (CDC 2022) de Estados Unidos de Norteamérica:

- Los adultos mayores tienden a tener mayor grasa corporal que los adultos jóvenes con igual índice.
- Las mujeres tienden a tener mayor grasa corporal que los hombres a igual IMC.
- Los individuos musculosos y los atletas altamente entrenados tienen menor grasa corporal que los no atletas con igual índice.
- A igual IMC, la cantidad de grasa corporal puede ser mayor o menor dependiendo del grupo étnico.

La obesidad es un factor de riesgo para muchas enfermedades, particularmente para las enfermedades cardiovasculares. En la cohorte del estudio Framingham hay evidencia que el aumento de 1 punto en el índice de masa corporal eleva el riesgo de insuficiencia cardíaca un 5 % en hombres y un 7 % en mujeres (Kenchaiah 2002). Sin embargo también se asocia a la obesidad a una mayor tasa de supervivencia entre los pacientes con insuficiencia, teniendo un 10 % de reducción en la mortalidad cada 5 puntos de aumento del índice de masa corporal (Fonarow 2007). A esta contraintuitiva situación entre los pacientes con insuficiencia cardíaca se la conoce como paradoja de la obesidad. A pesar de su evidencia, actualmente está en revisión ya que nuevos estudios han mostrado que, al ajustar

el índice de masa corporal mediante otras variables pronósticas, la mayor tasa de supervivencia en pacientes obesos desaparece (Butt 2023).

Desde el punto de vista fisiopatológico, la expansión del tejido adiposo visceral es acompañado por la infiltración de células del sistema inmune que inducen inflamación crónica, resistencia a la insulina y desregulación metabólica. La resistencia a la insulina estimula la endotelina-1 que promueve un tono vasoconstrictor elevado y aterogénesis (Triposkiadis 2022). De esta forma, la obesidad está asociada a un espectro metabólico que en su continuo puede llevar a la enfermedad cardiovascular y al fenotipo de insuficiencia cardíaca (Figura 3).

Figura 3: **Esquema de los factores de riesgo que contribuyen al desarrollo de un espectro fenotípico de obesidad.** Tomado de Triposkiadis et al. 2022.

3.2.2. Aterosclerosis

La aterosclerosis es una enfermedad vascular inflamatoria crónica que se define como el engrosamiento de la pared arterial debido a la formación de lesiones conocidas como placas de ateroma. Estas lesiones se localizan en la túnica íntima de arterias de mediano y gran calibre. La placa de ateroma contiene un núcleo lipídico compuesto principalmente por colesterol con una cubierta fibrosa y células inflamatorias. La aterosclerosis se produce por un desequilibrio en el metabolismo lipídico y la inadecuada respuesta del sistema inmune a la acumulación de lípidos en las arterias. Esta placa puede crecer hasta obstruir la luz arterial, impidiendo el

flujo y ocasionando hipoxia tisular en los tejidos ditales a la obstrucción o puede romperse y causar una trombosis vascular obstructiva.

En el año 2020 en la Argentina la enfermedad cardiovascular fue la principal causa de muerte, siendo la responsable del 25,8 % del total de defunciones (Ministerio de salud de la República Argentina 2023).

El desarrollo de la aterosclerosis es un proceso lento y progresivo que comienza en la niñez y se mantiene asintomático por varias décadas. Las complicaciones de la misma -infarto de miocardio, accidente cerebrovascular o isquemia vascular periférica- aparecen a partir de los 30 años habitualmente y se tornan mucho más frecuentes en edades avanzadas. Los factores de riesgos conocidos para el desarrollo de la enfermedad incluyen edad avanzada, hipertensión, diabetes, dislipemia, hiperhomocisteinemia, obesidad, tabaquismo y sedentarismo. En general, ninguno de estos factores de riesgo por sí mismo es suficiente para causar una lesión aterosclerótica (Singh 2002) y, por otro lado, algunos de ellos presentan mecanismos fisiopatológicos en común. La aterosclerosis es una enfermedad compleja que involucra un proceso fibroproliferativo inflamatorio que se desenvuelve en la capa más interna de la pared arterial a través de una serie de eventos patológicos inflamatorios involucrando al sistema inmune, la coagulación y mecanismos de manejo de colesterol y lípidos. La placa aterosclerótica evoluciona secuencialmente a partir de un daño en el endotelio, una disfunción endotelial, resultando en la deposición de lípidos por diferentes células hasta la formación de la placa característica.

3.2.2.1. Disfunción endotelial e inflamación

El endotelio vascular es un órgano endócrino ubicado estratégicamente entre la sangre y la pared vascular que lleva a cabo importantes funciones regulatorias (Singh 2002). Sostiene el balance entre la prevención y la estimulación de la agregación plaquetaria, la trombogénesis y la fibrinolisis, la vasoconstricción y la vasodilatación, la promoción y la inhibición de la proliferación y migración de las células hematopoyéticas (Pacher 2007). La ruptura de este delicado balance se conoce como disfunción endotelial y es una de las primeras manifestaciones de la aterosclerosis.

Los principales eventos que pueden causar disfunción endotelial incluyen, entre otros, a la tensión tangencial generada por el flujo turbulento de la sangre y las especies reactivas del oxígeno producidas por los factores de riesgo cardio-

vasculares como el cigarrillo. Las células endoteliales dañadas o excesivamente activadas secretan vasoconstrictores, tales como ET-1, dismuyen la producción de vasodilatadores, tales como el óxido nítrico, y secretan factores que afectan la diferenciación y el crecimiento de las células musculares lisas de la túnica media. Esto genera quimiotaxis de leucocitos y plaquetas que, a su vez, inducen la expresión de moléculas de adhesión -selectinas, integrinas y proteínas de la superfamilia de las inmunoglobulinas- que interactúan con ligandos específicos en la superficie de leucocitos y plaquetas. Moléculas de adhesión celular como ICAM-1, VCAM-1 y P-selectina median la adhesión de los monocitos al endotelio y la migración a la íntima. Las ICAMs son expresadas en varios tipos celulares, incluyendo leucocitos y células endoteliales. Una expresión defectuosa de las ICAMs no sólo se observa en la aterosclerosis, sino también en otras patologías que interfieren en la función inmune normal. TNF-α, IL-1, oxLDL y el aumento de la tensión tangencial cuando el flujo sanguíneo deja de ser laminar contribuyen a aumentar la expresión de ICAM-1. El aumento de ICAM-1 aumenta a su vez los depósitos de fibrinógeno y la adhesión de los monocitos, seguido por la migración subendotelial, evento crítico en la formación de la placa. La citoquina IL-1 está involucrada en la activación del factor de transcripción NF-κB que regula la transcripción de genes involucrados en la formación de MCP-1 y moléculas de adhesión como ICAM-1.

La ET-1 tiene actividad vasoconstrictora y mitogénica en las células de músculo liso de la túnica media que resulta en la liberación de radicales libres y de citoquinas proinflamatorias a la circulación. En los sitios de daño o inflamación, las citoquinas proinflamatorias, como IL-1 y TNF-α promueven la adhesión y la activación de los leucocitos. También se promueven activadores de neutrófilos como el factor estimulante de colonia de macrófago, el factor activador de plasminógeno e IL-8. Esto induce la activación de MAP quinasas, como ERK y p38 intracelulares, como transducción de señales de estrés y de factores de crecimiento, potenciando la activación de los neutrófilos aumentando su adhesividad (Takahashi 2001). Los neutrófilos activados contribuyen al deterioro y al daño endotelial produciendo más ROS. El proceso inflamatorio es perpetuado por las células endoteliales dañadas gracias a las actividades antitrombóticas y la reducción de la expresión de factores activadores de plasminógeno. Como el daño endotelial es considerado el estímulo para la migración de los leucocitos, la aterosclerosis es clasificada como una enfermedad inflamatoria.

3.2.2.2. Especies reactivas del oxígeno (ROS)

El sistema renina-angiotensina-aldosterona es el encargado de regular la presión sanguínea, el volumen de líquido extracelular y mantener el balance sodio-potasio fisiológico. La angiotensina II es el principal mediador del sistema. Un desbalance en sus niveles sistémicos o locales promueve la aterosclerosis por la formación de ROS en macrófagos, células endoteliales y células de la vasculatura lisa aumentando la actividad NADH/NADPH oxidasa, la que oxida a los lípidos en las partículas de LDL en el segmento del vaso afectado. Simultáneamente, el aumento en la expresión de citoquinas como TNF-α, IL-1 y PDGF, estimula aún más la producción de ROS y la proliferación celular (Figura 4) (Singh 2002). La angiotensina II también ejerce efectos proaterogénicos, proinflamatorios y procoagulantes en plaquetas y monocitos.

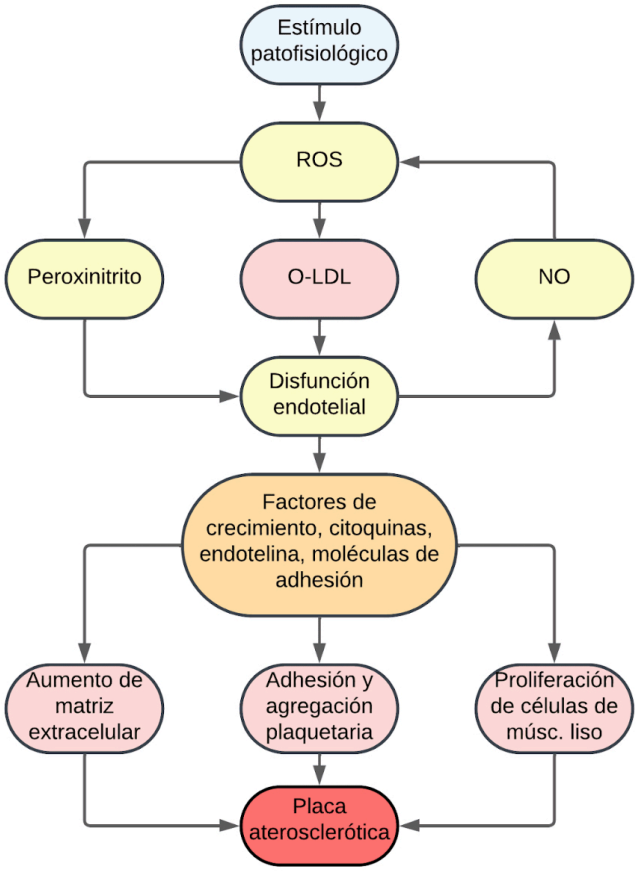


Figura 4: **Diagrama esquemático de la formación de la placa aterosclerótica.** Extraído de Raja B Singh et al, 2002. ROS: especies reactivas del oxígeno, NO: óxido nítrico, O-LDL: Lipoproteína de baja densidad oxidada. Realizado con Lucidchart.

El óxido nítrico (NO) es un mensajero intercelular ubicuo que modula el flujo sanguíneo, la trombosis y la actividad neuronal. El óxido nítrico reacciona, sin

necesidad de catalización enzimática, con el superóxido (O2•-) formando el oxidante fuerte peroxinitrito (ONOO-) altamente tóxico para la célula. Las enzimas superóxido dismutasas (SOD) remueven el superóxido previniendo su acumulación y minimizando su toxicidad en condiciones fisiológicas. El NO posee un efecto importante anti-aterosclerótico, con efecto antiagregante plaquetario, antioxidante, antiproliferativo, anti-inflamatorio y vasodilatador. Dentro de la familia de las óxido nítrico sintasas (NOS), en el sistema cardiovascular se presenta una forma constitutiva, la NOS endotelial (eNOS) y una forma inducible (iNOS). La inducción ocurre usualmente en ambiente oxidativo por citoquinas proinflamatorias, por lo que la alta disponibilidad de NO (Figura 4) forma mediadores proaterogénicos, como el peroxinitrito, que modifican lípidos y proteínas que median la lesión aterosclerótica. Las ROS y las especies reactivas del nitrógeno a su vez activan a NFκB (Pacher 2007).

La agregación plaquetaria se promueve por la disminución del efecto anticoagulante de óxido nítrico y por el aumento de la molécula procoagulante inhibidor 1 del activador de plasminógeno (PAI-1) (Morrow 2020). La Angiotensina II también promueve la proliferación celular de la vasculatura lisa, factores de crecimiento autócrinos, la expresión de enzimas proinflamatorias (fosfolipasa A2 y la NADP/NADPH oxidasa) e induce la transcripción de protooncogenes. Todas estas vías tienen en común ser proaterogénicas por ser generadoras de ROS.

3.2.2.3. Lípidos

En el plasma humano los lípidos más comunes son el colesterol, triglicéridos, fosfolípidos y ácidos grasos. Estos lípidos insolubles son transportados en el plasma en complejos lipoproteicos micelares incluyendo a la lipoproteína de muy baja densidad (VLDL), lipoproteína de baja densidad (LDL) y lipoproteína de alta densidad (HDL). Dependiendo de cuál de estas partículas se encuentra desregulada, los desórdenes en el metabolismo de lípidos se expresan como hiperlipidemia o hiperlipoproteinemia (Poznyak 2020). La lipoproteína LDL está formada por una estructura externa de lípidos, colesterol y fosfolípidos, y un núcleo que consiste en el colesterol y triglicéridos que transporta. En la parte externa de una partícula LDL se encuentra una proteína esencial para mantener su estructura: la apolipoproteína B (ApoB). La ApoB es sintetizada por el hígado y está presente en quilomicrones, VLDL, IDL, LDL y partículas Lp(a). La lipoproteína HDL está

compuesta, al igual que el LDL, de colesterol, triglicéridos y numerosas apolipoproteínas (Apo-AI, Apo-AII, Apo-AIV, Apo-AV, Apo-C1, Apo-CII, Apo-CIII y Apo-E). La apolipoproteína AI es la principal encargada de mantener la estructura de HDL y colabora en funciones enzimáticas que permiten el transporte, reciclaje y degradación del colesterol desde los tejidos periféricos hasta su eliminación en el hígado. ApoE es la principal proteína componente de las VLDL y funciona como ligando en la eliminación mediada por receptor de los quilomicrones y la VLDL remanentes en el hígado.

Las lipoproteínas son modificadas por diferentes reacciones químicas en la pared arterial, de las cuales las principales son la oxidación llevada a cabo por las ROS y la glicosilación no oxidativa en aquellos pacientes con diabetes mellitus. Las lipoproteínas oxidadas (oxLDL) son más aterogénicas que la LDL nativa y lleva al reclutamiento de macrófagos al lugar de la lesión. El sistema inmune identifica estas lipoproteínas modificadas como exógenas y los leucocitos se infiltran para eliminar estos oxLDL de la íntima arterial. Este es uno de los motivos por los cuales la aterosclerosis es vista también como una forma de enfermedad autoinmune (Sima 2018). Los macrófagos con las lipoproteínas modificadas fagocitadas forman células espumosas y entran en apoptosis dejando una capa lipídica en la íntima que progresa a placa aterosclerótica. Finalmente, otro tipo de LDL modificado es el LDL modificado enzimáticamente (E-LDL) que activa a los macrófagos y al sistema del complemento, también induce una citoquina quimioatractante (MCP-1) selectivamente que estimula la secreción de IL-6 y la proliferación de las células musculares lisas vasculares (Singh 2002).

3.2.2.4. Angiogénesis

La angiogénesis es el crecimiento de nuevos capilares a partir de estructuras vasculares preexistentes. Tiene un rol crucial en la progresión de la aterosclerosis y, en particular, en su complicación aguda llamada ruptura de placa. La placa aterosclerótica puede presentar microhemorragias a partir de estos neovasos, o romperse por su fragilidad, lo que lleva a la exposición de las estructuras subendoteliales al torrente sanguíneo, que a su vez desencadena la formación de trombos oclusivos. La oclusión de un vaso por trombosis resulta en el cuadro clínico conocido como síndrome coronario agudo o infarto de miocardio.

El factor de crecimiento endotelial vascular A (VEGF-A) es inducido por hipoxia

en el engrosamiento de la íntima y activa receptores de las células endoteliales de los vasos sanguíneos cercanos a la placa aterosclerótica. Las células activas, también llamadas células punta, liberan proteasas para abrirse paso hacia la zona de hipoxia. Las proteínas mediadoras de inflamación, tales como IL-1 y TNF-α, aumentan la expresión de moléculas de adhesión (VCAM e ICAM) en la superficie endotelial. VCAM-1 y E-selectina solubles son secretados y son mediadores en la angiogénesis. Las células forman brotes que luego formarán aros hasta convertirse en un vaso completo. Esta neovascularización permite el suministro de nutrientes y promueve la infiltración de los macrófagos, el depósito de lípidos y la inflamación en la progresión de la lesión aterosclerótica (Camaré 2017).

3.2.2.5. Progresión de la placa y calcificación

Luego de meses o años de evolución, la placa aterosclerótica genera zonas de inflamación crónica que lentamente comienzan a depositar calcio. Los mediadores del proceso inflamatorio y un elevado contenido de lípidos en el contexto de una lesión aterosclerótica inducen una diferenciación osteogénica de las células de músculo liso vasculares. Este proceso no es exclusivo de la placa aterosclerótica, ya que se puede observar en otros tejidos luego de un proceso de inflamación crónico. Estas células sufren una transdiferenciación a células tipo osteoblasto elaborando vesículas calcificantes y secretando factores que disminuyen la capacidad de reabsorción mineral de las células tipo osteoclasto (Liu 2015). La vía del receptor activador del ligando de NF-κB y osteoprotegerina podría ser la conexión entre la osteoporosis y la calcificación coronaria (Demer 2008). La calcificación ocurre en serie con la progresión de la aterosclerosis severa. Usualmente comienza como micronódulos (0.5 a 15.0 μm) y luego progresa a partículas de mayor tamaño que forman depósitos laminares (mayores a 3 mm) en las arterias.

La Tomografía Computarizada (TC) se utiliza habitualmente para la identificación y medición del calcio coronario. El score Agatston ha sido el método tradicionalmente utilizado para expresar la carga de placa calcificada en el conjunto del árbol arterial coronario y está basado en el análisis corte a corte de las imágenes adquiridas sobre un estudio de TC sin contraste intravenoso. Se ha establecido un valor arbitrario de 130 UH para separar las calcificaciones verdaderas de otros pixeles de alta densidad. En cada corte, el usuario señala una región de interés (ROI) alrededor de un grupo de placas que se encuentra en el curso de una arteria

coronaria y el programa calcula el área de todos los píxeles por encima de 130 UH y la multiplica por un factor de ponderación (Ci), que depende de la máxima densidad de la placa. El score Agatston es la suma de los scores individuales de todas las placas (Agatston 1990).

3.2.2.6. Sexo

Es importante mencionar que las guías actuales para el diagnóstico, la investigación y el tratamiento de la enfermedad cardiovascular aún no discriminan entre los sexos y están basadas en estudios con mayor cantidad de hombres (Woodward 2019), por lo que las mujeres tienen mayor posibilidad de experimentar retrasos en el diagnóstico y menores posibilidades de recibir atención médica (Mosca 2004, Gurgoglione 2023). Algunos trastornos del embarazo, como la preeclampsia antes de las 34 semanas, aumentan el riesgo de aterosclerosis por la disfunción endotelial resultante (Powe 2011). La menopausia está relacionada a un aumento del colesterol ligado a lipoproteínas de baja densidad, la disminución de la concentración de estrógenos y el aumento de los andrógenos generan mayor riesgo de enfermedad cardiovascular, también el uso de anticonceptivos orales combinados (Poznyak 2023, Curtis 2006). Otros factores como ovario poliquístico (Daan 2014), una menarca precoz, primer embarazo a edad temprana, antecedentes de aborto, muerte fetal, parto prematuro y bebés con bajo peso al nacer se asocian con un aceleramiento en la aterosclerosis (Geraghty 2021). Algunos algoritmos para predicción del riesgo no tienen en cuenta los factores específicos por sexo, por lo que tienden a subestimar el riesgo en mujeres, aún cuando la enfermedad cardiovascular es la principal causa de muerte en el sexo femenino a nivel mundial, la mujer continúa estando en desventaja frente al hombre en prevención primaria, secundaria y tratamiento (Woodward 2019).

3.2.2.7. Aspectos clínicos del desarrollo de la aterosclerosis

La detección temprana y precisa de aquellos individuos con alto riesgo de desarrollar enfermedad cardiovascular continúa siendo un desafío. El enfoque tradicional para identificar a individuos con mayor riesgo de desarrollar enfermedad cardiovascular o muerte por esta causa se basa, en gran medida, en la identificación de variables individuales (Diabetes, colesterol, tabaquismo, etc.), un grupo de variables convertidas en un score (Framingham Risk Score, SCORE, QRISK®,

etc.) o biomarcadores individuales (Proteína C reactiva de alta sensibilidad (PCR-Hs), péptidos natriuréticos tales como el BNP, etc.). A lo largo de los años se han desarrollado estrategias basada en el uso de diferentes marcadores clínicos y de laboratorio implicados en la fisiopatología de la aterosclerosis (Chiorescu 2022). La estratificación del riesgo para los pacientes susceptibles a un evento cardiovascular agudo (habitualmente definido como infarto de miocardio, accidente cardiovascular o muerte) es fundamental para la aplicación en ellos de terapias preventivas.

Las ecuaciones de riesgo actuales sintetizan múltiples factores de riesgo de enfermedad cardíaca e infarto. El FRS (Framingham risk score) es un método muy usado para estratificar el riesgo cardiovascular y ofrece una estimación individual sexo-específica de desarrollar un evento cardiovascular fatal o no fatal (infarto de miocardio, ACV) en 10 años (D'Agostino 2008). La Asociación Americana de Cardiología (American Heart Association) desarrolló recientemente la ecuación PREVENT para estimar el riesgo de eventos fatales y no fatales a 10 años (Khan 2024). La Sociedad Europea de Cardiología recomienda el uso del algoritmo SCORE2 (Systematic coronary risk evaluation 2) que estima la probabilidad de eventos fatales en las poblaciones europeas a 10 años. Aunque en el trabajo concluyen que "el futuro de la estimación de riesgo debería expresarse como años de exposición a un perfil de riesgo cardiovascular en lugar de riesgo sobre un período fijo de tiempo, como a 10 años, y cómo los avances en genética permitirían una estimación del riesgo individualizada desde la niñez" (Graham 2021). En la figura 5 se puede observar como ejemplo el cuadro que propone la Organización Mundial de la Salud para clasificar el riesgo de la población de la zona sur de Sudamérica cuando no se cuenta con datos de estudios de laboratorio.

A pesar de estos avances, estos enfoques poseen una capacidad limitada de predicción. Por ejemplo, el área bajo la curva (AUC) ROC para el Score de Framingham es de 0.61 (Rosenberg 2010, Wang 2020). Esta limitación es explicable por diferentes motivos, por ejemplo, el 20 % de los pacientes con enfermedad coronaria no tiene factores de riesgo tradicionales y el 40 % sólo tiene uno (Khot 2003). Además, los factores de riesgo predisponen a la enfermedad cardiovascular, pero no son necesariamente la causa de la misma. Es claro entonces que existe la necesidad de desarrollar nuevas metodologías para mejorar la detección de aquellos pacientes de alto riesgo.

Las imágenes cardiovasculares, por otra parte, son útiles para detectar precozmente el desarrollo de esta enfermedad, pero requieren de equipamientos costosos

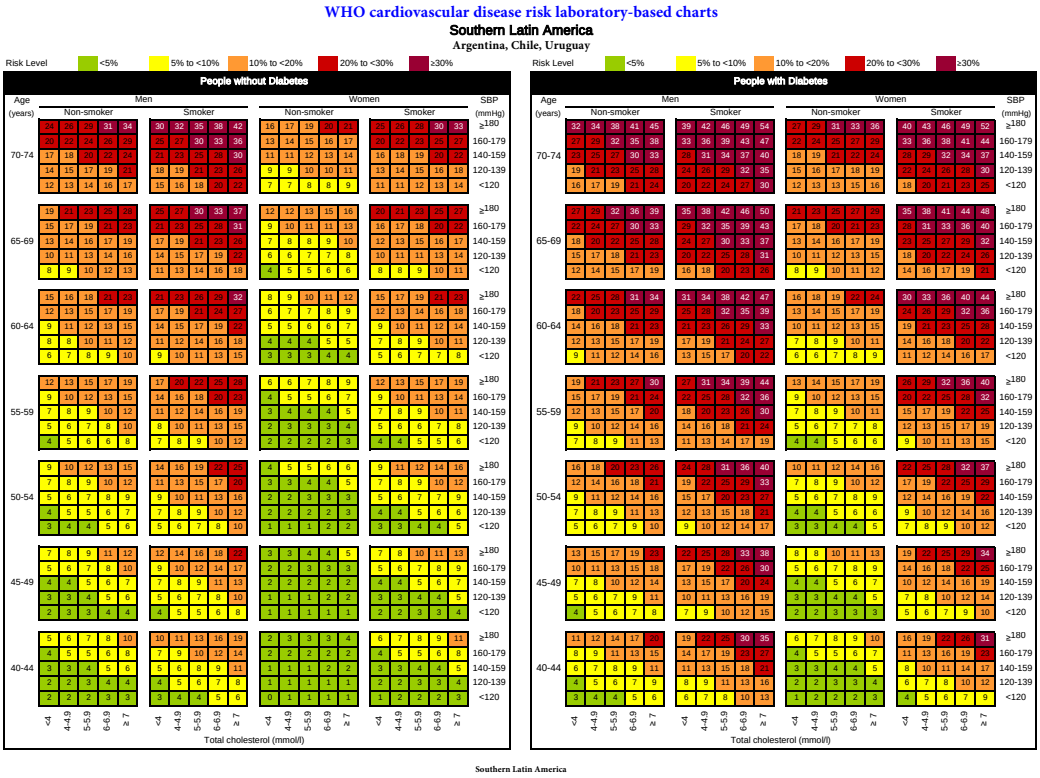


Figura 5: **Cuadros para la clasificación de riesgo cardiovascular.** Cuadros propuestos por la Organización Mundial de la Salud para la clasificación de riesgo cardiovascular en la población de Argentina, Chile y Uruguay. Izquierda: Individuos sin diabetes. Derecha: Individuos con diabetes.

y personal especializado. Esto limita en parte su aplicación clínica. Asimismo, las alteraciones tisulares estructurales detectables en una imagen corresponden a daños establecidos, y los mismos pueden ser en muchos casos irreversibles. A pesar de estas limitaciones, en los últimos años se han establecido aplicaciones para la detección y estratificación de la enfermedad coronaria. Una de ellas, y la más popular, es la detección y cuantificación de la presencia de calcificaciones en las arterias coronarias. La aterosclerosis, como proceso inflamatorio crónico, sufre a lo largo del tiempo el depósito de sales de calcio como se explicó previamente. Estas calcificaciones presentan una correlación estrecha con la presencia de aterosclerosis y con su evolución clínica. Mientras que la calcificación por sí misma no tiene una manifestación clínica específica, se encuentra asociada a una mayor cantidad de eventos cardiovasculares y mayor mortalidad (Budoff 2010, Hou 2012, Saremi 2012). Esto es debido a a dos motivos. En primer lugar, la presencia de calcio coronario se asocia con la presencia de placas lipídicas que pueden romperse y generar trombosis intravascular. En segundo lugar, la zona calcificada en la arteria coronaria puede sufrir un proceso conocido como erosión endotelial, en el cual existe una exposición de las capas intimales al flujo sanguíneo y se produce una trombosis plaquetaria, llevando a un síndrome coronario agudo.

El calcio coronario es un marcador de la presencia de enfermedad coronaria aterosclerótica. Su medición, por medio de radiografías o tomografías, es utilizado para evaluar el riesgo de eventos cardíacos graves a futuro (Shreya 2021) y reclasificar los pacientes en categorías más precisas. En la figura 6 se destaca en forma gráfica la correlación entre la categoría de score de calcio coronario y los eventos cardíacos (infartos) en las principales cohortes estudiadas: MESA (Bild 2002), Framingham (Hoffmann 2008), Heinz Nixdorf RECALL (Schmermund 2002) y Rotterdam (Oei 2002). Las guías clínicas de Estados Unidos y Europa consideran que el score de calcio coronario puede mejorar la evaluación de riesgo cardiovascular en los pacientes asintomáticos. Además, los análisis de costo-efectividad concluyeron que la evaluación del score de calcio coronario es costo efectiva en la población asintomática, aunque las aseguradoras de salud aún no cubren las imágenes para este propósito (Greenland 2018).

El esfuerzo de la comunidad médica por mejorar los clasificadores de riesgo en enfermedades coronarias se hace patente a lo largo de una variedad de trabajos (Panahiazar 2015, Taslimitehrani 2016, Peng M. 2023) demostrando el interés de encontrar una herramienta con mayor grado de sensibilidad y especificidad para

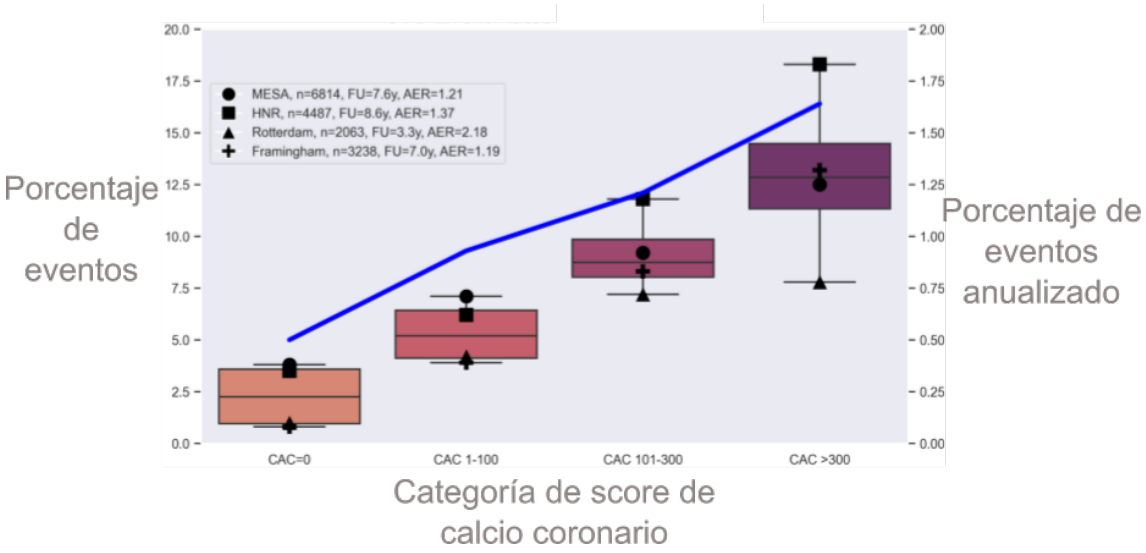


Figura 6: **Número de eventos cardiovasculares de acuerdo a la severidad del score de calcio coronario.** Gráfico de la cantidad de eventos cardiovasculares en porcentajes en cada cohorte (MESA: Multi-Ethnic Study of Atherosclerosis, HNR: Heinz Nixdorf RECALL, Rotterdam, Framingham) por categoría de score de calcio coronario. CAC=0: Calcificación no identificable. CAC 1 a 100: Calcificación leve. CAC 101 a 300: Calcificación moderada. CAC >300: Calcificación severa. FU: Seguimiento. AER: Tasa de eventos anualizada. Datos tomados de Greenland 2018.

el manejo de los pacientes.

Dentro de los biomarcadores usados en la clínica se encuentra el péptido natriurético cerebral (BNP), nombrado así por haber sido descubierto en el cerebro porcino. Tanto la forma activa como la forma inactiva, NT-proBNP, de esta hormona peptídica se pueden medir en el suero sanguíneo. El péptido natriurético es secretado en las paredes auriculares y ventriculares del corazón en respuesta a aumento de volúmenes intravasculares y de las presiones intracavitarias. Sus acciones son inducir la natriuresis y la diuresis. En contextos patológicos tiene efectos de inhibir la activación neurohormonal y la remodelación cardíaca (Alcidi 2022). Los individuos con niveles de BNP en plasma mayor a 100 pg/mL o 300 pg/mL de NT-proBNP deben continuar siendo evaluados para diferenciar una enfermedad cardiovascular de otras patologías que aumenten el volumen sanguíneo.

Los péptidos natriuréticos también son utilizados en diagnóstico y pronóstico en pacientes con insuficiencia cardíaca crónica debido a que correlacionan estados más avanzados de la enfermedad con niveles más altos de péptido natriurético. Cada 100 pg/mL de aumento en plasma de BNP existe un aumento del 35 % de riesgo relativo de muerte (Doust 2005). También, los niveles elevados de BNP agudos lograron predecir el aumento de estancia en el hospital en pacientes hospitalizados por insuficiencia cardíaca. Aunque es una metodología con una alta sensibilidad, no es específica de insuficiencia cardíaca, por lo tanto, no se utiliza como una única guía en el manejo clínico (Novack 2023). Asimismo, se debe tener especial cuidado en pacientes con comorbilidades. Los niveles de péptido natriurético cerebral en sangre son más altos en pacientes con falla renal, diabetes y síndrome coronario agudo y más bajos en personas con un índice de masa corporal en el rango de obesidad (Novack 2023).

Otro biomarcador utilizado en la atención clínica es la proteína C reactiva, una proteína plasmática de fase aguda que aumenta su nivel en respuesta a la IL-6 para activar el sistema del complemento (Sheriff 2021). A pesar de ser un marcador de inflamación e infección inespecífico, puede orientar el riesgo, el diagnóstico y el seguimiento de la enfermedad cardiovascular. En aterosclerosis, los niveles de proteína C reactiva de alta sensibilidad (hs-CRP) en sangre señalan la cantidad de inflamación de bajo grado que está presente en las paredes arteriales. La medición clasifica el nivel de hs-CRP en bajo ($<$1.0 mg/L), intermedio (1.0 – 3.0 mg/L) o alto ($>$3.0 mg/L). Las personas que pueden beneficiarse de esta información son las que se encuentran ya clasificadas como de riesgo cardiovascular intermedio.

Debido a que las personas de riesgo alto se encuentran igualmente en tratamiento y las personas de bajo riesgo no se encontrarán con un riesgo tal que requiera cambios en su estilo de vida, aún con niveles de hs-CRP alto. En Estados Unidos (AHA, CDC y NACB) la recomendación es no medir hs-CRP en la población general para determinar su riesgo cardiovascular, sino medir a los pacientes con riesgo intermedio (10–20 % de riesgo a 10 años) para decidir si comenzar la prevención primaria con estatinas (Ridker 2003, Arnett 2019).

3.2.2.8. Aspectos genéticos

Los procesos fisiopatológicos implicados en el desarrollo de las enfermedades son modulados por factores genéticos y ambientales. Algunos estudios (Marenberg 1994, Zdravkovic 2002) indican que la herencia genética explica entre un 38 y 57 % de los eventos cardíacos fatales. Hasta el momento se han identificado más de 300 loci independientes asociados con la enfermedad aterosclerótica en estudios de asociación de genoma completo (GWAS)(Aragam 2022, Tcheandjieu 2022). Además, se han priorizado 114 genes a través de estudios de asociación de transcriptoma completo (TWAS) (Li 2022). También se han encontrado 5 firmas moleculares de lesiones ateroscleróticas avanzadas mediante transcriptomas masivos de las placas (Mokry 2021) relacionadas al metabolismo, la respuesta inmune y en la homeostasis de la matriz extracelular, las cuales correlacionan con rasgos histológicos y síntomas clínicos. Sin embargo, a pesar de los avances, la descripción comprensiva y unificadora de las firmas genéticas identificadas con la causa o la asociación con aterosclerosis continúa siendo imprecisa y parcial (Örd 2023). Es de notar que, si bien estas mutaciones se asocian con mayor riesgo de eventos cardiovasculares, las mismas no son, en general, determinantes.

La incorporación de la genética en las ecuaciones de riesgo ofrece la oportunidad de refinar los riesgos potenciales más tempranamente y así lograr estrategias de reducción de riesgo individualizadas. Tener un progenitor con antecedentes de enfermedad cardiovascular prematura eleva las chances de desarrollarla en un 50 %, independientemente de los factores de riesgo clínicos (Lloyd-Jones 2004). Estudios de comparación de gemelos monocigóticos y dicigóticos muestran que la variación en el desarrollo de enfermedad coronaria es atribuible a variantes genéticas frecuentes (Zdravkovic 2002), lo que agrega evidencia a que la genética podría ser aditiva en la predicción del riesgo (Crous-Bou 2016, O'Sullivan 2022). En

los estudios de asociación del genoma completo (GWAS) se confirmó la base poligénica de las enfermedades cardiometabólicas, mostrando que muchas variantes (SNVs) de bajo riesgo colectivamente identifican a pacientes con un riesgo cardiovascular significativo (Mars 2020). De esta manera se desarrollaron los conocidos como scores de riesgo poligénico (*polygenic risk scores*, PRS), los cuales resultan de la suma ponderada por el peso de su efecto de las múltiples variantes de nucleótido único asociadas a la enfermedad a través del genoma (O'Sullivan 2022). Se ha recabado vasta evidencia de los aportes del riesgo poligénico a la mejora en la clasificación de los pacientes (Tikkanen 2013, Inouye 2018, Elliott 2020). Esta mejora es detectable antes de la aparición de los factores de riesgo clínicos y es apreciable en todo el espectro de edades, poblaciones y ancestrías (Weale 2021, Lu 2022, O'Sullivan 2022). En el estudio de Riveros-Mckay y otros se concluye que adicionar el score de riesgo poligénico mejora la habilidad predictiva y la utilidad clínica de las herramientas existentes de pronóstico de riesgo para enfermedad cardiovascular a todas las edades y para ambos sexos, pero es especialmente pronunciada para hombres de mediana edad (40 a 54 años) (Riveros-Mckay 2021). La evidencia de los estudios clínicos realizados demuestra que la clasificación de riesgo genético alta puede cambiar el manejo clínico de los pacientes. Entre los individuos de mediana edad asintomáticos considerados de riesgo intermedio por los factores de riesgo convencionales un riesgo poligénico alto de enfermedad cardiovascular ayudó en la reclasificación de la prescripción de estatinas. Además entre los asintomáticos jóvenes puede ayudar a intensificar los esfuerzos a un cambio de vida más saludable y un potencial inicio del consumo de estatinas más temprano para mitigar el riesgo a largo plazo (Aragam 2020).

Aunque el score de calcio coronario y el riesgo poligénico representan tecnologías totalmente diferentes y no han sido comparadas directamente ambas demostraron mejorar la capacidad predictiva de las ecuaciones que solo utilizan los factores de riesgo. Además la evidencia sugiere que las habilidades predictivas parecen ser, al menos, comparables (Patel 2021, Saad 2021) aunque aún la evidencia no es concluyente (Khan 2023).

Las variantes génicas comentadas hasta ahora son variantes germinales, es decir, son heredadas y por lo tanto pueden ser encontradas en todas las células del cuerpo del individuo ya que estaban presentes en las gametas de los progenitores. Las variantes somáticas, sin embargo, ocurren mitóticamente en cualquier célula del organismo (exceptuando las germinales) luego de la fecundación que dio ori-

gen a ese individuo y no son heredables entre generaciones. Estos cambios en el ADN suelen ser silentes, ya sea por no producir ninguna alteración en el funcionamiento celular, o encontrarse en una células aislada sin repercusión orgánica. Si los cambios acumulados aportan a la célula y a sus clones una ventaja frente a las otras células del organismo (variantes conductoras o *drivers*) la variante puede conducir a manifestaciones clínicas. Particularmente interesante es el caso de ciertas variantes con pérdida de función de los genes DNMT3A, TET2 y ASXL1 que confieren una ventaja selectiva a las células madre hematopoyéticas. Estos cambios son detectables en los clones circulantes en sangre periférica ya que estas células mantienen la capacidad de diferenciarse en granulocitos, monocitos y linfocitos. Más del 10 % de las personas mayores de 70 años portan este tipo de variantes aumentando 10 veces el riesgo de desarrollar algún cáncer hematológico frente a las personas que no las portan (Jaiswal 2014). A las personas portadoras de este tipo de variantes y que no presentan ninguna anormalidad hematológica se las definió como portadoras de hematopoyesis clonal de potencial indeterminado (CHIP por sus siglas en inglés) (Steensma 2015). El 75 % de los casos de CHIP se encuentran en los genes DNMT3A, TET2 y ASXL1. Otras variantes se encontraron en el gen JAK2 el cual está asociado a mayores tasas de trombosis, PPM1D y TP53 responsables de respuesta a daño celular, y en los genes de splicing SRSF2 y SF3B1 (Haring 2022).

Jaiswal y otros encontraron una asociación significativa entre enfermedad cardiovascular o infarto de miocardio temprano y CHIP en humanos de una manera dosis dependiente. Además, observaron experimentalmente un empeoramiento de la aterosclerosis en un modelo murino con la variante más frecuente de TET2 (Jaiswal 2017). Las variantes en DNMT3A, TET2 y ASXL1 aumentan el riesgo de evento coronario en 2 veces mientras que la variante V617F en JAK2 se asoció a un incremento del riesgo de 12 veces (Haring 2022). La secuenciación de ARN de las células portadoras de las variantes de pérdida de función en TET2 mostraron un aumento de la expresión de citoquinas proinflamatorias como IL-1β e IL-6 (Jaiswal 2017). La búsqueda poblacional de las personas portadoras de CHIP podría ayudar a los médicos a la identificación y el manejo de pacientes con mayor riesgo cardiovascular. Una aplicación terapéutica posible sería la modulación de las vías inflamatorias mediante fármacos específicos. También la inhibición de JAK2, o de las integrinas río abajo, para reducir las complicaciones de la enfermedad cardiovascular trombótica (Haring 2022).

3.3. Insuficiencia cardíaca

La insuficiencia cardíaca es un síndrome clínico debido a la incapacidad del corazón de bombear sangre en los volúmenes adecuados que demanda el metabolismo de los diferentes órganos. Esta incapacidad generada por anormalidades cardíacas estructurales o funcionales provoca signos y síntomas característicos, incluyendo congestión pulmonar, retención de agua sistémica y baja perfusión tisular. Estas alteraciones pueden, en muchos casos, ser graves y llevar a la muerte del paciente por diferentes complicaciones.

La insuficiencia cardíaca es muy frecuente, con una prevalencia de 1 a 3 % en la población adulta de países industrializados y es especialmente prevalente en edades avanzadas, llegando al 4-5 % en mayores de 70 años. La incidencia estimada es de 1 a 20 casos cada mil habitantes por año (Savarese 2023). La etiología más frecuente en el mundo es la miocardiopatía isquémica producto de alteraciones ateroscleróticas en la irrigación del corazón, siendo responsable en el 50 % de los casos. En Sudamérica le sigue como principal causa la cardiomiopatía chagásica (Stanaway 2015). Dentro de otras causas que pueden desencadenar la insuficiencia cardíaca se incluyen la hipertensión arterial, la enfermedad valvular, la cardiomiopatía dilatada idiopática (habitualmente producto de alteraciones genéticas o por alteraciones estructurales luego de infecciones virales), la cardiomiopatía inducida por quimioterapia y las cardiopatías congénitas (Savarese 2023).

La insuficiencia cardíaca puede presentar grados de severidad clínica muy variados llegando, en muchas oportunidades, al fallecimiento a los pocos años del diagnóstico (Emelia 2019). Para poder clasificar la gravedad y afectación se han desarrollado diferentes escalas. De ellas, la más conocida y utilizada es la propuesta por la Asociación Cardíaca de New York (NYHA classification), la cuál agrupa a los pacientes de acuerdo a la gravedad de los síntomas clínicos. Aquellos pacientes en clase funcional I no presentan prácticamente limitaciones para la actividad física. En el otro extremo se encuentran los pacientes en clase funcional IV, que debido a las alteraciones extremas en la función cardíaca presentan síntomas en reposo (Dolgin 1994). Además de clasificar a los pacientes de acuerdo a su sintomatología, la clase funcional es una forma de categorizar el riesgo de muerte de los pacientes. De esta forma, los pacientes en clase funcional I tienen una mortalidad anual de 1-2 %, mientras que aquellos pacientes en clase funcional IV tienen

una mortalidad anual mayor al 8-10 %.

Además de la capacidad funcional, la insuficiencia cardíaca también se clasifica de acuerdo al volumen de sangre que eyecta en cada latido, lo que se denomina habitualmente fracción de eyección (FEy). Cuando la eyección es menor o igual al 40 % se la denomina insuficiencia cardíaca con fracción de eyección reducida, medianamente reducida entre 41 y 49 % y preservada con FEy ≥50 % (Bozkurt 2021). El 50 % de los afectados tiene insuficiencia cardíaca con FEy reducida.

La insuficiencia cardíaca con FEy del ventrículo izquierdo reducida es el resultado de varios procesos fisiopatológicos que culminan en la pérdida parcial de la capacidad contráctil del corazón, ya sea regional (afectando sólo una zona del ventrículo) o global (afectando todo el ventrículo). Estos procesos corresponden a diferentes aspectos del aparato contráctil cardíaco desregulados que conducen al deterioro de la fracción de eyección. Desde la activación neurohormonal sistémica hasta el sustrato metabólico energético para lograr la contracción cardíaca, pasando por la regulación de los iones implicados y variantes patogénicas en proteínas de importancia mecánica, muchos de los procesos necesarios para el bombeo eficiente del corazón se encuentran alterados. Las consecuencias, sin embargo, son similares en todos los casos: congestión pulmonar por efecto retrógrado de flujo sanguíneo, congestión sistémica y baja perfusión tisular por bajo flujo anterógrado. Estas alteraciones conducen a fenómenos fisiopatológicos asociados con las consecuencias clínicas, la fracción de eyección se encuentra entonces fuertemente asociada a la mortalidad de estos pacientes.

La acción mecánica de la contracción depende de la acción coordinada de los cardiomiocitos, que son mayoritariamente proteínas contráctiles (actina y miosina) y mitocondrias. La contracción comienza (Figura 7) con un potencial de acción que genera la liberación de cationes de calcio desde el canal de Ca^{++} tipo L sensible al voltaje, el que a su vez estimula la liberación de calcio mediante los receptores de rianodina (RyR) ubicados en el retículo sarcoplasmático. El Ca^{++} se une a la troponina C e induce cambios conformacionales en la actina, exponiendo los sitios de unión a la miosina, permitiendo que actina y miosina se unan activando la contracción cardíaca (sístole). La contracción se produce porque la interacción de la actina y la miosina produce una conformación tridimensional más corta de la estructura fibrilar. Posteriormente, la relajación o diástole requiere la captación activa del Ca^{++} hacia el retículo gracias a las Ca^{++} ATPasas del retículo endo/sarcoplasmático (SERCA-2a) y hacia el exterior mediante intercambiadores

sodio-calcio (Ge 2019), generando el desacople de actina-miosina y el retorno al estado relajado.

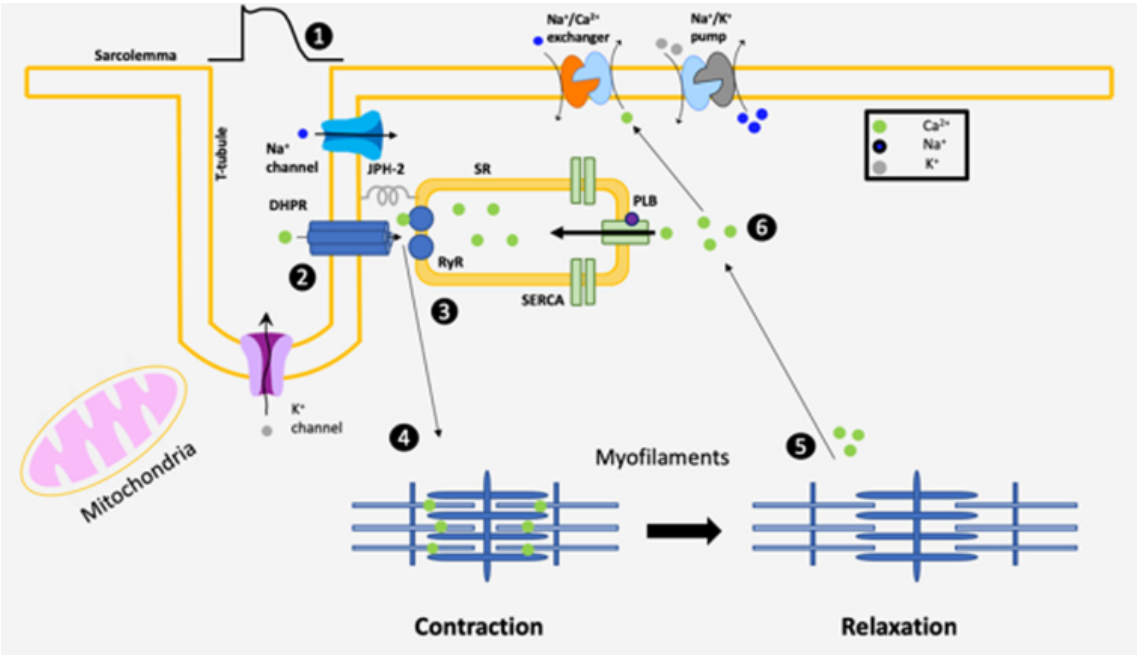


Figura 7: **Acoplamiento excitación-contracción en el cardiomiocito.** El potencial de acción cardíaco (1) se inicia por la entrada de iones de sodio Na^+ por sus canales cambiando el potencial intracelular de negativo a positivo. Esto estimula la liberación de Ca^{++} desde los canales de calcio tipo L (LTCCs) (2) sensibles al voltaje que se encuentran en invaginaciones de la membrana plasmática llamadas túbulos T. El aumento de Ca^{++} estimula la liberación de más Ca^{++} desde los receptores de rianodina (RyR) (3) ubicados cercanamente en el retículo sarcoplasmático. La contracción (4) y la relajación (5) de los miofilamentos son dependientes de la unión y de la disociación del Ca^{++} de la troponina. La limpieza del Ca^{++} interior ocurre hacia el retículo sarcoplasmático (SR) vía la ATPasa de Ca^{++} del retículo sarco/endoplasmático (SERCA) o hacia el exterior vía el intercambiador sodio-calcio (Na^+ Ca^{++} exchanger) (6). Imagen tomada de Ge et al 2019.

En la insuficiencia cardíaca con FEy reducida se encuentran anormalidades en el flujo de calcio que reducen la amplitud del calcio transitorio y el calcio en el retículo, lo que perjudica la contracción y la relajación. Estas anormalidades pueden ser una reestructuración de los túbulos T (Pinali 2017, Frisk 2021), desacoplamiento de los receptores de rianodina (Waddell 2023) y menor actividad de la ATPasa de calcio del retículo sarco/endoplasmático (Mora 2023, Kho 2023). Aunque esto ni siquiera soslaya el nivel de complejidad, ya que si se tienen en cuenta las proteínas con función mecánica que están involucradas en la estructura de los túbulos T, como la JPH-2 (junctophilin-2) (Howe 2021); los reguladores de la SERCA, como el fosfolamban (PLB) (De Genst 2022) o su regulación postraduccional (glutationalizacion (Stammers 2015), sumoilación (Kho 2011, Peng 2023)) se continúan agregando capas de dificultad en la fisiología de la insuficiencia car-

díaca.

En las etapas iniciales de la enfermedad se activan los sistemas adrenérgicos y renina-angiotensina-aldosterona como mecanismo de compensación. Pero la hiperactividad neurohormonal lleva a una desensibilización donde disminuyen los adrenorreceptores, por lo tanto disminuyen los niveles de AMP cíclico para, finalmente, disminuir los niveles intracelulares de Ca^{++} (Iravanian 2008).

Otro punto en la patofisiología de la enfermedad es el metabolismo energético. El continuo trabajo mecánico del corazón depende de la generación constante de ATP. Hasta el 70 % de la energía es utilizada para la contracción y el 30 % para mantener las bombas de iones, especialmente la SERCA, por lo tanto, cualquier decaimiento en el proceso puede llevar rápidamente a la disfunción contráctil. En normoxia el 95 % del ATP generado en el corazón deriva de la fosforilación oxidativa en la mitocondria, utilizando como sustrato mayoritariamente ácidos grasos y en menor medida glucosa, cuerpos cetónicos y lactato; el otro 5 % corresponde a la glicólisis anaeróbica (Doenst 2013). En la insuficiencia cardíaca se observa un cambio en el sustrato usado, aumentando la dependencia de la glucosa y disminuyendo la de los ácidos grasos. Este desbalance no se debe a una reducción en la disponibilidad de los ácidos grasos, sino a una alteración en la transcripción de genes involucrados en su oxidación. Los receptores activados por proliferadores de peroxisomas (PPARs) son los reguladores centrales del metabolismo de lípidos en el corazón, siendo el PPARα el factor de transcripción más expresado en cardiomiocitos y el responsable de controlar la expresión de genes involucrados con la oxidación de los ácidos grasos (Da Dalt 2023). Entre otros factores de transcripción y promotores, PPARα se encuentra regulado negativamente en la insuficiencia cardíaca (Rowe 2010). Esto genera el paradigma del agotamiento energético a pesar de una correcta disponibilidad del sustrato. Además la oxidación de piruvato derivado de la glucosa disminuye en el metabolismo anaeróbico, acumulando lactato y piruvato. El miocardio de pacientes con insuficiencia cardíaca tiene reducida la expresión de la enzima piruvato deshidrogenasa, de los transportadores de piruvato mitocondrial y de las aminotransferasas de piruvato, sugiriendo que el metabolismo y el transporte del piruvato se encuentran reducidos (Lopaschuk 2021).

Los defectos en la función mitocondrial también son causas fisiológicas de la insuficiencia cardíaca. Cualquier disminución o interrupción en la cadena de transporte de electrones resulta en un inadecuado funcionamiento del manejo de iones,

un aumento de especies reactivas del oxígeno que llevan a apoptosis de miocitos, remodelación patológica del miocardio y disfunción contráctil (Wu 2022).

3.4. Redes Neuronales - Aprendizaje profundo

En los primeros tiempos de la inteligencia artificial el campo rápidamente solucionó problemas intelectualmente difíciles para los humanos, pero relativamente sencillos para las computadoras, problemas que pueden ser definidos por una lista formal, por reglas matemáticas. El verdadero desafío para la inteligencia artificial fue resolver situaciones que eran fáciles de resolver para las personas, pero difíciles de describir formalmente, problemas que se resuelven intuitivamente, que parecen automáticos, como reconocer un automóvil en una foto. La solución a estos problemas fue permitir a las máquinas aprender de la experiencia y entender el mundo como conceptos jerarquizados, en los que cada concepto es definido por la relación de conceptos más sencillos (neuronas) (Hopfield 1982). Si graficamos cómo estos conceptos se van construyendo unos sobre otros el gráfico es "profundo", con muchas capas. Por esto se llama a este enfoque aprendizaje profundo (Goodfellow 2016). El aprendizaje profundo es un subtipo del aprendizaje automático. Para los algoritmos de aprendizaje automático es muy importante la representación del dato que se le brinda, diferentes tipos de información relevante llamados rasgos (features en inglés). Por ejemplo, si quisiéramos reconocer autos en imágenes podríamos seleccionar el rasgo ruedas. La elección de la representación tiene un efecto enorme en la performance del modelo, muchos problemas se pueden resolver fácilmente diseñando los rasgos correctos para extraer, pero en otros es muy difícil, hasta imposible, saber qué rasgos elegir debido a la sofisticación de las representaciones posibles. En el ejemplo del auto, las ruedas pueden variar según la posición del auto, la perspectiva de la imagen o cómo la luz refleja sobre estas. El aprendizaje profundo resolvió este problema central ya que expresa las representaciones en función de otras representaciones más sencillas (conexionismo) logrando la variedad de matices necesaria, este tipo de aprendizaje automático permite al sistema computacional aprender con experiencia y datos (Goodfellow 2016). Con muchas imágenes de automóviles el sistema aprende simultáneamente muchas variantes de muchos rasgos no planteados de antemano. Una parte del éxito actual de estos algoritmos de aprendizaje se debe a que podemos proveerlos

de la cantidad de datos necesaria para que sean productivos. La cantidad de datos digitalizados y su interacción en red es enorme y continúa aumentando, lo que permite una buena generalización a partir de nueva información luego de observar sólo una parte del universo de los datos, y la medicina no escapa a esta tendencia (Minor 2017). Otra parte del éxito se debe a la capacidad computacional que permitió aumentar el tamaño de los modelos y al advenimiento de las unidades de procesamiento gráfico (GPU), gracias a la industria del video juego, que permitió aumentar la velocidad de entrenamiento (Chellapilla 2006). Estos procesadores permiten el cálculo simultáneo de una enorme cantidad de procesos, lo que acelera los tiempos de entrenamiento de los modelos.

El nombre de la unidad básica de procesamiento que interacciona para alcanzar el complejo sistema en red, la neurona, está inspirado en las neuronas biológicas por su capacidad de activación al vencer un umbral de estímulo. Matemáticamente, los estímulos son los valores de entrada que pueden provenir de un archivo de entrada con datos o de otras neuronas. La neurona (Fórmula 1) realiza una suma ponderada de los valores de entrada (x_j). La ponderación es el peso (w_j) que cada estímulo tiene sobre la respuesta de la neurona (y) más un término de sesgo (b). El umbral que va a determinar si la neurona se activa o no es definido por una función, la función de activación (f). Una función sencilla que logra eliminar la linealidad de la neurona y permite la suma entre neuronas para formar la red. Es importante recalcar la relación no lineal entre neuronas, una propiedad muy conveniente a la hora de aplicar la red neuronal en redes genéticas.

$$y = f(WX + b) = f(w_1 x_1 + w_2 x_2 + \cdots + b) = f\left(\sum_j w_j x_j + b\right) \tag{1}$$

Las neuronas que reciben la misma información se dice que se encuentran en la misma capa. Así todas las neuronas de una misma capa entregan el resultado de su cómputo a la siguiente capa, de manera secuencial y jerarquizada. La manera en que se organizan las diferentes capas es la arquitectura de la red. A las capas que se encuentran entre la capa de entrada y de salida se les llama capas ocultas. Por ejemplo, en la figura 8 se observa una red neuronal totalmente conectada (*feed foward*) con tres capas ocultas. Las flechas representan la conectividad entre las neuronas o nodos de la red, ambas neuronas de la capa de entrada brindan su información hacia todas las neuronas de la primera capa oculta. A su vez, todas ellas

pasan su resultado a todas las neuronas de la siguiente capa, lo que evidencia un flujo de información direccional que permite la jerarquización de la información.

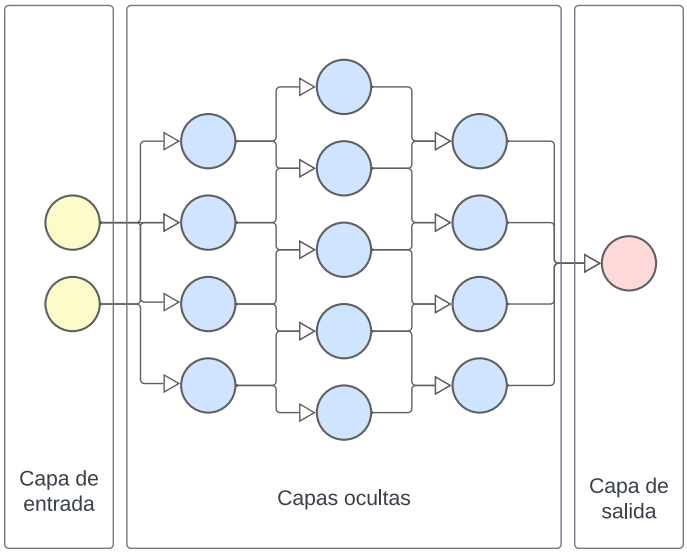


Figura 8: **Representación esquemática de una red neuronal artificial totalmente conectada.** Su arquitectura consta de 1 capa de entrada con 2 neuronas (en amarillo), 3 capas ocultas (2 de ellas con 4 neuronas y la central con 5 neuronas, todas ellas en color celeste) y 1 capa de salida con 1 neurona (representada en rojo).

Una red neuronal artificial es un modelo matemático que logra encontrar similitudes dentro de los datos para resolver un problema, como por ejemplo, de clasificación. Estas similitudes van a permitir que un nuevo dato pertenezca a una clase determinada debido a sus semejanzas con ese grupo: automóviles con automóviles, helicópteros con helicópteros, etc... Matemáticamente esto lo logra ajustando sus parámetros. La riqueza actual de estos modelos yace en la cantidad de parámetros que se pueden utilizar y que, claro está, puede ajustarlos por su cuenta, hacer un aprendizaje automático (*machine learning*). Existen tres tipos de aprendizaje: supervisado, no supervisado y reforzado (que no será utilizado en esta tesis). En el aprendizaje supervisado los datos que se brindan a la red tienen su etiqueta, la foto del automóvil va acompañada del título “automóvil” que previamente alguien identificó y señaló. Por lo tanto, las bases de datos son más complicadas de conseguir y pueden conllevar el sesgo del etiquetador. En el aprendizaje no supervisado se encuentran los algoritmos de agrupamiento (*clustering*) y de reducción de la dimensionalidad. Al no tener el conocimiento a priori de la etiqueta el agrupamiento se efectúa sólo por características inherentes a los datos y la evaluación e interpretación de los agrupamientos son subjetivos, dependiendo de la experticia del operador.

El aprendizaje supervisado de una neurona es un proceso iterativo y se modela como la actualización del vector de ponderaciones (W) en cada iteración.

$$w'_j = w_j + \mathrm{Lr}\,(\delta - y)\,x_j \tag{2}$$

donde w'_j es el elemento en la posición j del vector de ponderaciones W actualizado. La tasa de aprendizaje (Lr) es una constante entre 0 y 1 que modula el error o la diferencia entre la respuesta esperada (δ) y la salida de esa iteración (y). x(x_j) es el elemento en la posición j del vector de entrada X.

Para modelar la diferencia entre el resultado esperado y el resultado real global en cada iteración se necesita una función de error (también llamada de costo o de pérdida) que va a señalar cuál es el error para cada una de las combinaciones de los parámetros (Por ejemplo el error cuadrático medio (*MSE*)). La necesidad de ir reduciendo la diferencia entre el resultado esperado y el resultado observado, que parte de parámetros iniciales al azar, nos lleva a buscar el mínimo de la función de costo. Para ello (Fórmula 3) se hace la derivada de la función de costo en función de los parámetros (dE/dW), resultando en un vector que señala la pendiente de la función en el punto determinado en esa iteración o también llamado: el gradiente (∇) de la función. La actualización de los parámetros (W') en una iteración va a corresponder a restarle a los parámetros de la iteración anterior (W) el gradiente de la función de error, ya que se quiere avanzar en el sentido de disminuir la pendiente hasta que el error sea nulo.

$$W' = W - \mathrm{Lr}\,\nabla f \tag{3}$$

En una red neuronal el error de cada capa depende del error de la capa anterior, así que, para obtener el gradiente, se parte desde el error de la última capa operando de forma recursiva capa tras capa hasta la capa de entrada, retropropagando los errores. Entonces, “si se está mirando en la dirección correcta, lo único que se debe hacer es seguir caminando”. La tasa de aprendizaje (Lr) es el hiperparámetro que indica cuánto moverse en la superficie de la función de error en la siguiente iteración para repetir el cálculo, hasta que la pendiente desaparezca y, por lo tanto, haber llegado a un mínimo. El algoritmo de retropropagación de errores (Rumelhart 1986) es lo que permite a la red autoajustar sus parámetros, aprender, de manera eficiente.

Durante el entrenamiento de una red neuronal se divide el set de datos. Inicialmente, a la red se le muestra un porcentaje mayoritario de las fotos con su etiqueta en cada iteración, esto es llamado grupo de imágenes de entrenamiento. Con este grupo la red autoajusta sus parámetros. El resto de las fotos que la red no utilizó para entrenar es el set de validación y es utilizado para evaluar la performance del modelo. Este grupo es presentado a la red sin su etiqueta para que realice una predicción la cual se compara con la etiqueta conocida. Existen diferentes métricas para la evaluación de la red, una de ellas es la exactitud de la clasificación (Classification Accuracy). La exactitud es un número entre 0 y 1 que representa el número de predicciones correctas sobre el número total de predicciones realizadas. Es una métrica muy sensible al desbalance de los datos, pero funciona muy bien si el número de muestras está equilibrado entre las clases. Otra métrica muy utilizada y que brinda una descripción más completa de la performance del modelo es la matriz de confusión. Como su nombre lo indica es una matriz, en ella se ordenan tanto los verdaderos y falsos negativos como los verdaderos y falsos positivos. Las predicciones correctas (verdaderos positivos y negativos) se encuentran en la diagonal principal de la matriz, los falsos positivos y negativos, o las "confusiones", se encuentran esparcidas en los componentes de la matriz que no pertenecen a la diagonal principal.

Una red neuronal convolucional es un tipo de aprendizaje profundo que aplica convoluciones matemáticas (estrictamente son correlaciones) en una o más capas internas de la red, esta arquitectura permite realizar una gran cantidad de operaciones para lograr aprender de relaciones no lineales entre los datos de entrada y de salida (Gu 2018).

Un problema que surge de los modelos con mucha profundidad es el desvanecimiento del gradiente. El algoritmo de retropropagación de errores, comienza a fallar a medida que se agregan más capas, ya que el gradiente se va haciendo más y más pequeño a medida que nos alejamos de la capa de salida, el resultado es que en las capas más cercanas a la entrada el aprendizaje es nulo (Kolbusz 2017). Una de las maneras que se abordó este problema es colocando una conexión que saltee algunas capas, un atajo que no entra en el algoritmo de retropropagación y puede mantener el gradiente en capas sucesivas (Identity shortcut connection), así aparecen el aprendizaje residual profundo (He 2016) y las redes residuales (ResNet) utilizadas en la presente tesis (Kaiming 2015).

3.4.1. Antecedentes

El aprendizaje profundo se utiliza desde hace un tiempo para resolver múltiples problemas de la bioinformática (Zhou 2015, Jin 2021). Particularmente en transcriptómica de célula única existen programas para evadir el ruido inherente a la técnica, reducir la dimensión y encontrar la señal para el agrupamiento de células y para una expresión génica diferencial mediante *autoencoders* (Eraslan 2019, Grønbech 2020, Tran 2022, Pandey 2023) o redes neuronales profundas dentro de un modelo probabilístico bayesiano (López 2018).

En transcriptómica masiva se utilizó para predecir la expresión diferencial de genes a partir de rasgos en el ARNm y en la zona promotora (Tasaki 2020), a partir de un subgrupo de genes (Subramanian 2017) o a partir de la expresión de un grupo de factores de transcripción (Magnusson 2022). También se aprovechó la reducción de la dimensionalidad de los *autoencoders* para aprender representaciones latentes biológicamente relevantes de un *subset* de genes del transcriptoma para encontrar las diferencias de expresión entre tejidos (Azevedo 2021). Holzscheck y otros utilizaron redes neuronales lineales para hallar información valiosa en envejecimiento en transcriptomas provenientes de la piel. Utilizaron una estrategia que consistió en no permitir conexiones entre neuronas de diferentes vías metabólicas preseleccionadas hasta conseguir una neurona que refleje las características esenciales de cada vía en el proceso de envejecimiento. Una neurona final integra la información y predice una edad para la expresión génica brindada (Holzscheck 2021). En otro trabajo utilizaron redes adversarias generativas (GANs) en datos de secuenciación de ARN masiva para estudiar la progresión de la enfermedad de Alzheimer en un modelo de ratón (Park 2020). Un punto en común a todos los trabajos es la búsqueda de la reducción de la dimensionalidad de la transcriptómica.

En el año 2018 se presenta el trabajo de Lyu y Haque en el cual utilizaron redes neuronales convolucionales para clasificar 33 tipos de tumores prevalentes a partir de datos de secuenciación de ARN. Por primera vez se utilizaron datos transcriptómicos del Pan-Cancer Atlas llevados a imágenes de 2 dimensiones para entrenar una red neuronal convolucional (Lyu 2018). En el mismo año, Ma también aplicó datos de expresión génica disponibles en bases de datos ordenados en imágenes para clasificar los grados de muestras de glioblastomas difusos con redes neuronales convolucionales (Ma 2018).

En el caso de las enfermedades cardiovasculares las redes neuronales artificiales se aplicaron al diagnósticos de infartos y arritmias a partir de electrocardiogramas (Khan 2001). También se han utilizado en la interpretación de imágenes de resonancia magnética y radiografías. Varios trabajos utilizan redes neuronales convolucionales para clasificar pacientes en categorías de riesgo cardiovascular a partir de la determinación del calcio coronario en las imágenes (Cano-Espinosa 2018, Lessmann 2018, De Vos 2019, Chao 2021, Zeleznik 2021). Moon y otros propusieron un modelo de procesamiento del lenguaje natural que predice si una persona es susceptible a enfermedad cardiovascular a partir de identificar factores de riesgo, síntomas, mecanismos y genes asociados a enfermedad cardiovascular a partir de la búsqueda de palabras clave de consulta en bibliografía disponible en PubMed (Moon 2023). En otro trabajo, con el objetivo de encontrar blancos terapéuticos para la progresión de la aterosclerosis, entrenaron una red neuronal artificial totalmente conectada con 5 capas ocultas con la expresión de 5 genes asociados al progreso de la placa aterosclerótica. El estudio se basó en datos de microarreglos de la base de datos GEO y tuvo un filtrado previo de genes a partir de otros algoritmos (Miao 2024).

Hipótesis

Existen perfiles de expresión génica característicos de cada patología, que resultan de la contribución de todos los transcriptos, no solo de los genes diferencialmente expresados, y que pueden ser captados por modelos de aprendizaje automático.

Objetivos

Visto que la actual estrategia para el manejo clínico de los pacientes continúa siendo elegir algunos pocos biomarcadores dentro de las vías metabólicas implicadas en la patología y que las enfermedades poligénicas tienen una correlación genotipo-fenotipo baja, proponemos un paradigma diferente: no utilizar pocos marcadores sino aprovechar la herramienta de las redes neuronales para lograr una clasificación a la luz de la expresión de todos los genes representados en la muestra.

El objetivo principal es desarrollar una metodología capaz de clasificar, por medio de aprendizaje profundo, transcriptomas de diferentes patologías y controles directamente desde sangre entera. Para esto se plantean los siguientes objetivos particulares:

1. Generar una base de datos de transcriptomas de secuenciación propios de sujetos sanos (controles) y de diferentes patologías (insuficiencia cardíaca y aterosclerosis).

2. Generar agrupamientos no supervisados y correlacionar los mismos con los datos clínicos disponibles.

3. Analizar los transcriptomas por medio de redes neuronales y obtener un algoritmo con capacidad de clasificación.

4. Correlacionar los hallazgos de la clasificación con las características clínicas y la evolución de los pacientes.

Materiales y Métodos

6.1. Estudios clínicos

En base a los objetivos propuestos, se realizaron estudios clínicos a fin de poner a prueba nuestras hipótesis. Los mismos se realizaron siguiendo las pautas del método científico en seres humanos voluntarios (Saidon 2005, Thorat 2010) y bajo un estándar internacional ético de buenas prácticas clínicas (Ministerio de Salud Argentina Resolución 1490/2007, International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) General Considerations for Clinical Studies). Cada estudio estuvo a cargo de un investigador principal. Se diseñó el protocolo de estudio, el cual explica y contiene la información en relación a los objetivos, requisitos de elegibilidad, la cantidad de participantes necesarios para el estudio, qué tratamiento se administra, de qué forma, su dosis y frecuencia; qué tipo de información debe recopilarse y cuándo. Además se redactó un consentimiento informado. Estos documentos fueron presentados y aprobados por un comité de ética autorizado. Al momento del enrolamiento en el estudio, los participantes deben atravesar un proceso mediante el cual, luego de recibir información veraz acerca del mismo de manera clara y precisa, de tal forma que pueda ser entendida al grado de que pueda establecer sus implicaciones en su propia situación clínica, documenten por medio de un formulario de consentimiento informado escrito, firmado y fechado, su colaboración voluntaria al estudio (Saidon 2005).

Se realizaron 3 estudios clínicos con diferentes objetivos y poblaciones, incluyendo un estudio en sujetos sanos, un estudio en pacientes en quienes se realizó una evaluación de riesgo cardiovascular y un estudio en pacientes con insuficiencia cardíaca (Figura 9).

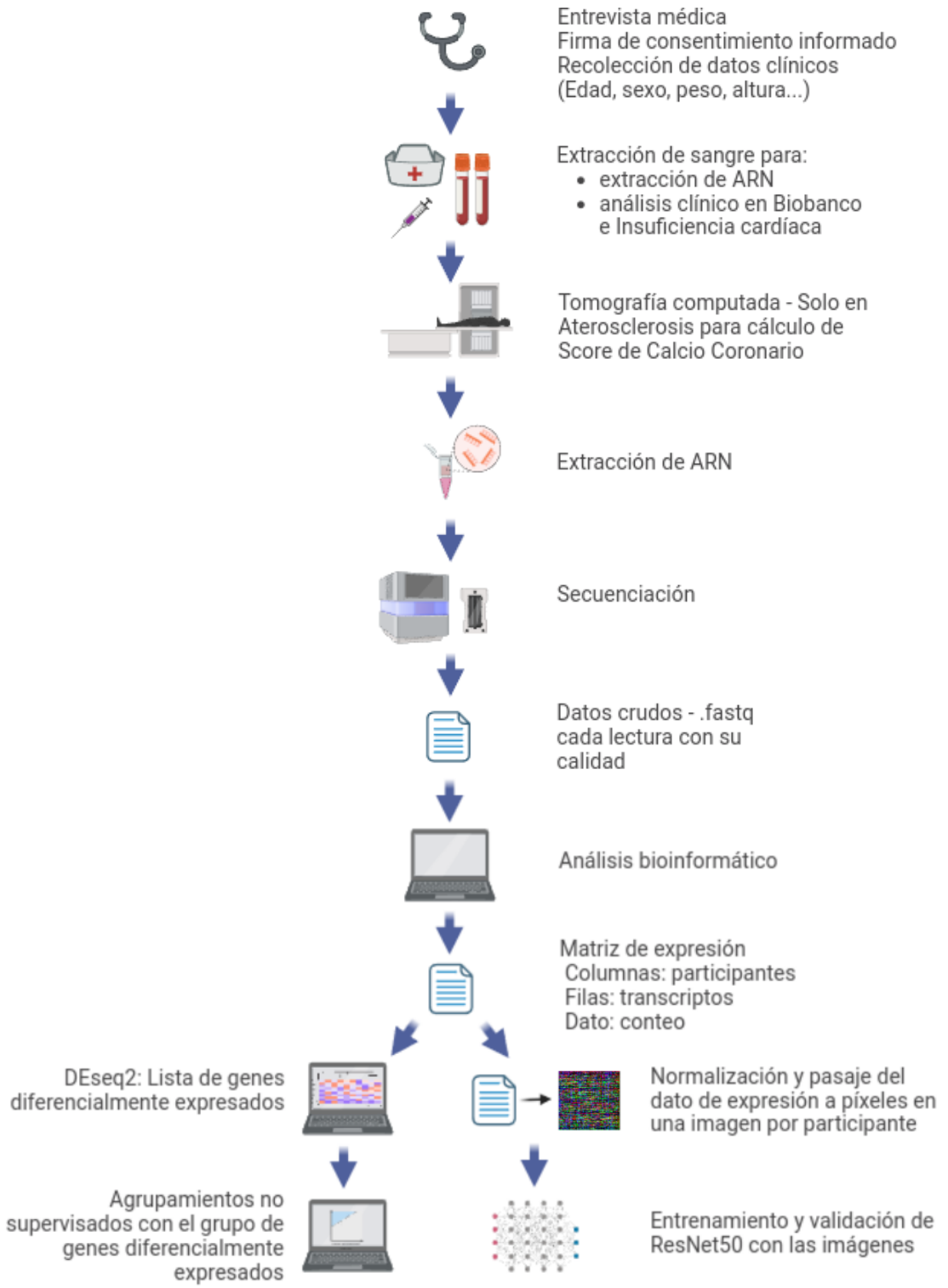


Figura 9: **Esquema del flujo de trabajo bioinformático.** Los 2 ó 3 primeros pasos se llevaron a cabo en el centro de salud correspondiente a cada estudio clínico. La extracción de ARN se llevó a cabo en el Instituto de Neurociencias, Fleni - CONICET. La secuenciación en Illumina, Reino Unido. Y todo el procesamiento de los datos se llevó a cabo en el Instituto de Neurociencias, Fleni - CONICET.

6.1.1. Estudio de sujetos sin enfermedad aguda

El objetivo de este estudio fue determinar el patrón del transcriptoma normal en sangre periférica de sujetos sin enfermedad activa conocida por medio de una secuenciación profunda del ARN de sangre de vena periférica, sin separación celular, incluyendo a los eritrocitos, leucocitos, plaquetas y todo ARN contenido en otras estructuras celulares o vesiculares presentes en el plasma. El objetivo fue considerar a la sangre como un tejido fluido, pero sin diferenciación de las estructuras celulares que la conforman. En este estudio se extrajeron muestras de participantes sanos para lograr asociar la expresión de decenas de miles de genes con la edad, el sexo y los factores de riesgo presentes en la población general.

Se realizó un muestreo de conveniencia incluyendo a empleados de la Fundación para la Lucha contra las Enfermedades Neurológicas de la Infancia (FLENI). Participaron 356 sujetos sanos sin enfermedades activas (aguda o subaguda). Para ser elegidos para el estudio los sujetos debieron cumplir los siguientes criterios:

- Mayor de 18 años.
- Firma de consentimiento informado.
- No presentar enfermedades recientes (en los últimos 3 meses), cardiovasculares (Infarto agudo de miocardio, insuficiencia cardíaca, fibrilación auricular), ni respiratorias (Asma, EPOC, neumonía), ni neurológica (Demencia, Parkinson, accidente cerebrovascular/AIT, migraña, ni ninguna otra que requiera tratamiento farmacológico), ni gastrointestinal (Enfermedad inflamatoria crónica: Crohn, colitis ulcerosa; ni úlcera gástrica o duodenal, ni diagnóstico de H. pylori reciente), ni metabólica (Diagnóstico reciente o cambio reciente en el tratamiento).
- No utilizar en forma crónica o en las últimas 96 horas por cualquier enfermedad ninguno de los siguientes medicamentos: antiinflamatorios, inmunosupresores, antialérgicos, antidiarreicos, antieméticos.
- No presentar síntomas o signos de infección activa en los últimos 14 días, incluyendo infecciones de las vías aéreas superiores, gastrointestinales, dermatológicas o sospecha de infección viral.
- No tener diagnóstico por hisopado nasal de CoViD en los últimos 90 días.

A los individuos se les realizó una entrevista previa para descartar criterios de no inclusión. Posteriormente se les realizó un exámen físico donde se relevaron variables clínicas como el peso, la altura y la tensión arterial. Luego de un ayuno mínimo de 8 horas se les extrajo, por punción de vena antecubital con técnica estándar, una muestra de sangre para un análisis de laboratorio clínico básico y una muestra para la transcriptómica en un tubo TempusTM Blood RNA Tube (Thermo Fisher Scientific). Estos tubos contienen una solución que lisa inmediatamente todas las estructuras celulares y preserva el ARN, estabilizándolo por 48 horas a temperatura ambiente, 2 semanas a 4 grados e indefinidamente a -20°C.

6.1.2. Score de calcio coronario

El objetivo primario de este estudio fue evaluar la precisión diagnóstica de un algoritmo de inteligencia artificial (*Deep learning*) para identificar la presencia de calcificación coronaria estudiada con tomografía computada (TC) y discriminar diversos grados de calcificación a partir del transcriptoma de sangre entera en pacientes asintomáticos y sin antecedentes cardiovasculares. Los objetivos específicos primarios fueron:

- Análisis de los transcriptomas y su clasificación según la presencia y extensión de la calcificación arterial coronaria por TC.
- Entrenamiento y validación del algoritmo de inteligencia artificial (Deep learning) utilizando los transcriptomas.
- Determinar el grado de precisión del algoritmo para predecir diversos grados de calcificación coronaria a partir del transcriptoma del grupo testeo.

Dentro de este estudio observacional se reclutaron 200 pacientes que asistieron al Instituto Médico ENERI y la Clínica La Sagrada Familia para la realización de una evaluación de riesgo cardiovascular. Se los invitó a participar en el estudio y, como parte del protocolo, se les realizó una TC de tórax de baja dosis sin contraste. Se recolectaron datos como el peso, la altura y la tensión arterial mediante una entrevista médica en consultorio. De manera automatizada, se determinó el grado de calcificación coronaria en las imágenes tomográficas. Luego se obtuvo sangre por vena antecubital para el análisis del transcriptoma. Por 5 años se recolectarán datos sobre la incidencia de eventos cardiovasculares fatales y no fatales mediante un seguimiento clínico telefónico.

Los siguientes fueron los criterios de inclusión:

- Mujeres mayores de 50 años y menores de 75, hombres mayores de 40 años y menores de 75.
- Firma de consentimiento informado.
- No presentar insuficiencia renal o hepática, ni enfermedad pulmonar activa (asma agudizado, EPOC agudizado o fibrosis pulmonar), ni enfermedad cardiovascular conocida (infarto previo de miocardio, insuficiencia cardíaca o antecedentes de intervenciones vasculares/valvulares, coronarias, periféricas o cerebrales).
- No cursar hiper o hipotiroidismo.
- No presentar insuficiencia suprarrenal, ni haber tenido cirugías en los últimos 3 meses o algún traumatismo severo los últimos 6 meses. No padecer enfermedad oncológica, ni ninguna patología bajo tratamiento inmunosupresor, ni ninguna otra enfermedad grave con pronóstico de vida estimado menor a 12 meses.
- No haber tenido un diagnóstico de CoViD en los últimos 3 meses.
- No tener embarazo en curso o puerperio menor a 12 meses.

De la misma manera que en el estudio anterior, se extrajo una muestra de sangre para la transcriptómica en un tubo Tempus.

La cuantificación de calcio coronario (CC) se realizó en una tomografía sin contraste de tórax. El tomógrafo utilizado fue un multidetector de 16 cabezales modelo MX8000 IDT de Philips. El cálculo del grado de calcificación coronaria se realizó con un software específico de detección automática. La cuantificación se realizó utilizando el score de Agatston 21. Habitualmente se clasifican los pacientes según los siguientes grados de score de Agatston:

- 0: Patología no identificable
- 1 a 99: Patología leve
- 100 a 399: Patología moderada
- Mayor a 400: Patología severa

6.1.3. Insuficiencia cardíaca

El objetivo primario del estudio fue evaluar la utilidad diagnóstica del análisis transcriptómico de sangre entera asistido por inteligencia artificial en pacientes con antecedentes de insuficiencia cardíaca y deterioro de la fracción de eyección del ventrículo izquierdo. Los objetivos específicos fueron:

- Análisis de los transcriptomas y su clasificación según la fracción de eyección del ventrículo izquierdo.
- Entrenamiento del algoritmo de inteligencia artificial (*Deep learning*).
- Determinar el grado de precisión diagnóstica del algoritmo para identificar diversos grados de deterioro ventricular izquierdo a partir del transcriptoma.

Para este estudio observacional de cohorte prospectivo se reclutaron 116 participantes con diagnóstico previo de insuficiencia cardíaca y deterioro de la función ventricular izquierda que concurrieron a la institución médica Instituto Cardiovascular de Buenos Aires para la realización de un seguimiento en una unidad especializada de insuficiencia cardíaca. Se recolectaron datos clínicos en una entrevista médica y se extrajo sangre para el análisis del transcriptoma y para un análisis bioquímico de rutina. Se recolectarán datos durante 5 años sobre la incidencia de muerte por todas las causas, eventos cardiovasculares fatales y no fatales.

Para poder participar en el estudio de insuficiencia cardíaca los pacientes cumplieron con los siguientes criterios de inclusión:

- Ser mayor de 18 y menor de 80 años.
- Firma de consentimiento informado.
- No ser catalogado por el médico tratante en el estadio A de la clasificación de insuficiencia cardíaca ACC/AHA.
- Tener una fracción de eyección del ventrículo izquierdo menor a 50 %.
- No tener enfermedad renal crónica, ni insuficiencia renal o hepática conocida, ni enfermedad pulmonar activa (asma agudizado, EPOC agudizado o fibrosis pulmonar), ni enfermedad cardiovascular activa (infarto agudo de

miocardio diagnosticado hace menos de 180 días, antecedentes de internación por insuficiencia cardíaca menores a 30 días, intervención coronaria o vascular periférica hace menos de 90 días).

- No estar en lista de espera para trasplante cardíaco.
- No cursar hiper o hipotiroidismo.
- No presentar insuficiencia suprarrenal, ni haber tenido cirugías en los últimos 3 meses o algún traumatismo severo los últimos 6 meses.
- No padecer enfermedad oncológica, ninguna patología bajo tratamiento inmunosupresor, ni ninguna otra enfermedad grave con pronóstico de vida estimado menor a 12 meses.
- No presentar diagnóstico de CoViD en los últimos 3 meses.
- No tener embarazo en curso o puerperio menor a 12 meses.

Todos los estudios fueron aprobados por los Comités de Ética de sus respectivas instituciones y todos los participantes firmaron el consentimiento informado respectivo a su estudio. Todos los datos de los participantes fueron anonimizados durante los análisis.

6.2. Procesamiento de la muestra biológica

Las muestras de sangre fueron procesadas en el laboratorio de FLENI para la extracción de ARN. Para ello se utilizó el kit Tempus™ Spin RNA Isolation Reagent de Thermo Fisher Scientific. El ARN se mantuvo a -80°C hasta el procesado. La preparación de la biblioteca de secuenciación desde sangre entera se llevó a cabo con el kit TruSeq™ Stranded Total RNA with Ribo-Zero™ Globin. Este kit remueve los ARN correspondientes a los ribosomas y la hemoglobina, el cual compone un porcentaje mayoritario en una muestra de sangre y no aporta información alguna. Brevemente, en el primer paso se depletó el ARN ribosomal y de la globina desde el ARN total purificado previamente mediante el Ribo-Zero Plus rRNA Depletion kit ™ según las especificaciones del fabricante. Luego se fragmentó y se desnaturalizó el ARN para colocar *primers* utilizando hexámeros para síntesis de ADN complementario al azar. En el tercer paso se retrotranscribió la primera hebra de

ADNc a partir de los hexámeros, luego se removió el molde de ARN y se reemplazó con la segunda hebra de ADN logrando así el ADNc doble cadena. El quinto paso agrega un nucleótido de adenina en el extremo 3' de los fragmentos para prevenir que se liguen entre ellos durante la reacción de ligado de los adaptadores. Una timina correspondiente en el nucleótido del extremo 3' del adaptador provee la complementariedad para el ligado y asegura una baja tasa de formación de quimeras. En este próximo paso se concatenaron los múltiples adaptadores índice al fragmento que lo prepara para la hibridación a una celda de flujo del secuenciador. Luego existe un paso de limpieza donde se removieron, con Agencourt AMPure XP beads, los restos de la biblioteca. Los fragmentos ya están listos para su amplificación mediante PCR, en el cual se amplifican selectivamente los fragmentos de ADN que tengan los adaptadores a ambos extremos de la molécula. La PCR se realizó con PCR Primer Cocktail que se une al final de los adaptadores. Una vez finalizada la amplificación hay un nuevo paso de limpieza con las esferas Agencourt AMPure XP y un chequeo de calidad y concentración. El chequeo se realizó a partir de 1 μl de una biblioteca con un kit DNA 1000 en un bioanalizador Agilent 2100.

6.3. Secuenciación

Las bibliotecas preparadas fueron secuenciadas en una plataforma Illumina NovaSeq 6000 (Illumina, San Diego, CA, EE. UU.) empleando la química de celda de flujo S4. Se realizó un pool de veinte bibliotecas, por cada carril de la celda de flujo, garantizando una cantidad equitativa de ADN de cada biblioteca. Resultaron 80 pacientes por cartucho. La secuenciación se llevó a cabo utilizando lecturas pareadas de 150 pb, con el objetivo de alcanzar una profundidad mínima de, al menos, 100 millones de lecturas por muestra. La identificación de bases y la evaluación de calidad se realizaron mediante el software Illumina Real-Time Analysis (RTA).

6.4. Flujo bioinformático

A fin de desarrollar un flujo boinformático acorde a nuestras necesidades e intereses, se realizó un extenso *pipeline* con todos los pasos necesarios para procesar los archivos .fastq originados en el secuenciador y que termina en los archivos de

imágenes para alimentar las redes neuronales (Figura 10).

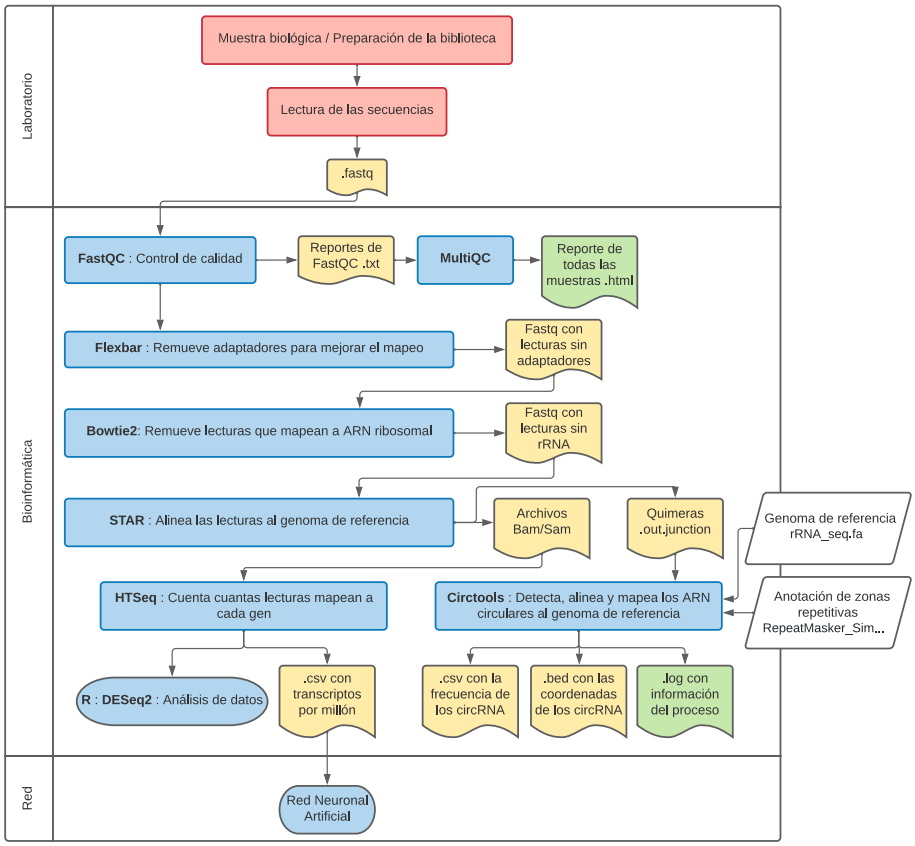


Figura 10: **Diagrama de flujo para** el proceso desde la recepción del RNA extraído hasta la red neuronal artificial.

Se analizó la calidad de las muestras secuenciadas mediante el programa FastQC (Babraham Institute. v.0.11.4) y se obtuvieron informes sobre la cantidad de lecturas, la calidad de la secuencia por base, la distribución de los largos de lectura, el contenido de G-C, la cantidad de lecturas duplicadas y la cantidad de adaptadores remanentes, entre otros parámetros. Dentro de los filtros de calidad se utilizó un Phred score superior a 30 para eliminar lecturas de baja calidad. El nivel de calidad Phred se define como una propiedad que está relacionada logarítmicamente con las probabilidades de error de las llamadas de cada base (P). Si Phred asigna un nivel de calidad de 30 a una base en concreto, las probabilidades de que esta base sea incorrecta es de 1 entre 1000 bases secuenciadas, o sea una precisión de 99.9 %.

Una vez obtenidas las lecturas y controlada su calidad se removieron los adaptadores remanentes mediante el programa Flexbar v.3.0 (Dodt 2012). Se utilizó el programa alineador Bowtie2 v.2.5.1 para mapear lecturas a ARN ribosomal y removerlo. De esta manera se logró un filtrado bioinformático del ARN ribosomal

remanente luego del filtrado biológico en el laboratorio. En el siguiente paso se ordenaron las lecturas y se buscó su ubicación dentro del genoma de referencia GRCh38 por Genome Research Consortium human build 38 o Hg38. Mapear es asignar a una lectura una posición en el genoma de referencia, para esto se utilizó el programa STAR (por *Spliced Transcripts Alignment to a Reference*) (Dobin 2013). Aunque los genomas están compuestos por secuencias de ácidos nucleicos linealmente ordenadas, las células eucariotas generalmente reorganizan la información en el transcriptoma cortando zonas intrónicas y empalmando exones no contiguos para crear transcriptos maduros (*splicing*). Mapear las lecturas de transcriptos al genoma de referencia presenta un gran desafío, principalmente dos tareas que son muy intensas computacionalmente: primero, el correcto alineamiento de lecturas que contienen diferencias en una sola base; segundo, el mapeo de secuencias derivadas de regiones genómicas no contiguas.

Este alineador también encuentra las uniones de *splicing* quimeras para obtener los ARN circulares, devolviendo un archivo .out.junction. Las lecturas mapeadas resultaron en un archivo de formato .sam, un archivo de texto tabulado que presenta un encabezado indicado con el símbolo "@" con información general del alineamiento y luego cada renglón corresponde a una lectura alineada. Cada línea está dividida en campos tales como nombre del fragmento, nombre de la referencia, posición de mapeo, calidad de mapeo, la longitud del fragmento y su secuencia, código CIGAR (código de cómo se encuentra alineada la base a la referencia. Por ejemplo, una "M" si el alineamiento coincide perfectamente o una "I" si se encuentra una inserción respecto de la referencia.), entre otros.

Una vez ubicadas las lecturas al genoma de referencia se pasó a contar cuántas lecturas pertenecían a cada gen mediante HTSeq-count de la librería de Python HTSeq (Anders 2015). Sólo se contaron las lecturas que mapearon estrictamente a un solo gen. Aquellas lecturas que mapeaban a múltiples posiciones o que solapaban con más de un gen fueron descartadas. Otro detalle importante es que se contaron fragmentos y no lecturas, por ser una secuenciación por final de lectura pareado en la cual las dos lecturas dan evidencia del mismo fragmento de ADNc y deben ser contadas una sola vez.

Los conteos sin normalizar se utilizaron para los análisis de expresión diferencial de genes ya que el DESeq2 corrige internamente por el tamaño de la biblioteca y largo del gen. Los genes HBA-T2 (hemoglobin subunit-α2), HBA-T3 (hemoglobin subunit-α1) y β-GLOBIN o CD113T-C (hemoglobin subunit β) fueron removidos

bioinformáticamente de todos los análisis, en caso que algunas lecturas hayan quedado presentes luego de la construcción de la librería.

6.5. Expresión diferencial de genes y contexto biológico

La regulación génica difiere entre tipo celulares y etapas del desarrollo, pero también en respuesta al ambiente y a diferentes estímulos. Entonces, cuando la expresión génica no está debidamente regulada, la homeostasis celular se perturba pudiendo alterar las funciones celulares y llegar, incluso, a generar una patología. Un análisis de rutina para los datos del conteo de genes por muestra de una secuenciación de ARN en dos estados diferentes es la detección de genes diferencialmente expresados en dichos estados. La expresión diferencial es la cuantificación y la inferencia estadística de los cambios sistemáticos entre condiciones. Los datos del conteo de genes es una tabla que asigna el número de fragmentos de secuencia que se mapean sin ambigüedad a cada transcripto (Love 2024). Estos datos no responden a una distribución normal y tienen una dependencia de la varianza con la media.

6.5.1. DEseq2

El paquete estadístico DEseq2 provee un método para testear expresión diferencial utilizando un modelo lineal generalizado (GLM) con una distribución binomial negativa. El marco estadístico ranquea los genes basándose en la estimación del tamaño del efecto, el logaritmo de las veces de cambio (LFC), y testea la expresión diferencial. Se ajusta un modelo lineal generalizado modelando los conteos K de lecturas para cada gen (g) en cada muestra (m) siguiendo una distribución binomial negativa con media μ_{gm} y dispersión α_g (Fórmula 4). Las medias se escalan con un factor de normalización que va a independizarlas de la profundidad de secuenciación. La función de link es el logaritmo en base 2.

$$Kg_m \sim BN(\mu_{g_m}, \alpha_g) \tag{4}$$

Los estimadores de este modelo paramétrico se calculan con un método de contracción (*shrinkage*) y no mediante máxima probabilidad, lo que le brinda mayor

estabilidad y reproducibilidad a los resultados. Para testear la significancia de la expresión diferencial el paquete utiliza el test de Wald. Los p valores obtenidos se ajustan por comparaciones múltiples mediante Benjamini y Hochberg (Love 2014). Para esta tesis se agregaron a la variable en estudio las covariables sexo y edad. A la covariable edad se la dividió en jóvenes (19 años hasta 35), adultos (entre 36 y 64 años) y adultos mayores (desde 65 años hasta 89). Por ejemplo, para el estudio de insuficiencia cardíaca (IC) el modelo planteado resultó el expuesto en la fórmula 5.

$$\log_2(\text{veces de cambio}) = \beta_0 + \beta_1\ \text{femenino} + \beta_2\ \text{edad}_2 + \beta_3\ \text{edad}_3 + \beta_4\ \text{IC} + \varepsilon_m \quad (5)$$

6.5.2. Ontología génica

Las ontologías usualmente consisten en términos, clases o conceptos con relaciones que operan entre ellos. La ontología génica (GO) describe el cuerpo de conocimiento biológico para los genes en tres aspectos: función molecular, componente celular y proceso biológico. En este trabajo nos enfocamos en el proceso biológico definido como el objetivo biológico para el cual un gen o genes contribuyen (Ashburner 2000).

El análisis de enriquecimiento de ontología génica encuentra los términos ontológicos sobrerrepresentados o subrepresentados para un conjunto de genes que se encuentran sobreexpresados o subexpresados bajo ciertas condiciones. Por ejemplo, partiendo del genoma humano, la lista de referencia sería de 20 mil genes codificantes aproximadamente (es necesario aclarar que el genoma humano cuenta con más de 60,000 genes anotados, pero la función de la gran mayoría de los genes no codificantes no es conocida y por ello no son utilizados habitualmente en la construcción de ontología génica). Si 440 genes están involucrados en un término ontológico, entonces el 2.2% de los genes de la lista de referencia mapean a ese término. Si se analiza una lista de 500 genes se espera que 11 genes (500 x 2.2%) en esa lista estén involucrados con ese término. Si hay más genes que los esperados para ese término entonces se habla de sobrerrepresentación y, de igual modo, si hay menos se habla de subrepresentación de ese término en el subconjunto de genes.

El método estadístico usado para encontrar el estimador es el test binomial. Se asume en la hipótesis nula que la probabilidad de encontrar genes para una

categoría particular ($p_{(c)}$) es la misma que para la lista de referencia y tiene una distribución binomial (Mi 2013).

En este trabajo se realizó el análisis de enriquecimiento de ontología génica para el subconjunto de genes que el DEseq2 entregó como resultado con un p valor inferior a 0.05 (Cuando la lista de genes del DEseq2 lo permitió por ser extensa se eligió un p valor más estricto: 0.01). Para esto se utilizó el paquete estadístico de R llamado clusterProfiler (Wu 2021).

6.5.3. GAGE: generally applicable gene set enrichment for pathway analysis

Se realizó un análisis de enriquecimiento de genes mediante GAGE en el lenguaje computacional estadístico R. La estrategia de los métodos de enriquecimiento de un grupo de genes utiliza el conocimiento previo de vías de procesos biológicos (genes anotados juntos gracias a que intervienen en la misma vía biológica). Este análisis determina si estos conjuntos de genes, definidos a priori, muestran una expresión diferencial significativa en diferentes condiciones experimentales. A diferencia de otros métodos que requieren una clasificación previa de los genes, GAGE evalúa directamente los cambios en la expresión génica a nivel de conjuntos de genes.

Para ser utilizado, en primer lugar GAGE necesita la identificación de las vías canónicas (set de genes curado) y el set experimental derivado de la expresión diferencial al asumir que los genes de las vías canónicas están regulados de manera heterogénea y los experimentales están regulados hacia la misma dirección, sobre o subexpresados. Luego, para testear si un grupo de genes está correlacionado significativamente con una condición experimental se examinan las veces de cambio del nivel de expresión génica en la condición experimental vs la condición control. También determina si la media de las veces de cambio de un set de genes es significativamente diferente de la totalidad de los genes (del *background set*). Esto lo realiza mediante una prueba t de Student a dos colas (Fórmula 6).

$$t = \frac{(m - M)}{\sqrt{\frac{s^2}{n} + \frac{S^2}{n}}} \tag{6}$$

Donde m es la media de las veces de cambio de un set particular de genes y M la de todos los genes del set de datos. La desviación estándar se nota como s en

el set particular y como *S* en el set background. Y *n* es el número de genes en el set particular. Finalmente, se calculan los valores p de las comparaciones con las réplicas y, por último, se calcula un valor p global para las múltiples muestras de un set de genes en un meta-test. Para calcular el valor q debe corregir los valores p por múltiples comparaciones utilizando fdrtool que calcula la tasa de falsos positivos (FDR) basado en la distribución nula empírica (Luo 2009).

La enciclopedia de genes y genomas de Kioto es un compendio de conocimiento para el análisis sistemático de funciones génicas (Base de datos llamada GENES) que conecta información genómica con información funcional de un orden superior (Base de datos llamada PATHWAY). PATHWAY es una base de datos con un conjunto de representaciones gráficas de procesos celulares (Kanehisa 2000) en las cuales se pueden mapear las expresiones diferenciales de interés.

En el desarrollo de esta tesis se trabajó con los datos de expresión diferencial obtenidos en el DEseq2 y se buscó el marco biológico a través de GO. También se realizaron estudios de enriquecimiento mediante GAGE con GO y KEGG. Se utilizó la herramienta KEGG PATHWAY mapper para graficar algunos resultados interesantes de destacar.

6.6. Análisis no supervisados

A medida que se agregan dimensiones al espacio matemático de los datos, el volumen crece de manera que entre los datos se encuentra una mayor cantidad de volumen vacío. El matemático Richard Ernest Bellman llamó a este fenómeno la maldición de la dimensionalidad (Bellman 1961). Por ejemplo, dos puntos en una recta tienen una porción de la recta que ocupan y otra porción que no, espacio matemático desconocido. Dos puntos en un plano ocupan proporcionalmente menor lugar obteniendo mayor cantidad de espacio vacío. Lo mismo puede imaginarse en tres dimensiones donde la zona que no ocupan los dos puntos es mucho mayor que en una recta. Así, la cantidad de observaciones necesaria para describir el espacio usualmente crece exponencialmente con la dimensionalidad. Encontrar similitudes entre datos para organizarlos se dificulta cuando estos pueden ser diferentes en tantos sentidos, lo que genera que las técnicas de agrupamiento fallen o sean poco eficientes.

PCA y UMAP son métodos exploratorios ampliamente usados para la repre-

sentación de expresión génica (Diaz 2021, Luecken 2019). La técnica de PCA se utilizó con los conteos crudos de todos los transcriptos (en el orden de los 60 mil) previamente a todos los análisis para la detección y eliminación de valores extremos. En el caso de los agrupamientos no supervisados, para aliviar a las técnicas de PCA y UMAP, se decidió utilizar los genes obtenidos en el análisis de expresión diferencial como forma de reducir la dimensión matemática del análisis. De esta manera, se redujo el número de genes a pocos miles o a decenas en algunos casos.

6.6.0.1. Análisis de componentes principales (PCA)

El análisis de componentes principales es una técnica lineal de reducción de la dimensionalidad que prioriza la representación global de las observaciones (Jolliffe 2016). Es una técnica exploratoria que procura hallar las combinaciones lineales de las variables originales que maximizan la varianza. El test minimiza el error cuadrático entre las distancias de las observaciones en el espacio original y las distancias en el espacio de baja dimensión. Se busca el vector que maximiza la varianza y luego se observa, dentro de las infinitas direcciones ortogonales a éste, cuál maximiza la varianza y así sucesivamente hasta tener una sola opción posible. De esta manera se obtienen nuevas variables (los componentes principales) que son combinaciones lineales de las originales ordenadas decrecientemente según la varianza. Se obtienen coordenadas a partir de los vectores de los componentes principales para graficar las observaciones en el espacio de baja dimensión. Es una buena técnica para el reconocimiento de patrones ya que preserva la estructura general de los datos, las correlaciones de las variables y las varianzas generales de las observaciones. Sin embargo, no preserva las distancias entre los datos, particularmente las distancias pequeñas. No se encuentra optimizada para la cuantificación de la separación de clases (calcular el centro de masa de cada clase en el espacio del componente principal y reportar la distancia euclídea). La alternativa para la separación de clases, el Análisis de discriminantes lineales (LDA), no puede ser aplicada en este caso porque el desbalance en la cantidad de genes excede ampliamente la cantidad de muestras de cada clase (las covarianzas estimadas no tienen la totalidad del rango y no pueden ser invertidas)(Martinez 2021).

6.6.0.2. Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)

UMAP es otra técnica de reducción de la dimensionalidad, pero que, al contrario de PCA, trata de preservar las distancias locales ante las distancias globales (la varianza) (McInnes 2020). Transforma la distancia entre observaciones en una probabilidad de que sea un vecino. Luego, en un espacio de menos dimensiones, recapitula esa probabilidad respetando, sobre todo, las distancias cortas. Busca una representación de las observaciones del espacio de alta dimensión en uno de baja dimensión (2 o 3) que minimice la función de pérdida. Esta función pondera las diferencias de probabilidades medida desde las probabilidades inducidas por las observaciones del espacio original, pondera alto las observaciones cercanas en el espacio original. La función de pérdida se optimiza computando el gradiente descendente, calculando el gradiente de la distancia entre observaciones en cada iteración. Se utilizan hiperparámetros que van a designar:

- El grado de localía (no es igual la distancia para ser vecino en un cluster más denso que en uno más disperso). Por lo tanto, la ventana que indica la distancia a la que una observación es vecina de otra (compromiso local-global) se ajusta según los datos.
- La distancia mínima deseada entre puntos en el espacio reducido (estético).
- El número de épocas de entrenamiento cuando se optimiza la representación de baja dimensión.

Para este trabajo también se utilizó *t-distributed stochastic neighbor embedding* (t-SNE) dentro de los métodos que preservan la información local, pero UMAP se comportó mucho mejor frente a la maldición de la dimensionalidad, formando agrupamientos más claros y que coincidían mejor con las variables clínicas.

6.7. Normalización

Los conteos crudos mapeados a un cierto transcripto no son comparables entre muestras o condiciones porque la profundidad de la secuenciación o los tamaños de las bibliotecas (el tamaño total de lecturas mapeadas) varían de muestra en muestra. Los conteos crudos que mapean a diferentes transcriptos dentro de una

muestra tampoco son comparables porque los transcriptos más largos tienen más lecturas mapeadas a ellos comparados con transcriptos más cortos con un nivel de expresión similar. Entonces la normalización de los datos de secuenciación es necesaria para remover sesgos técnicos. Los transcriptos por millón (TPM) no tienen unidades (Fórmula 7) y además cumple con el criterio de tener un promedio invariante (El promedio de la abundancia de ARN de los genes dentro de una muestra debe ser constante, o sea, la inversa del número de transcriptos mapeados). Los transcriptos por millón (10^6) son las lecturas mapeadas al transcripto iésimo (Ti) sobre largo del transcripto iésimo (Li) divididas por la suma de todas las lecturas mapeadas a los n transcriptos luego de normalizarlas por el largo de cada transcripto (L). Para una muestra de ARN, si se secuencian 1 millón de transcriptos, el valor de TPM representa el número de transcriptos dado para un gen o isoforma (Zhao 2020).

$$TPM = 10^6 \times \frac{\mathrm{Ti/Li}}{(\mathrm{N1/L1} + \mathrm{N2/L2} + \ldots + \mathrm{Nn/Ln})} \tag{7}$$

6.8. Redes neuronales artificiales

Se entrenaron redes neuronales convolucionales con los datos transcriptómicos convertidos a imagen. Para lograr obtener una imagen se utilizaron los valores de los conteos de expresión normalizados (TPM) como el valor de los píxeles de la imagen. Los cromosomas ordenados como un cariotipo se distribuyeron uno tras otro desde el vértice superior izquierdo de la imagen hasta el vértice inferior derecho y se dividió la imagen en 1350 x 1350 píxeles (Figura 11). Para ello se dividieron a todos los cromosomas alineados en 1.822.500 partes a las cuales se les asignó una intensidad, entre 0 y 255 (intensidad de cada píxel), que corresponde a la expresión (normalizada al rango 0 a 255) en esa zona del genoma. Con esas imágenes etiquetadas se alimentó el entrenamiento de la red residual.

Para entrenar la red neuronal y validar los resultados se utilizó fast.ai (v2) un *frontend* de PyTorch y un GPU NVIDIA GeForce GTX 1080 Ti. Para disminuir la carga al GPU se utilizaron números en 16 bit en vez de 32.

Se probaron arquitecturas neuronales preestablecidas, incluyendo ResNet de 18, 34, 50 y 101 capas. Se eligió la ResNet50 entre las diferentes arquitecturas debido a su buen desempeño en todos los set de datos. Esta universalidad resultaría valiosa para agregar sencillez a una posible aplicación del diseño experimental en

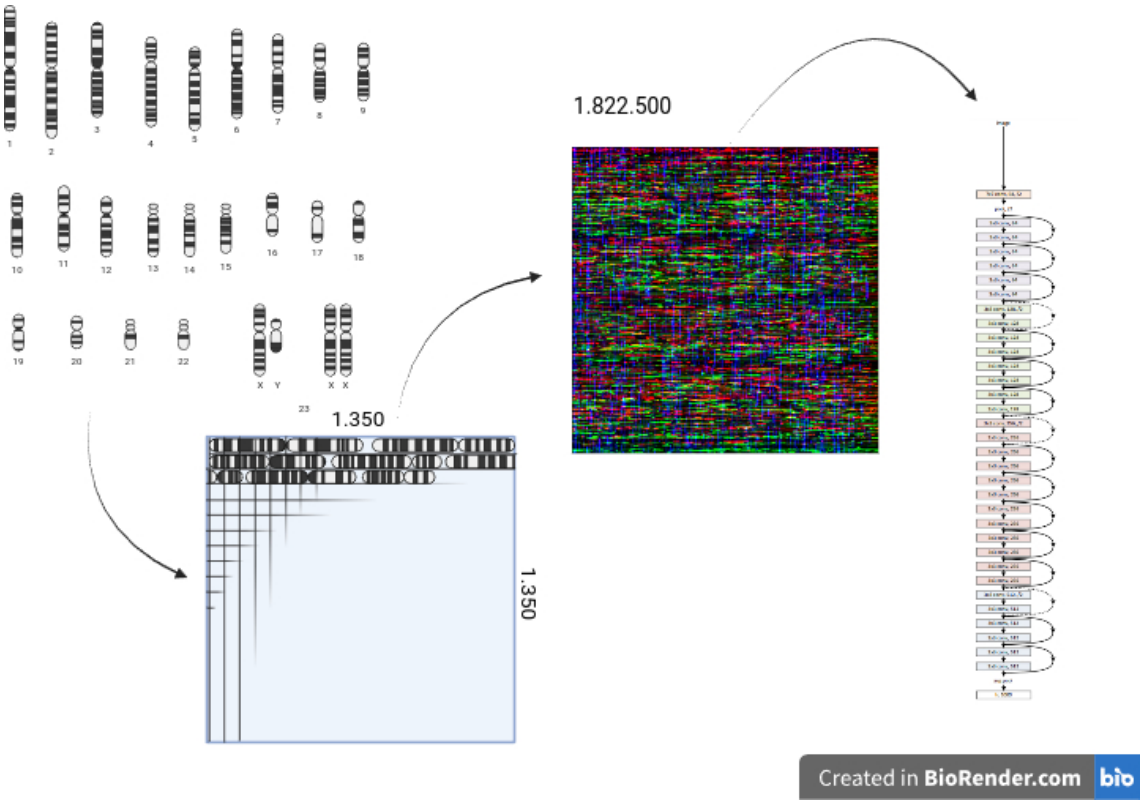


Figura 11: **Esquema de trabajo para la creación de la imagen utilizada en el entrenamiento de la Red Neuronal Artificial.** Desde el cariotipo hasta la ResNet pasando por una imagen real tomada de un participante anónimo.

la práctica.

Para cada entrenamiento se utilizó el 80 % de las imágenes, reservándose un 20 % para validar los parámetros encontrados. Con estas imágenes reservadas sin etiquetar se predijo el resultado de la clasificación y luego se lo comparó con el resultado conocido. Con estos resultados se elaboraron, en cada caso, las métricas para evaluar la performance del modelo: la exactitud de clasificación y la matriz de confusión. Se eligieron al azar 5 semillas para la elección de 5 grupos de imágenes diferentes de entrenamiento y validación dentro de cada grupo, pero entre los grupos de datos se mantuvieron las 5 semillas iguales. Los resultados se presentan como el promedio ± el desvío estándar. Se eligió siempre la corrida de la menor semilla para mostrar los resultados gráficos como representativos de todas las corridas.

Se probó de iniciar los entrenamientos con parámetros al azar en todas las capas y con los parámetros transferidos del entrenamiento de la red a partir de las imágenes de ImageNet y sólo la última capa con parámetros al azar (ImageNet es una base de datos de 14.197.122 imágenes etiquetadas y organizadas jerárqui-

camente). Se decidió iniciar el entrenamiento con el conocimiento transferido en todos los casos ya que los resultados fueron mínimamente mejores a los iniciados completamente al azar.

La tasa de aprendizaje utilizada en todos los casos fue entre $1x10^{-5}$ hasta $1x10^{-3}$. Se empezó a entrenar el primer tercio de los lotes de imágenes con una tasa de $1x10^{-5}$ y se comenzó a incrementar hasta que se llegó a la máxima $1x10^{-3}$ para los otros dos tercios se disminuyó hasta $1x10^{-5}$ nuevamente. Por este motivo, sólo se tomaron los parámetros de entrenamiento de las últimas épocas como modelo a ser validado.

Resultados

7.1. Secuenciación

Un total de 623 muestras alcanzaron los criterios de calidad en el conjunto de los 3 estudios clínicos realizados. Se obtuvieron 59171 genes con lecturas en, al menos, una de las muestras.

En la tabla 2 se presentan las lecturas obtenidas de la secuenciación de próxima generación (NGS) de los transcriptomas de los tres estudios clínicos realizados. El estudio de score de calcio coronario tiene la media más alta de lecturas por muestra, seguido por el estudio de sujetos sanos. El estudio de insuficiencia cardíaca presentó la mayor variabilidad en las lecturas y la menor cantidad de lecturas promedio.

En la figura 12 se observa la distribución de las lecturas por muestra para cada estudio clínico.

La base de datos GENCODE v46 cuenta con los datos de 63140 genes (feature = gene) anotados. En las 623 muestras se encontraron 51382 genes con alguna lectura en, al menos, un participante. En la primera columna de la 3 se observa la composición de los genes anotados en GENCODE v46.

El consorcio GTEx considera que un gen está expresado si tiene 5 o más lecturas. Este umbral reduce el ruido drásticamente mientras mantiene a la muestra

Estadística	**Sujetos sanos**	**Calcio coronario**	**Insuficiencia cardíaca**
Mediana	68.2 M	69.5 M	57.2 M
Media	67.9 M	69.6 M	58.2 M
Desviación estándar	18.9 M	19.8 M	20.1 M
Mínimo	18 M	34 M	17 M
Máximo	143 M	119 M	135 M
Total de muestras	334	196	108

Tabla 2: Resumen estadístico de las lecturas (M = millones de lecturas).

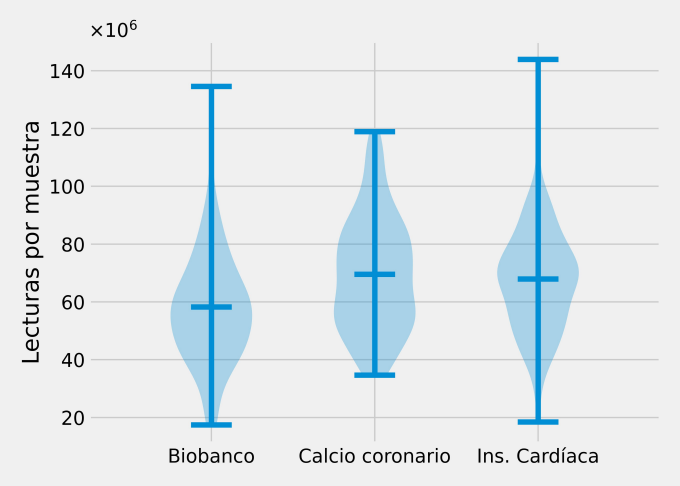


Figura 12: **Media de la cantidad de lecturas por muestra en los estudios clínicos realizados.** Gráfico de violín con la media de la cantidad de lecturas por muestra en los estudios clínicos realizados: Sujetos sin enfermedad aguda (Biobanco), score de calcio coronario e insuficiencia cardíaca.

Tipo de gen	Genes anotados (Total)	Hasta 1 lectura		Entre 1 y 5 lecturas		Más de 5 lecturas	
		Datos propios	GTEx	Datos propios	GTEx	Datos propios	GTEx
proteína	20089	17112	17167	1049	763	16063	16317
lncRNA	19258	15253	15143	3008	5468	12245	9364
Pseudogen	14481	11983	12905	4775	6651	7208	5514
miscRNA	2217	1910	1990	1013	1306	888	434
snRNA	1910	1603	1842	911	1134	692	181
miRNA	1879	1349	1528	656	937	693	90
snoRNA	942	649	789	323	406	326	145
Mitocondrial	24	19	24	5	5	14	19
Ribozima	8	5	5	3	3	2	2
Artificio	19	3	6	1	4	2	2
Otros	2313	1496	2209	389	759	1116	1243
Total de genes	63140	51382	53608	12133	17436	39249	33311

Tabla 3: Resultado de identificar en GENCODE v46 el tipo de gen hallado en las 623 muestras de todos los estudios propios y las 803 del consorcio Adult Genotype Tissue Expression (GTEx).

inclusiva (Melé 2015). Para analizar si los hallazgos de esta tesis coinciden con los criterios de expresión planteados por GTEx se analizaron cuántos y qué tipo de genes se encuentran en nuestros datos y en las muestras de tejido sanguíneo de GTEx. El consorcio contó con 803 muestras de sangre post mortem. Los 51382 genes identificados en nuestro estudio con, al menos, 1 lectura en alguna muestra corresponden al 81.38 % de los genes anotados, mientras que GTEx reportó 53608 genes (84.91 %). Cuando contamos los genes que tuvieron 5 lecturas o más en alguna de las muestras GTEx obtuvo un 60.01 % de los genes totales y en nuestros datos se alcanza un 70.70 %. Lo que señala que la cantidad de genes expresados en sangre es de alrededor del 85 % de los genes totales actualmente conocidos. Pero, teniendo en cuenta que una parte significativa de ellos sólo es detectada con entre 1 y 4 lecturas, una cifra de expresión que parecería biológicamente más relevante u "operativa" ronda más cercana al 70 %.

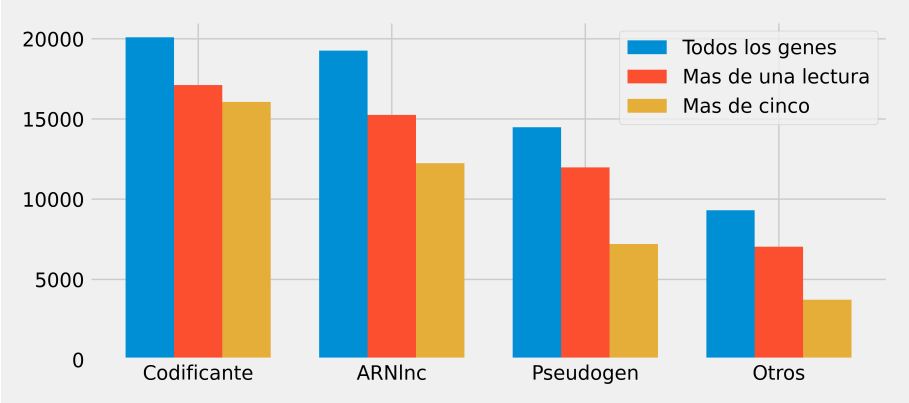


Figura 13: **Cantidad de genes encontrados en los 3 estudios clínicos.** Gráfico de barras mostrando la cantidad de genes encontrados en los 3 estudios clínicos distinguiendo entre Codificante, genes que codifican a proteína, ARNlnc, genes que resultan en ARN largo no codificante, Pseudogen (procesados o no) y la categoría Otros que resume a todos los otros ARNs encontrados como los no codificantes pequeños, los que están a la espera de ser confirmados experimentalmente, artificios y misceláneos. Los colores diferencian la población de genes utilizada para el análisis: Todos los genes anotados en el GENCODE v46 en azul. Los que sobrevivieron por tener al menos una lectura en alguna de las 623 muestras en rojo y, al menos, 5 lecturas en amarillo.

Respecto al tipo de gen expresado en ambos estudios se encontró una proporción similar de genes codificantes. Sin embargo, en el caso de los lncRNA nuestros datos alcanzaron un 80.28 % de genes frente al 61.85 % de GTEx. En la figura 13 se observa un gráfico de barras con los tipos de genes encontrados en los datos propios. Se hace evidente la disminución más abrupta en la clase pseudogen y otros en la población de genes que tienen más de 5 lecturas en alguna de las muestras (Barra amarilla) que en codificante y largo no codificante.

7.2. Experimentos de control

A medida que pasan los años se va produciendo un declive en las funciones fisiológicas de los tejidos de todo el cuerpo, resultando por sí mismo en un factor de riesgo para muchas enfermedades frecuentes (Savji 2013). El envejecimiento involucra una compleja red de vías metabólicas críticas en la respuesta homeostática al ambiente (Campisi 2019). El transcriptoma de una persona joven es esperable que sea diferente al de una persona adulta mayor (Peters 2015) y, por esta razón, se decidió incluir el análisis de la expresión diferencial entre estos grupos a esta tesis como un control; como un patrón que identifica cómo deberían verse los resultados en dos situaciones a sabiendas diferentes.

Con este mismo propósito se realizaron análisis poniendo en contraste a hombres y mujeres, dado que el dimorfismo sexual que presenta nuestra especie se debe en gran medida a la expresión diferencial de genes presentes en ambos sexos (Gershoni 2017). También por todo esto, los análisis estadísticos fueron controlados por sexo y edad.

7.2.1. Edad

Para evitar un umbral arbitrario que convierta la variable continua edad en una variable dicotómica y para encontrar diferencias de expresión génica más pronunciadas, se agrupó a los participantes de todos los estudios (pacientes y controles) por edad con un criterio biológico en tres grupos y se eliminó al grupo de edad intermedia.

- Jóvenes (19 años hasta 35)
- Adultos jóvenes (entre 36 y 64 años)
- Adultos mayores (65 años hasta 89)

Al eliminar del análisis a los adultos jóvenes se obtuvieron un total de 202 muestras, 99 pertenecientes a menores de 35 (media de 29±4 años) y 103 mayores de 65 años (media de 71±4 años). Se contabilizaron 103 participantes femeninas y 99 masculinos.

7.2.1.1. Expresión diferencial de genes

Se encontraron 12836 genes diferencialmente expresados por edad con un p ajustado menor a 0.01. De estos, la mayoría tiene un $\log_2$(VC) (veces de cambio) menor a 1 (1996 genes tienen un $\log_2$(VC) mayor a 1). Con una amplitud desde -5.62 hasta 4.64 $\log_2$(VC). En la figura 14 se observa el mapa de calor para los genes diferencialmente expresados en las filas y los participantes en las columnas. La barra verde señala el grupo joven y la barra coral el grupo adulto mayor. En el cuadrante superior izquierdo se aprecian los genes sobreexpresados en los sujetos jóvenes, mientras que estos mismos genes se encuentran subexpresados en los sujetos adultos mayores (cuadrante superior derecho). Lo opuesto sucede en los cuadrantes inferiores. Como es esperable, el cambio general en la expresión génica es de muchos genes con baja sobre o subexpresión.

7.2.1.2. Agrupamientos no supervisados

Se estudiaron los más de 12 mil genes resultantes del DEseq2 mediante análisis de componentes principales (Figura 15, izq.) y se observó un ordenamiento según la edad en el primer componente principal, el cual explica el 22 % de la varianza. En naranja se observa el grupo de menor edad y en azul el de mayor edad en ambos modelos. Aparece un subgrupo de adultos mayores que se separa de la nube mayor que son los mismos sujetos que se separan en UMAP (Figura 15, der.), sin embargo, no se encontró otra variable medida como responsable de ese pequeño agrupamiento. Se probaron sexo, índice de masa corporal y tabaquismo. En ambos métodos de agrupamiento no supervisados se llega a un ordenamiento por edad sin una verdadera clusterización.

7.2.1.3. Red neuronal

Se entrenó 5 veces independientes una red neuronal residual de 50 capas con 161 imágenes y se retiraron 40 para la validación. En el proceso de entrenamiento se le mostraron a la red las 161 imágenes 30 veces (en lotes de 4 imágenes por cuestiones de capacidad de memoria), configurando 30 épocas. En cada época se calcula el error de entrenamiento y, con las imágenes de validación, el de validación. Con estos datos se elaboró un gráfico (figura 16, izquierda) en el que se puede apreciar la mejora en cada iteración: con los nuevos valores de los parámetros el error o pérdida calculado es cada vez más cercano a cero.

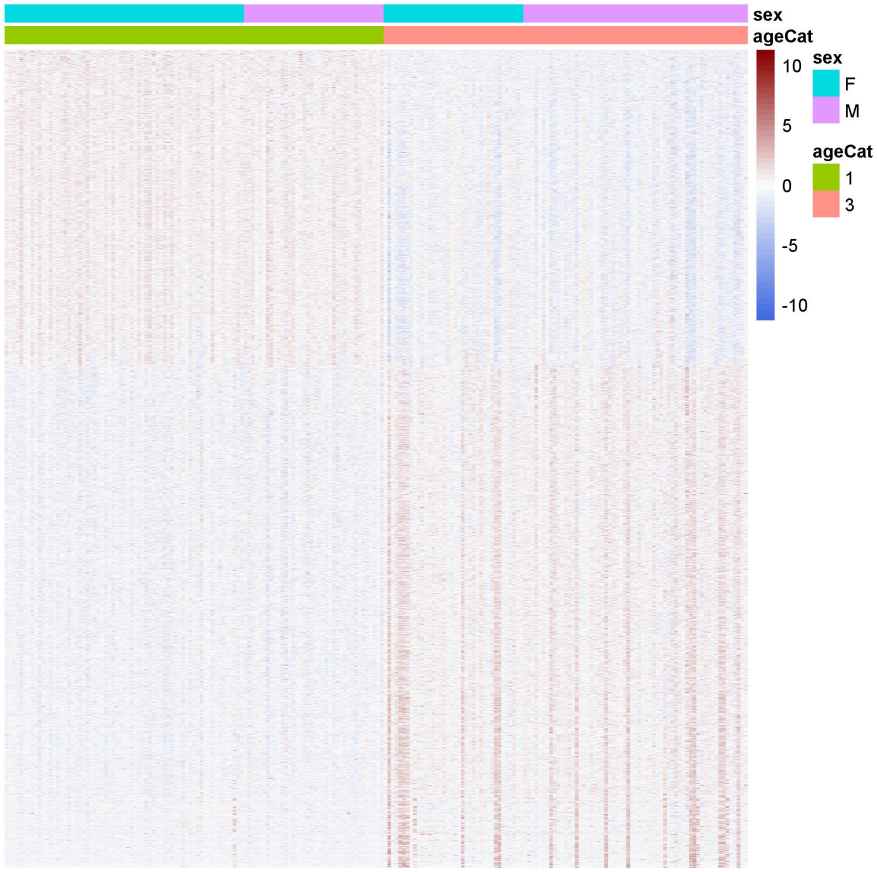


Figura 14: **Mapa de calor de la expresión diferencial según edad.** Mapa de calor de la expresión diferencial de genes entre jóvenes (ageCat 1: personas entre 19 y 35 años) y adultos mayores (ageCat 3: participantes entre 65 y 89 años). Las participantes del sexo femenino (F) se marcan en esmeralda y las columnas de sujetos de sexo masculino (M) en rosa. Se muestran los 12836 genes con p ajustado menor a 0.01 en las filas y los 201 participantes en las columnas. Expresión aumentada en rojo, expresión disminuida en azul.

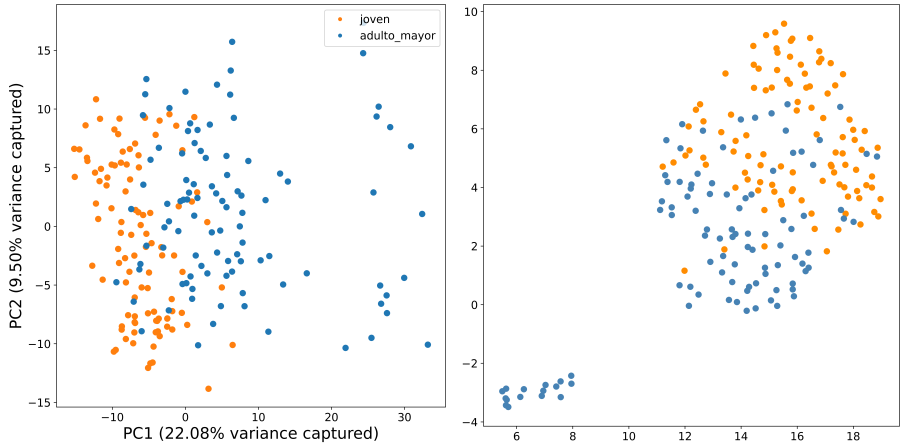


Figura 15: **Análisis de agrupamientos no supervisados según edad.** Izq: Gráfico de análisis de los componentes principales (PCA) para los participantes de todos los estudios clínicos con los genes diferencialmente expresados entre jóvenes (19 a 35 años) y adultos mayores (65 a 89 años). Der: Uniform Manifold Approximation and Projection (UMAP) distancia mínima = 0.7, número de vecinos = 10. En ambos gráficos se diferencian los jóvenes en naranja y los adultos mayores en azul.

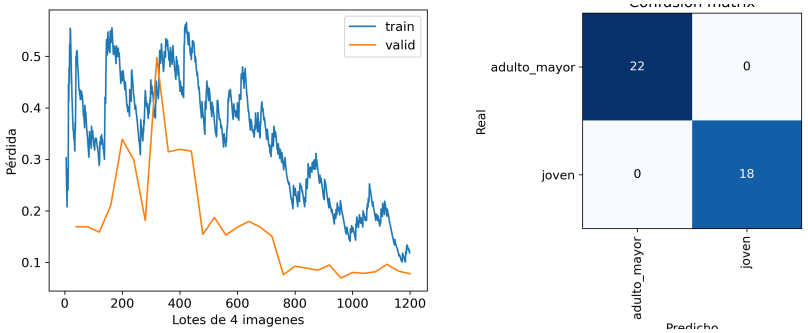


Figura 16: **Análisis de redes neuronales según edad.** Gráfico de la función de pérdida o error por lote de imágenes para 30 épocas en el entrenamiento de edad para todos los estudios. Der.: Matriz de confusión resultado de una de las validaciones.

El promedio de la exactitud de clasificación de las 5 validaciones (classification accuracy) fue de 0.980 ($\pm$0.011). En la figura 16, a la derecha, se observa la matriz de confusión para una de las corridas. En el caso de esta corrida de ejemplo, las predicciones de las 40 imágenes coincidieron con la verdadera etiqueta de la imagen. La red fue capaz de clasificar correctamente a los 22 adultos mayores y a los 18 jóvenes. En general, en todos los set de validación presentados la red se equivocó en la clasificación de 1 o, en un caso, 2 participantes. Logrando una excelente performance en la clasificación.

7.2.2. Sexo

Para agrupar a los participantes por sexo se unieron todos los estudios resultando en un total de 623 muestras de las cuales 310 fueron masculinas y 313 femeninas, con un promedio de edad de 52$\pm$14 y 48$\pm$14 años respectivamente.

7.2.2.1. Expresión diferencial de genes

Una vez realizado el modelo lineal generalizado se obtuvieron unos 4996 genes que se expresaron diferente entre hombres y mujeres con una significancia ajustada de 0.01. Dentro de estos, unos 186 genes tuvieron una expresión mayor (o menor) a 2 veces el grupo de comparación (un $\log_2$(VC) mayor a 1 (o menor a -1)).

En el mapa de calor presentado en la figura 17 se pueden observar los genes mayormente expresados en mujeres en el cuadrante superior izquierdo y los mayormente expresados en hombres en el cuadrante inferior derecho. En este caso, hay más participantes que en el caso de la edad, por lo que cada cuadrado representando la expresión de un gen de un participante es más pequeño y la visibilización es menos evidente. Además, hay proporcionalmente menor cantidad de genes con una diferencia de $\log_2$(VC) mayor a 1, por lo que no se encuentran muchos cuadrados con color rojo o azul intenso correspondientes a diferencias de expresión más intensa, resultando en un mapa de calor más pálido que en el caso de la edad.

7.2.2.2. Agrupamientos no supervisados

En el caso de los agrupamientos no supervisados (Figura 18) se observa un agrupamiento claro tanto para el análisis de componentes principales (izq.) como

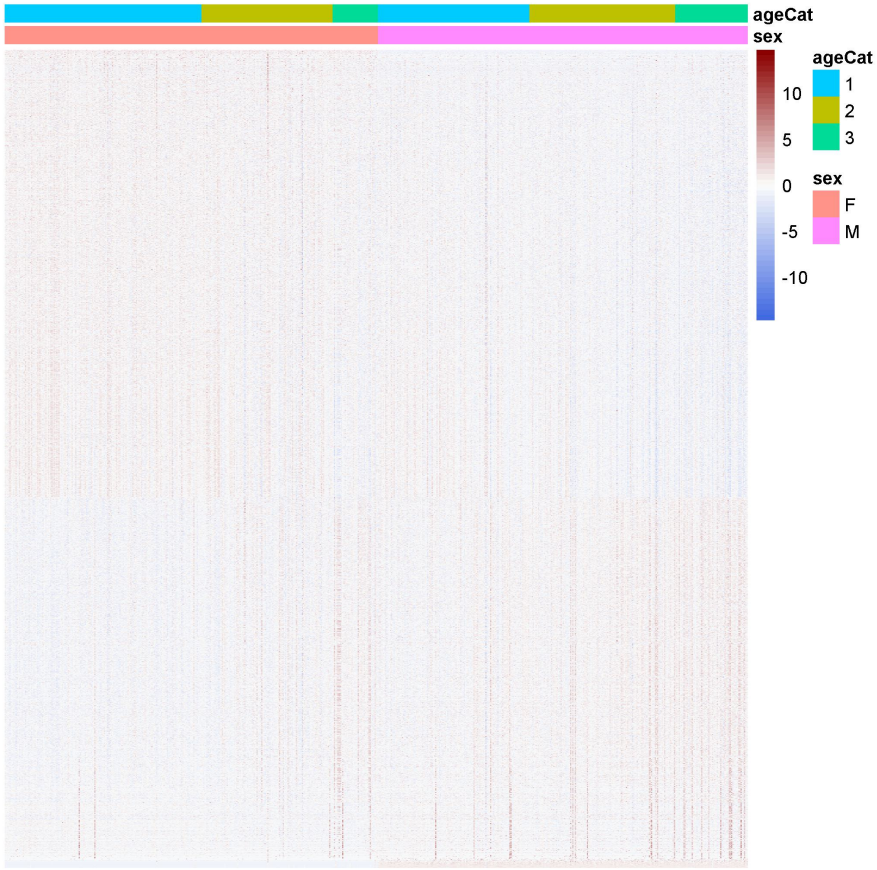


Figura 17: **Mapa de calor de la expresión diferencial de genes entre sexos.** Mapa de calor de la expresión diferencial de genes entre participantes femeninas (barra coral) y masculinos (barra magenta). Los sujetos con menos de 50 años se distinguen en cian (1), los mayores o iguales a 50 años y menores o iguales a 65 años en mostaza (2) y los sujetos mayores de 65 años en verde (3). Se muestran los 4996 genes con p ajustado menor a 0.01 y $\log_2$(VC) <-1 y >1 en las filas y los 623 participantes en las columnas. Expresión aumentada en rojo, expresión disminuida en azul.

para UMAP (der.). En PCA una combinación de los componentes principales 1 y 2 logran agrupar a hombres (en naranja) y mujeres (en azul) por separado. En ambos casos se observan unos pocos casos mal clasificados. En el caso de PCA, 4 mujeres fueron clasificadas como sujetos masculinos, pero ningún hombre fue mal clasificado; en el de UMAP unas decenas de hombres fueron clasificados como mujeres y unas 3 mujeres como hombres. Entonces, la representación global (la varianza) parecería representar mejor las diferencias entre hombres y mujeres.

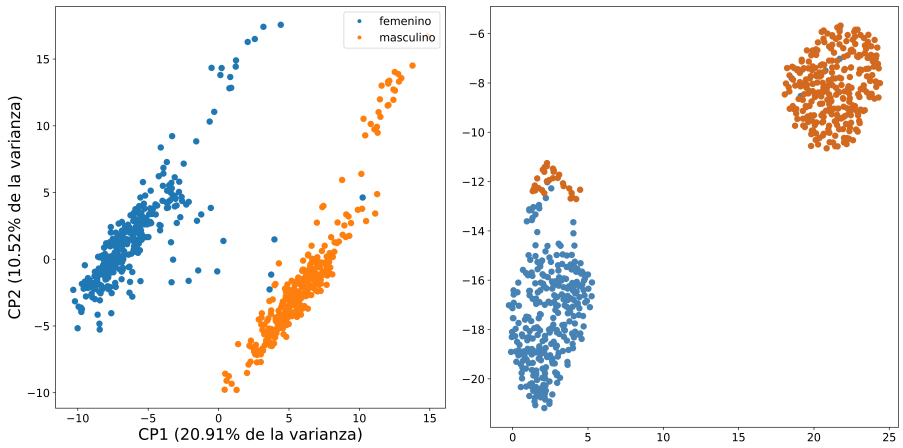


Figura 18: **Gráficos de análisis no supervisados aplicados a sexo.** Gráfico de análisis de los componentes principales (PCA) para participantes femeninas y masculinos de todos los estudios; CP: componente principal. Der: Uniform Manifold Approximation and Projection (UMAP) distancia mínima = 0.7, número de vecinos = 50, épocas = 500. En ambos gráficos se diferencian los sujetos masculinos en naranja y los femeninos en azul.

7.2.2.3. Red neuronal

Se utilizaron 499 imágenes para entrenar a la ResNet50 y se reservaron 124 para validar, obteniéndose una exactitud de clasificación promedio de 0.990 (±0.007). En un 99 % de los casos la red clasificó correctamente. Como se observa en la matriz de confusión de una de las corridas, a la derecha de la figura 19, sólo 2 mujeres fueron predichas como sujetos masculinos. Como en el caso de la edad, en general hubo muy pocas equivocaciones, no superando los 2 sujetos mal clasificados, superando aún a PCA y sin filtrado previo de genes.

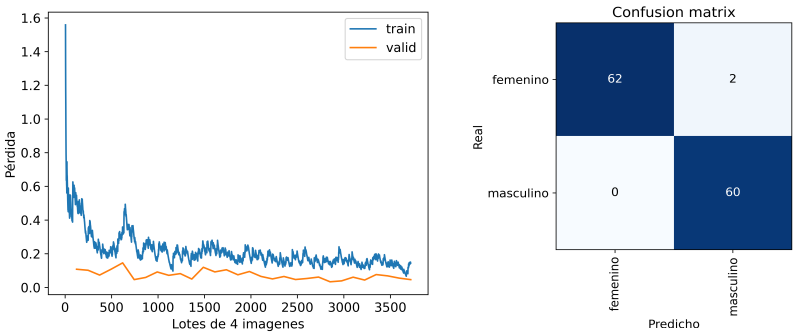


Figura 19: **Análisis por redes neuronales aplicados a sexo.** Izq.: Gráfico de la función de pérdida o error por lote de 4 imágenes para 30 épocas en el entrenamiento para clasificar por sexo para todos los estudios. Der.: Matriz de confusión resultado de una de las validaciones para las imágenes correspondientes a los 60 masculinos y 64 femeninos.

7.3. Factores de Riesgo

7.3.1. Obesidad

Con el dato de la altura y el peso se calculó el índice de masa corporal para todos los participantes. Para dicotomizar la variable IMC se utilizaron los valores más extremos de la distribución, por lo que se retiró a los participantes con peso intermedio. Los que obtuvieron un índice entre 25 y 30 inclusive fueron clasificados como personas con sobrepeso (ver tabla 1) y fueron eliminados del análisis. Se tomaron como control las personas con un índice menor a 25 y como obesos a quienes obtuvieron un índice mayor a 30. En total, 162 sujetos presentaron un IMC mayor a 30, con un promedio de edad de 53 ± 12 años y 227 controles con un promedio de edad de 46 ± 14 años. Se contabilizaron 208 mujeres (edad promedio 47 ± 13 años) y 181 hombres (edad promedio 51 ± 14 años).

7.3.1.1. Expresión diferencial de genes

Al realizar el análisis en el DEseq2 se le asignó al modelo lineal generalizado como covariable a los sujetos con insuficiencia cardíaca, debido a la paradoja de la obesidad en esta enfermedad. Se obtuvieron 570 genes diferencialmente expresados que se pueden observar en el mapa de calor de la figura 20. En la parte izquierda de la figura se encuentran los controles señalados con la barra lila y en la zona derecha los obesos con los genes menos expresados que los controles en azul en el cuadrante superior y los más expresados en rojo en el cuadrante inferior.

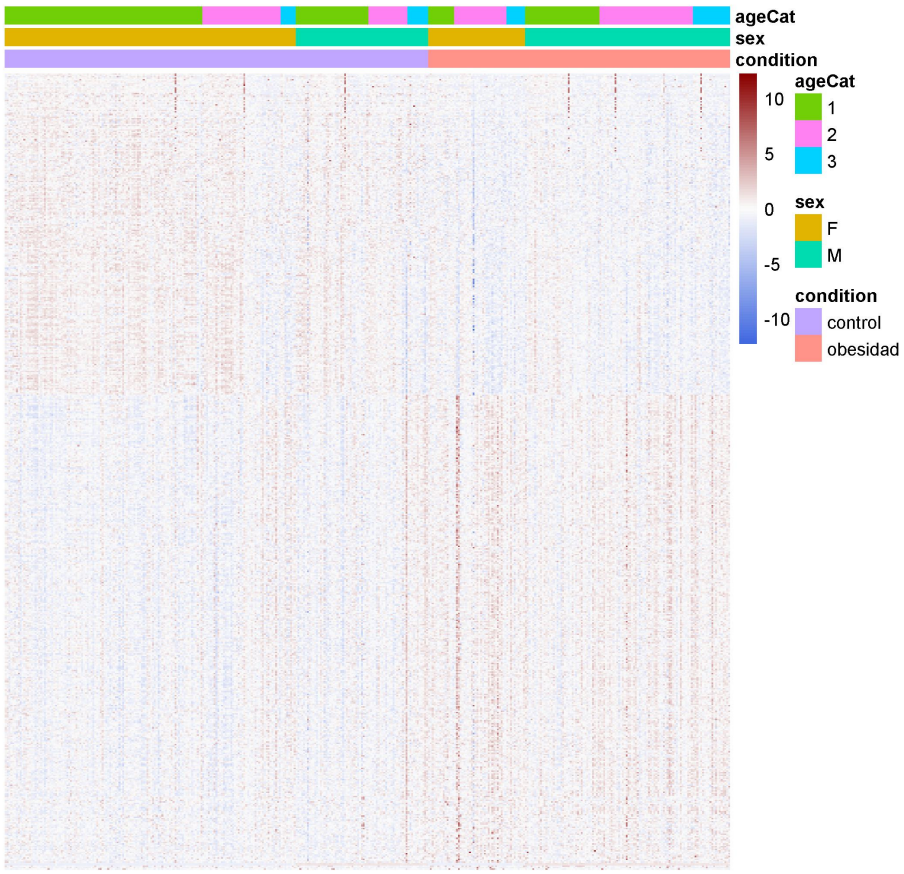


Figura 20: **Mapa de calor de la expresión diferencial de genes entre obesos.** Mapa de calor de la expresión diferencial de genes entre obesos (barra coral) y controles (barra lila). Las participantes del sexo femenino (F) se marcan en mostaza y las columnas de sujetos de sexo masculino (M) en esmeralda. Los sujetos <50 años se distinguen en verde (1), 50 ≥sujetos ≤65 en rosa (2) y los sujetos >65 en cian (3). Se muestran los 570 genes con p ajustado menor a 0.01 y $\log_2$(VC) <-1 y >1 en las filas y los 389 participantes en las columnas. Expresión aumentada en rojo, expresión disminuida en azul.

Con intención de tener una aproximación al significado biológico de los genes hallados se utilizó la herramienta de ontología génica GO enrichment analysis (Mi 2013). El Consorcio de GO provee una representación computacional del actual conocimiento científico sobre las proteínas y los ARNs no codificantes para mejorar la comprensión de cómo los genes individualmente contribuyen a la biología de un organismo desde el nivel molecular pasando por las vías metabólicas, hasta el nivel de organización celular o de organismo.

Al analizar el conjunto de genes diferencialmente expresados entre obesos y controles se encuentra el término "respuesta ante nutrientes", del cual derivan a su vez términos hijos como: regulación del apetito, respuesta a la restricción calórica, utilización de carbohidratos, entre otros en el mismo sentido. Este término describe el cambio de expresión génica como resultado de un estímulo que refleja la presencia, ausencia o concentración de nutrientes. Un término pertinente para una condición que, en la mayoría de los casos, se debe al consumo en exceso de nutrientes en la dieta. Algunos ejemplos dentro de los 11 genes encontrados para este término se encuentran: la subunidad catalítica de la glutamato-cisteína ligasa (GCLC), la ATPasa transportadora de calcio de membrana 1 (ATP2B1), el receptor coactivador nuclear 1 (NCOA1), la lipoproteín lipasa (LPL), el miembro de la familia de cadena larga de la acetil-CoA sintetasa 1 (ACSL1) y la endopeptidasa fosforregulada ligada al X (PHEX).

En la figura 21 se observan procesos biológicos que dieron resultados significativos. La mayoría relacionados a procesos de la sangre, como la coagulación, diferenciación de progenitores en células sanguíneas, procesos relacionados a la hemoglobina, como también otros relacionados a la respuesta inmune proinflamatoria como la cascada del complemento y la vía de señalización del receptor tipo toll.

Posteriormente realizamos análisis de ontología génica. Al utilizar el método de enriquecimiento de un grupo de genes para el análisis de vías biológicas GAGE (generally applicable gene set enrichment) se agregan procesos celulares de vías energéticas como la organización mitocondrial y la cadena de electrones en respiración celular o la generación de metabolitos precursores y energía, y procesos catabólicos de ácidos carboxílicos y de aminoácidos. Además de coincidir con las vías de respuesta inmune e inflamación ya mencionadas por el GO.

En la figura 22 se muestra en detalle la vía de señalización de PPAR (Peroxisome proliferator-activated receptor), un receptor hormonal nuclear que se activa

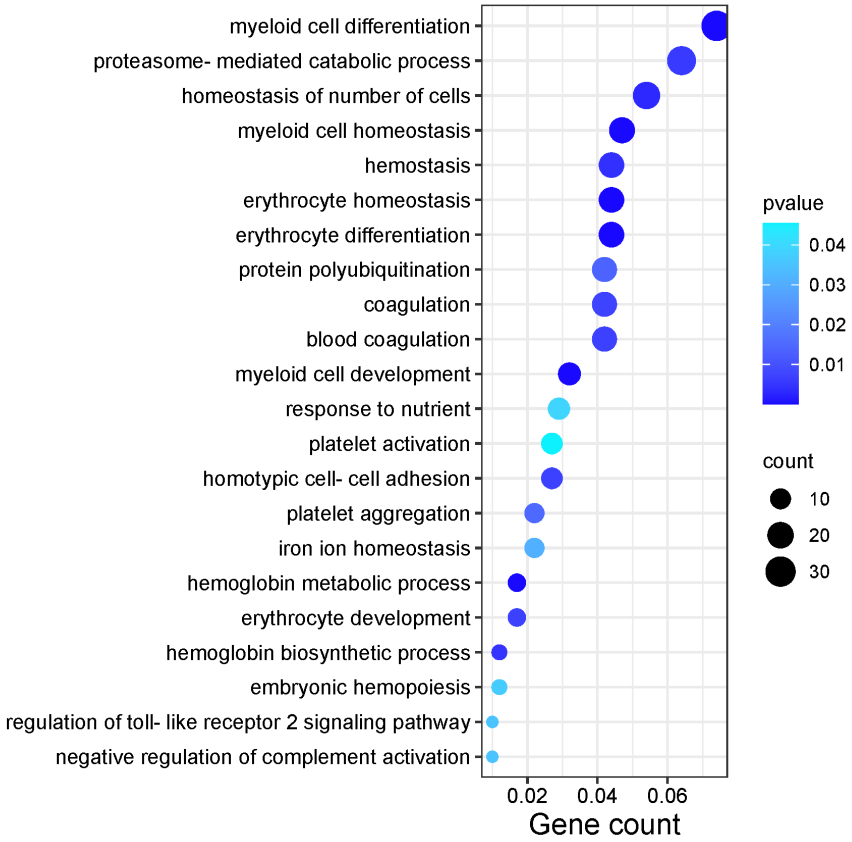


Figura 21: **Gráfico de burbuja para el análisis de ontología génica en obesidad.** Cada fila en el eje y describe un proceso biológico descrito en la ontología génica. En el eje x se expresa el ratio génico, qué proporción de los genes señalados en el proceso biológico se encuentran en la lista de genes diferencialmente expresados según el DEseq2. El tamaño de la burbuja señala la cantidad de genes involucrados en el proceso. El color indica la significancia, p valor ajustado hasta 0.05.

por ácidos grasos y sus derivados, como un ejemplo de vía energética diferencialmente expresada. En la imagen resalta en verde la lipoproteinlipasa, lo que significa que en el grupo de personas obesas se encuentra disminuida su expresión respecto al grupo control. La LPL es una enzima clave en el metabolismo lipídico, encargada de hidrolizar a los triglicéridos de los quilomicrones y VLDL, por lo tanto, su disminución genera un aumento de triglicéridos en la sangre y los tejidos (Balasubramanian 2024). Como se comentó anteriormente, los PPARs son los reguladores centrales del metabolismo de lípidos del corazón. Entonces, no resulta sorprendente que se encuentre desregulada esta vía del metabolismo energético en uno de los factores de riesgo de enfermedad cardiovascular en el cual el exceso de energía consumido en la dieta es el responsable.

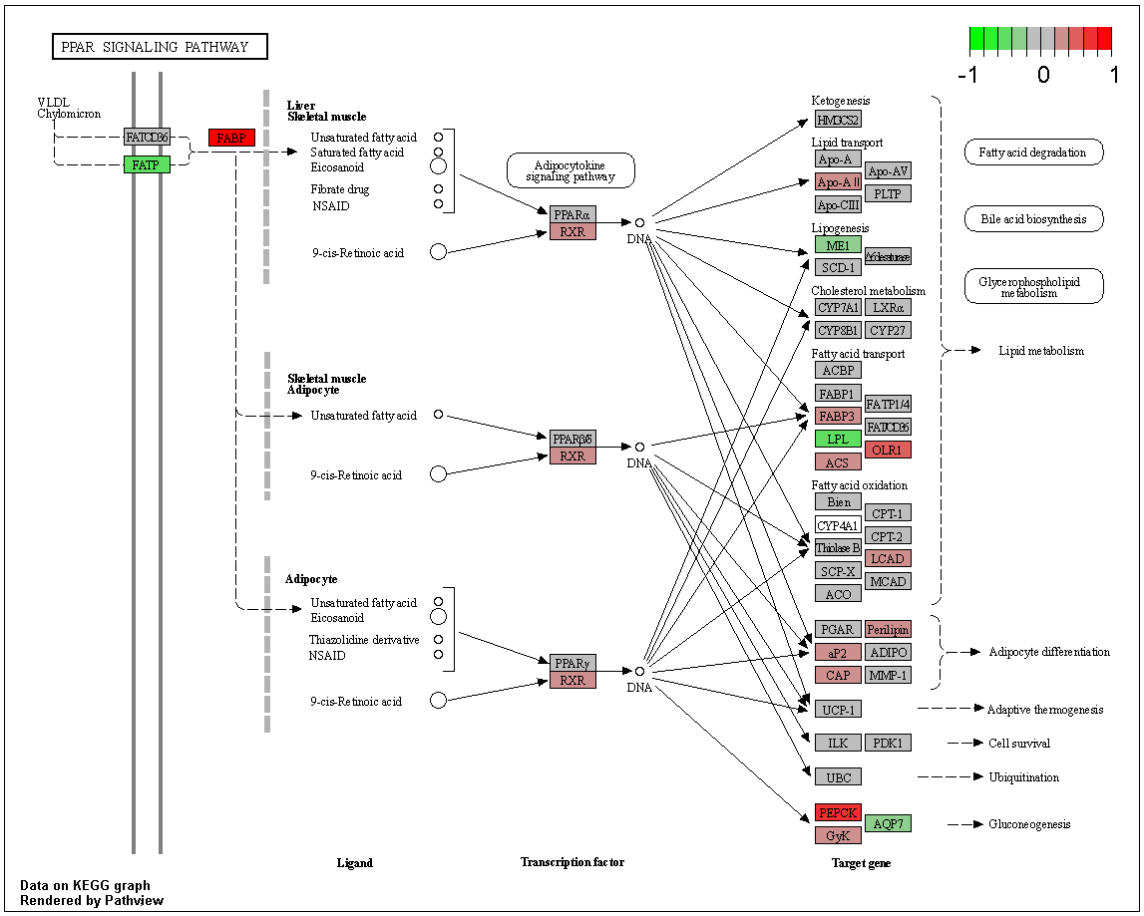


Figura 22: **Esquema de la vía de señalización de PPAR en obesos** Esquema de la vía de señalización de PPAR (Peroxisome proliferator-activated receptor) de KEGG (Kyoto Encyclopedia of Genes and Genomes) en el análisis de obesos vs controles. En verde genes menos expresados en el grupo obeso que en el grupo control. Gris: sin diferencias significativas. Rojo: mayor expresión que el grupo control.

En general, las vías y los procesos que mostraron diferencias significativas muestran un metabolismo energético e inflamatorio diferenciado en la expresión

génica entre personas obesas y controles.

7.3.1.2. Agrupamientos no supervisados

En el caso del análisis de componentes principales se observan dos agrupamientos (Figura 23, izquierda) de los cuales el primer componente explica el 18.09 % de la varianza y el segundo explica el 10.43 %. Aunque no es el índice de masa corporal la condición que estaría explicando los agrupamientos, sino el sexo. A pesar de haber sido elegidos solo los transcriptos diferencialmente expresados entre obesos y controles, la señal del sexo es más fuerte para agrupar a los y las participantes. Siendo el grupo superior izquierdo el femenino y el inferior derecho el masculino.

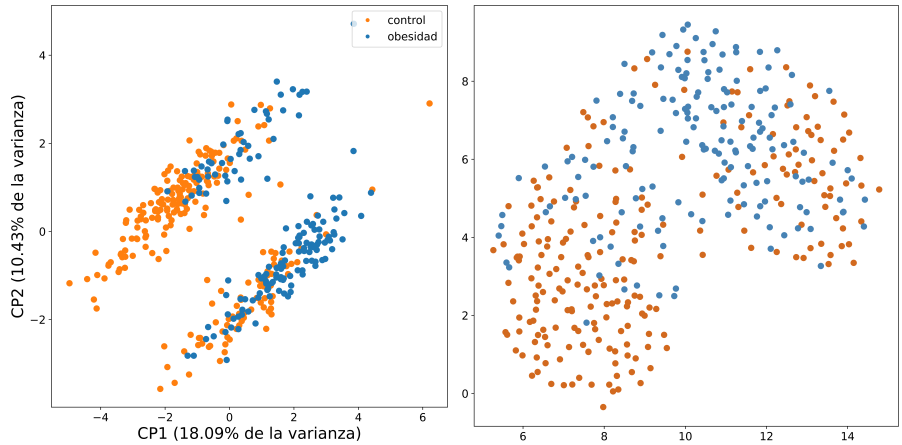


Figura 23: **Análisis de los componentes principales (PCA) en obesos.** Izq: Gráfico de análisis de los componentes principales (PCA) entre obesos y controles; CP: componente principal. Der: Uniform Manifold Approximation and Projection (UMAP): distancia mínima = 0.7, número de vecinos = 50, épocas = 500. En ambos gráficos se marcan los controles en naranja y los obesos en azul.

En el caso de UMAP (Figura 23, derecha) aparecen dos *clusters* no muy bien definidos, con una mayoría de controles (naranja) en uno y una mayoría de obesos (azul) en el otro, aunque con muchas muestras mal clasificadas por la condición. Al identificar las muestras por sexo se da el mismo caso que en PCA, los dos agrupamientos aparecen bien definidos, casi sin equivocaciones; uno con muestras femeninas y otro con muestras masculinas. También la señal de la condición sexual es más detectada en los agrupamientos no supervisados que la señal captada por el índice de masa corporal en este caso, aún recortando las variables por los genes diferencialmente expresados para esta condición.

7.3.1.3. Red neuronal

Con 308 imágenes, correspondientes a la expresión de todos los genes y etiquetadas diferencialmente en obesos y controles, se entrenó una ResNet50 durante 30 épocas obteniéndose una exactitud de clasificación de 0.793 (±0.033). Como puede observarse en la matriz de confusión de ejemplo de la figura 24, a la derecha, en aproximadamente un 80 % de los casos la red logra clasificar a las personas con índice de masa corporal control de los obesos con la información de los 49840 genes totales. A pesar de lo imperfecto de este índice se logra alcanzar una buena diferenciación génica entre los índices extremos (excluyendo los participantes con sobrepeso), captando la señal que los métodos de clusterización no lograron captar.

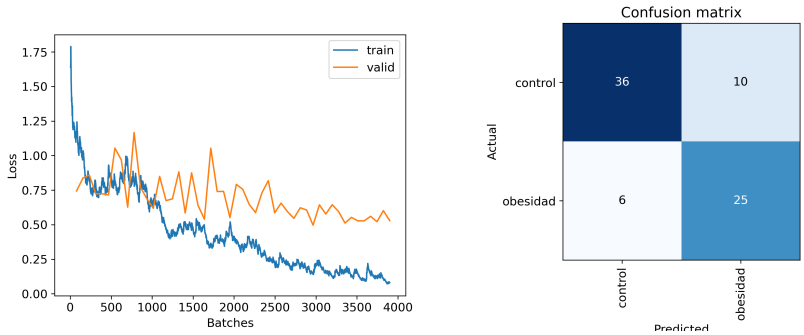


Figura 24: **Análisis por redes neuronales aplicados a obesidad.** Izq.: Gráfico de la función de pérdida o error por lote de 4 imágenes para 50 épocas en el entrenamiento para clasificar entre obesidad y control para todos los estudios. Der.: Matriz de confusión resultado de una de las validaciones de ResNet50 para las 77 imágenes correspondientes para obesos y controles.

A la izquierda de la figura 24 se detalla un entrenamiento de 50 épocas con las imágenes de la expresión génica de obesos y controles. En este caso se puede ver que la pérdida de entrenamiento continúa disminuyendo, mientras que la de validación se mantiene sin poder ser menor que 0.50, lo que podría deberse a un sobreajuste. Por este motivo se decidió hacer los entrenamientos hasta 30 épocas, donde, en general, no se observó este comportamiento.

Variable clínica	Media (SD)	Mediana	Mínimo	Máximo
Hemoglobina (g/dl)	14.0 (1.4)	13.9	8.3	18.2
HbA1C (%)	5.3 (0.7)	5.2	4.3	14.2
Glucosa (mg/dl)	93 (16)	90	71	204
Colesterol Total (mg/dl)	196 (41)	194	92	353
Colesterol HDL (mg/dl)	56 (15)	55	22	110
Colesterol LDL (mg/dl)	118 (36)	113	42	260
Triglicéridos (mg/dl)	115 (76)	96	33	648
Proteína C reactiva (mg/l)	2.3 (7.1)	1.3	0	122.7
Leucocitos (cél.$\times$mil/mm^3)	6.8 (1.6)	6.6	3.5	14.4
Creatinina en sangre (mg/dl)	0.75 (0.18)	0.74	0.39	1.76
Bilirrubina Total (mg/dl)	0.62 (0.30)	0.55	0.23	1.94
GOT-ASAT (UI/l)	20 (8)	18	0.2	96
GPT-ALAT (UI/l)	22 (16)	17	7	171
Presión Sistólica (mmHg)	116 (17)	113	78	172
Presión Diastólica (mmHg)	70 (10)	70	42	104
Índice de masa corporal	26 (5)	25	17	47

Tabla 4: Estadísticos muestrales de las variables clínicas obtenidas en el estudio de sujetos sin enfermedad aguda.

7.4. Resultados del estudio clínico de sujetos sin enfermedad aguda

Al cierre del estudio se alcanzó un total de 337 participantes, 211 mujeres y 126 hombres. Las características clínicas se describen en la tabla 4. La distribución etaria se desvía hacia los sujetos más jóvenes con un promedio de edad de 43 ($\pm$13) años (42 ($\pm$12) para las mujeres y 44 ($\pm$13) para los hombres). La amplia mayoría de los participantes (283) no padecen hipertensión arterial, 53 tienen hipertensión o están bajo tratamiento y sobre un participante no se obtuvieron datos al respecto.

De acuerdo a los criterios de inclusión, los sujetos de este estudio no presentaban patologías agudas en curso, pero una parte de ellos presentaban enfermedades metabólicas crónicas, o estados preclínicos de las mismas. Se analizaron entonces diferentes fenotipos de acuerdo al grado de afectación. Para incrementar las dife-

rencias esperables de encontrar, se dividieron los grupos en tercios y se realizaron las comparaciones entre los grupos extremos. Se analizaron tres fenotipos utilizando las siguientes variables:

1. Porcentaje de hemoglobina glicosilada, como marcador de diabetes.
2. Colesterol total, como marcador de dislipemia.
3. Proteína C reactiva, como marcador de inflamación.

7.4.1. Prediabetes

Se retiraron del análisis a todos los participantes del estudio diagnosticados con diabetes y/o con valores superiores a 6.4% (límite superior para un valor normal) de hemoglobina glicosilada y/o utilizaran insulina o metformina como variables subrogantes del diagnóstico de diabetes. Se dividió a los participantes restantes entre:

- Hemoglobina glicosilada baja, con valores inferiores o iguales a 4.9%.
- Media, con valores superiores a 4.9% y menores a 5.7%.
- Prediabéticos con valores mayores o iguales a 5.7% y menores o iguales a 6.4%.

Para dicotomizar la variable continua se retiraron del análisis las muestras con hemoglobina glicosilada media, quedando sólo los valores más extremos de la distribución. Resultaron un total de 52 participantes de los cuales 23 son prediabéticos y 29 tienen hemoglobina glicosilada baja. 23 participantes eran de sexo femenino con un promedio de edad de 41 ($\pm$15) años y 29 al sexo masculino con un promedio de edad de 41 ($\pm$14) años.

7.4.1.1. Expresión diferencial de genes

Al utilizar el modelo lineal generalizado del DEseq2 con sexo y edad como covariables se encontraron 185 genes diferencialmente expresados entre participantes con porcentaje de hemoglobina glicosilada baja y alta. De estos, sólo 29 tuvieron un $\log_2$(VC) mayor a 1. En la figura 25 se muestran los genes en un mapa de calor identificando a los participantes prediabéticos con la barra coral a la derecha y a los sujetos con hemoglobina glicosilada baja con la barra violeta a la

izquierda de la imagen. Se notan una mayoría de genes mayormente expresados al tener la hemoglobina glicosilada baja.

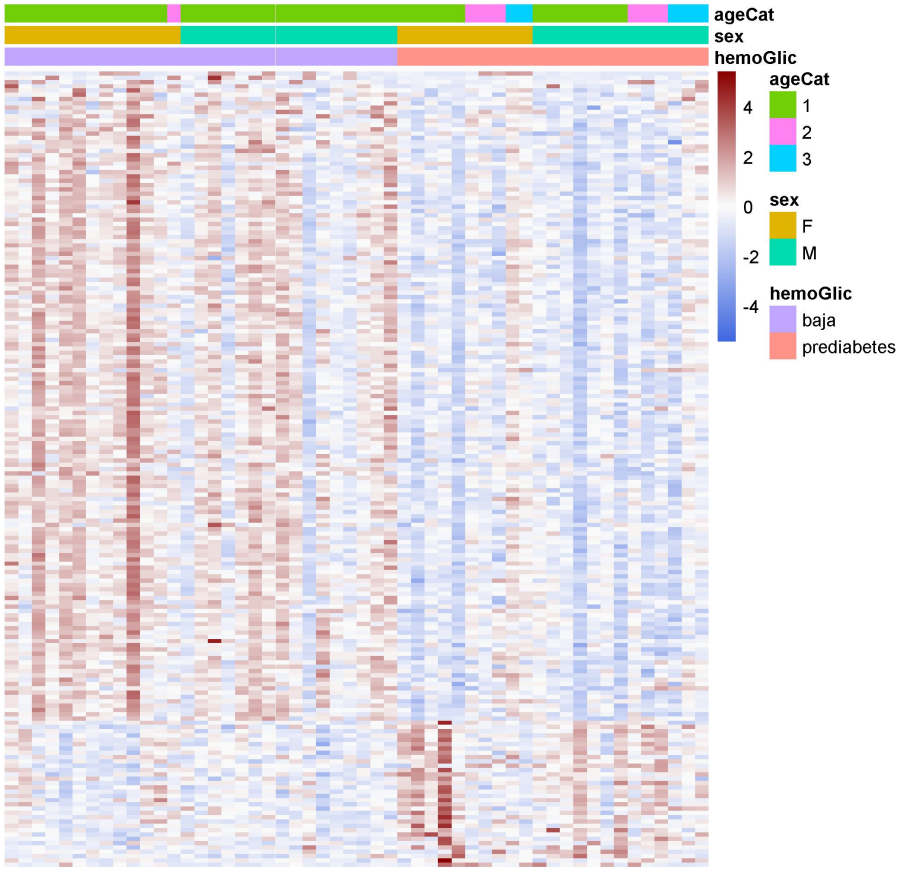


Figura 25: **Mapa de calor de la expresión diferencial de genes entre prediabéticos.** Mapa de calor de la expresión diferencial de genes entre prediabéticos (barra coral) y hemoglobina glicosilada baja (barra lila). Las participantes del sexo femenino (F) se marcan en mostaza y las columnas de sujetos de sexo masculino (M) en esmeralda. Los sujetos <50 años se distinguen en verde (1), 50 >= sujetos <= 65 en rosa (2) y los sujetos >65 en cian (3). Se muestran los 185 genes con p ajustado menor a 0.05 en las filas y los 52 participantes en las columnas. Expresión aumentada en rojo, expresión disminuida en azul.

Al analizar el grupo de genes diferencialmente expresados, el GO (Figura 26) los ubicó mayormente en procesos biológicos de división celular mitótica y en procesos inmunes. La alteración del ciclo celular, vital en la homeostasis del organismo, podría ser una respuesta al estrés metabólico inducido por la prediabetes. Así como la modulación de la respuesta inmune es relevante ante la inflamación crónica de bajo grado característica de esta condición.

En el análisis realizado mediante GAGE los procesos biológicos hallados en la ontología génica coinciden en encontrar significativas las vías de respuesta inmune y agregan las vías de “Respuesta inflamatoria”, “Coagulación”, “Hemostasis”

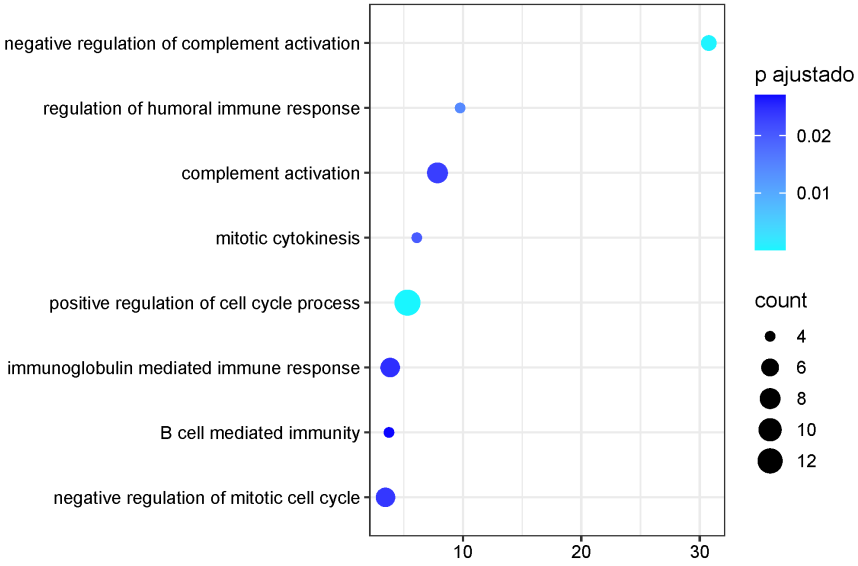


Figura 26: **Análisis de ontología génica en pacientes prediabéticos.** Gráfico de burbuja del análisis de ontología génica para los genes diferencialmente expresados de los participantes prediabéticos y controles del estudio de sujetos sin enfermedad activa. En el eje x se nota el enriquecimiento como el porcentaje de genes encontrado en el análisis de los genes totales de la vía.

y “Regulación de procesos catabólicos”. El mismo análisis con la Enciclopedia de Genes y Genomas de Kioto (KEGG) coincidió con las vías involucradas en procesos inflamatorios, inmunes y agregó la vía de la insulina.

En la figura 27 se detallan los genes diferencialmente expresados en la vía de la insulina donde aparece SOCS (suppressor of cytokine signaling proteins) sobreexpresado (en rojo). La familia SOCS consta de 8 proteínas que tienen la capacidad de unirse a residuos de los receptores de citoquinas para bloquear la capacidad de activar vías de señalización intramolecular evitando consecuencias dañinas de la estimulación excesiva de estas vías. SOCS3 es importante en la reducción de la actividad de citoquinas inflamatorias, pero su sobreexpresión resulta en la resistencia a la leptina y la insulina. La reducción de la entrada de insulina al músculo esquelético y al tejido adiposo favorece el desarrollo de diabetes tipo 2 (Pedroso 2019, Salminen 2021).

También se observa sobreexpresión de PP1 (protein phosphatase 1) en las personas prediabéticas en la vía de la insulina de la figura 27. La insulina en el músculo esquelético y en el hígado regula la síntesis de glucógeno suprimiendo la glucogenólisis y promoviendo la actividad de la glucógeno sintetasa (GYS) desfosforilándola mediante PP1 (Newgard 2000). En tejido adiposo blanco PP1 des-

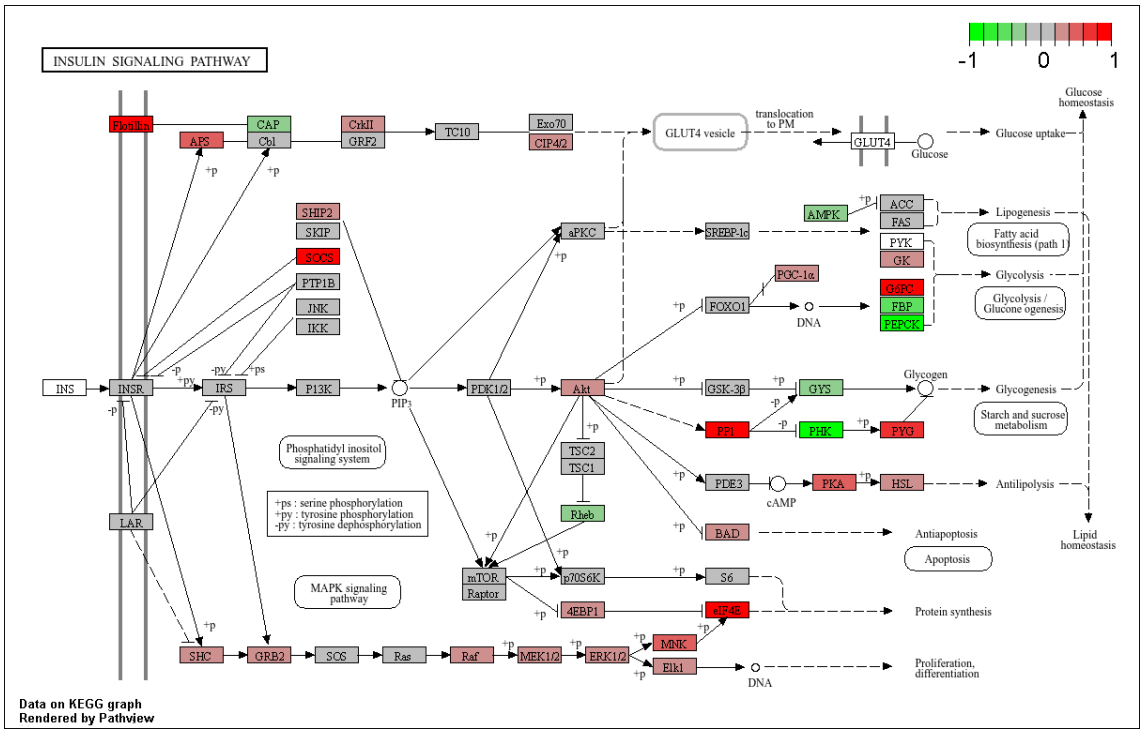


Figura 27: **Análisis KEGG de la vía de insulina en pacientes prediabéticos.** Esquema de la vía de señalización de la insulina de KEGG (Kyoto Encyclopedia of Genes and Genomes) en el análisis de prediabéticos vs porcentaje de hemoglobina glicosilada (HbA1C) baja. En verde genes subexpresados en prediabéticos. Gris: sin diferencias significativas. Rojo: genes sobreexpresados en prediabéticos.

fosforila proteínas regulatorias lipolíticas suprimiendo la lipólisis (Lee 2022). En oposición a SOCS se vió en animales que sobreexpresan PP1 una limpieza rápida de la glucemia y una mejora de la sensibilidad a la insulina en estado postprandial (Li 2019).

El hígado, los riñones y el intestino delgado tienen la capacidad de volcar glucosa a la sangre gracias a la unidad catalítica de la enzima glucosa-6-fosfatasa (G6PC) que sólo se expresa en estos tejidos. Se observa (Figura 27) que en el caso de los participantes prediabéticos del estudio de sujetos sin enfermedad aguda la G6PC se encuentra mayormente expresada que en los controles con hemoglobina glicosilada baja. Este gen se encuentra sobreexpresado cuando existe hiperglucemia independientemente del efecto de la insulina, ya que metabolitos de la glucosa inducen su transcripción y estabilizan su ARNm (Gautier-Stein 2012). Sobre el gen G6PC se tiene un particular interés dado que su represión es un componente del mecanismo de acción de la metformina, la principal droga administrada para la hiperglucemia (Moonira 2020). Su sobreexpresión en hígados de rata es suficiente para desencadenar desregulaciones hepáticas y periféricas asociadas a la diabetes

(Gautier-Stein 2012).

Según lo expuesto, se pudo ver una diferencia en la expresión de algunos de los genes relacionados con la diabetes de los participantes con hemoglobina glicosilada baja contra los que aún no entran en los parámetros de diagnóstico de la enfermedad, por lo que son considerados prediabéticos y no están medicados.

7.4.1.2. Agrupamientos no supervisados

Al correr los modelos de clusterización no se observaron agrupamientos ni en PCA ni en UMAP. Al parecer la señal de los 185 genes diferencialmente expresados en los 52 participantes no alcanzó a ser detectada por los algoritmos (Figura 28) que mostraron un patrón azaroso en la distribución de la condición de prediabetes de los participantes.

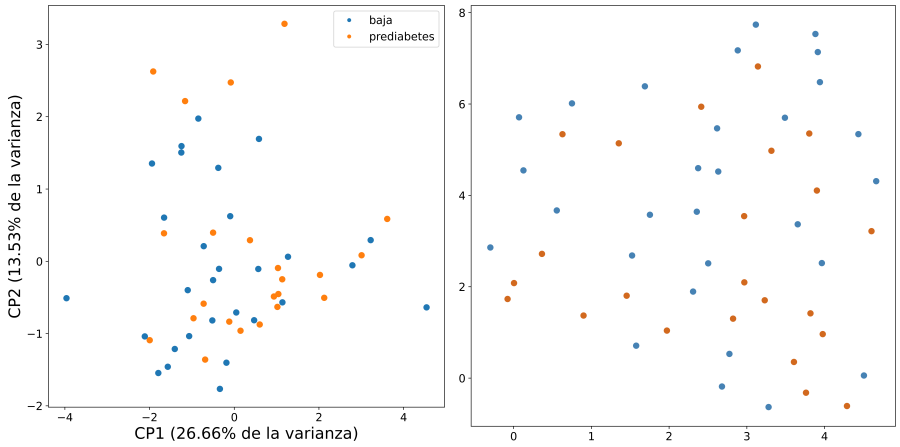


Figura 28: **Análisis de agrupamientos no supervisados en pacientes prediabéticos.** Izq: Gráfico de análisis de los componentes principales (PCA) entre prediabéticos y controles; CP: componente principal. Der: Uniform Manifold Approximation and Projection (UMAP): distancia mínima = 0.7, número de vecinos = 30, épocas = 500. En ambos gráficos se marcan los controles en azul y los prediabéticos en naranja.

7.4.1.3. Red neuronal

Sin embargo, la red neuronal residual de 50 capas logró clasificar bien entre 8 y 9 participantes de los 10 utilizados en la validación (Figura 29, derecha). Se entrenó con 42 imágenes, un número muy bajo para entrenamientos visuales, por lo que se bajaron las épocas a 8 para no sobreajustar el modelo. Un ejemplo de entrenamiento se puede observar en el gráfico de la función de error vs las épocas a la izquierda de la figura 29. El promedio de la exactitud de clasificación

para las 5 corridas fue 0.880 (±0.045). Al tener un número tan bajo de imágenes para la validación cada imágen tuvo un peso muy grande en la exactitud final, queda pendiente a futuro analizar con mayor cantidad de imágenes para tener un resultado más robusto en este ítem.

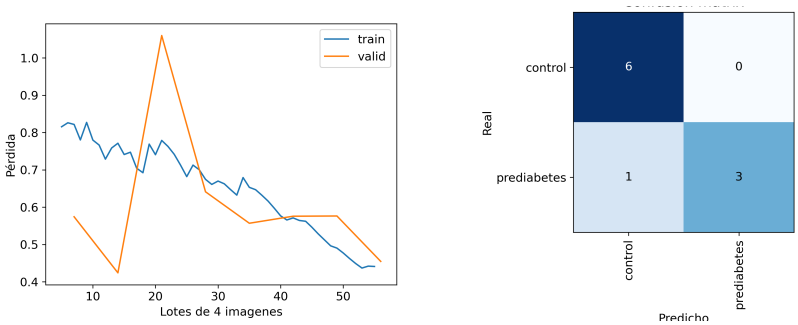


Figura 29: **Análisis de redes neuronales en pacientes prediabéticos.** Izq.: Gráfico de la función de pérdida o error por lote de 4 imágenes para 7 épocas en el entrenamiento para clasificar entre prediabetes y control para el estudio de participantes sin enfermedad activa. Der.: Matriz de confusión resultado de una de las validaciones de ResNet50 para las 10 imágenes correspondientes para prediabéticos y controles.

7.4.2. Dislipemia

En el caso de dislipemia primero se retiraron del análisis a los participantes que consumían algún tipo de fármaco para reducir el colesterol (estatinas o ezetimibe), ya que se consideró que la ingesta de hipolipemiantes podría introducir ruido en la expresión génica. Luego se ordenó a los participantes en tres grupos de acuerdo a los tercios encontrados:

- Colesterol total alto, definido como mayor o igual que 207 mg/dl.
- Colesterol total medio, definido como entre 207 y 183 mg/dl.
- Colesterol total bajo, definido como menor o igual que 183 mg/dl.

A fin de magnificar las diferencias significativas entre dos grupos para los análisis se descartó al grupo intermedio. Se obtuvieron un total de 237 muestras (120 con colesterol alto y 117 con colesterol bajo) con un promedio de edad de 41 (+/- 11) años para las 154 muestras de participantes femeninas y de 41 (+/- 12) años para las 83 masculinas.

7.4.2.1. Expresión diferencial de genes

Ajustando el modelo en DEseq2 por sexo y edad se encontraron 26 genes diferencialmente expresados con un límite de p ajustado de 0.05, no encontrándose ningún proceso biológico significativo en GO para este grupo de genes, probablemente por el bajo número de genes involucrados. En la tabla 5 se listan los 26 genes hallados, 15 de ellos codifican para proteínas, 3 son pseudogenes procesados, 1 gen de la cadena variable de una inmunoglobulina que sufre recombinación somática antes de la transcripción y de 7 genes no se encuentran datos. Estos genes y pseudogenes están involucrados en regulación génica, síntesis de proteínas, respuesta inmune y función mitocondrial. También se observan algunos genes ubicados en el cromosoma Y con roles en funciones específicas masculinas, como la espermatogénesis.

Al realizar el análisis mediante GAGE con GO sólo se encuentran 3 vías significativamente (p ajustado menor a 0.05) enriquecidas: “Hemostasis” (GO:0007599), “Coagulación” (GO:0050817) y “Coagulación sanguínea” (GO:0007596).

En el análisis con las anotaciones de la Enciclopedia de Genes y Genomas de Kioto no se encuentran vías con una significancia menor a 0.05 del valor q (p ajustado por la tasa de falsos positivos en testeos de comparaciones múltiples (FDR)). Se encontraron vías enriquecidas con una significancia ajustada de 0.28 como la “vía de la insulina”, “vía de señalización de adipocitoquinas”, la “vía de señalización del calcio” y en “contracción del músculo liso vascular”. También en procesos biológicos del aparato digestivo como: “secreción pancreática”, “secreción de ácido gástrico”, “secreción de bilis” y en procesos de respuesta inmune: “procesamiento y presentación de antígenos”, “vía de señalización del receptor de células B” y “vía de señalización del receptor tipo toll”.

Dentro de la vía de la insulina (Figura 30) se encuentra subexpresada AMPK (proteína quinasa activada por AMP) en las personas con colesterol total alto. La AMPK es un importante sensor de energía y se activa en respuesta a la falta de glucosa, restricción calórica o un incremento en la actividad física. Se encarga de cambiar las vías metabólicas hacia la formación de ATP y la inhibición de su consumo. Su activación a largo plazo promueve la génesis de mitocondrias, reduce la acumulación intramuscular de lípidos y mejora la acción de la insulina. En el tejido graso inhibe la síntesis de ácidos grasos, eleva la β-oxidación e inhibe la lipólisis. La desregulación de la AMPK está asociada a desórdenes metabólicos,

Ensemble ID	Nombre	$\log_2$(FC)	p ajust.	Tipo
ENSG00000067048.17	DDX3Y	-2.415	6.4×10^{-7}	codifica para proteína
ENSG00000183878.16	UTY	-2.686	1.6×10^{-5}	codifica para proteína
ENSG00000240661.3	–	-2.506	1.6×10^{-5}	–
ENSG00000211658.2	IGLV3-27	-1.303	1.6×10^{-5}	IG_V_gene
ENSG00000067646.12	ZFY	-1.941	2.1×10^{-5}	codifica para proteína
ENSG00000198692.10	EIF1AY	-2.778	4.2×10^{-5}	codifica para proteína
ENSG00000131002.13	–	-2.884	5.1×10^{-5}	–
ENSG00000231535.8	–	-2.567	6.2×10^{-5}	–
ENSG00000012817.16	KDM5D	-2.641	0.2×10^{-4}	codifica para proteína
ENSG00000198695.2	MT-ND6	-0.934	0.3×10^{-4}	–
ENSG00000129824.16	RPS4Y1	-2.495	0.7×10^{-4}	codifica para proteína
ENSG00000215048.13	HLA-DPB1	-2.070	0.001	codifica para proteína
ENSG00000114374.13	USP9Y	-2.321	0.002	codifica para proteína
ENSG00000247627.2	MTND4P12	-1.055	0.002	pseudogen procesado
ENSG00000234810.4	–	0.500	0.003	–
ENSG00000140265.14	ZSCAN29	-0.133	0.005	codifica para proteína
ENSG00000131018.25	SYNE1	-0.223	0.011	codifica para proteína
ENSG00000178162.8	–	-0.535	0.011	–
ENSG00000151468.11	CCDC3	0.590	0.012	codifica para proteína
ENSG00000225840.2	–	-4.202	0.013	pseudogen procesado
ENSG00000257473.7	HLA-DQA2	2.267	0.014	codifica para proteína
ENSG00000147206.17	NXF3	-1.239	0.022	codifica para proteína
ENSG00000169330.9	MINAR1	-0.412	0.030	codifica para proteína
ENSG00000254481.1	PTP4A2P2	0.371	0.030	pseudogen procesado
ENSG00000241431.1	–	-0.780	0.041	–
ENSG00000230385.1	–	0.410	0.044	–

Tabla 5: Genes diferencialmente expresados hallados por el DEseq2 entre participantes con colesterol total alto (mayor o igual que 207 mg/dl) y bajo (menor o igual que 183 mg/dl) del estudio de sujetos sin enfermedad aguda.

síndrome metabólico, resistencia a la insulina y a la diabetes tipo 2 (Szkudelski 2019). También se vió que un grado de inflamación bajo, pero crónico, regula disminuyendo la expresión de AMPK en múltiples tejidos (Ruderman 2013).

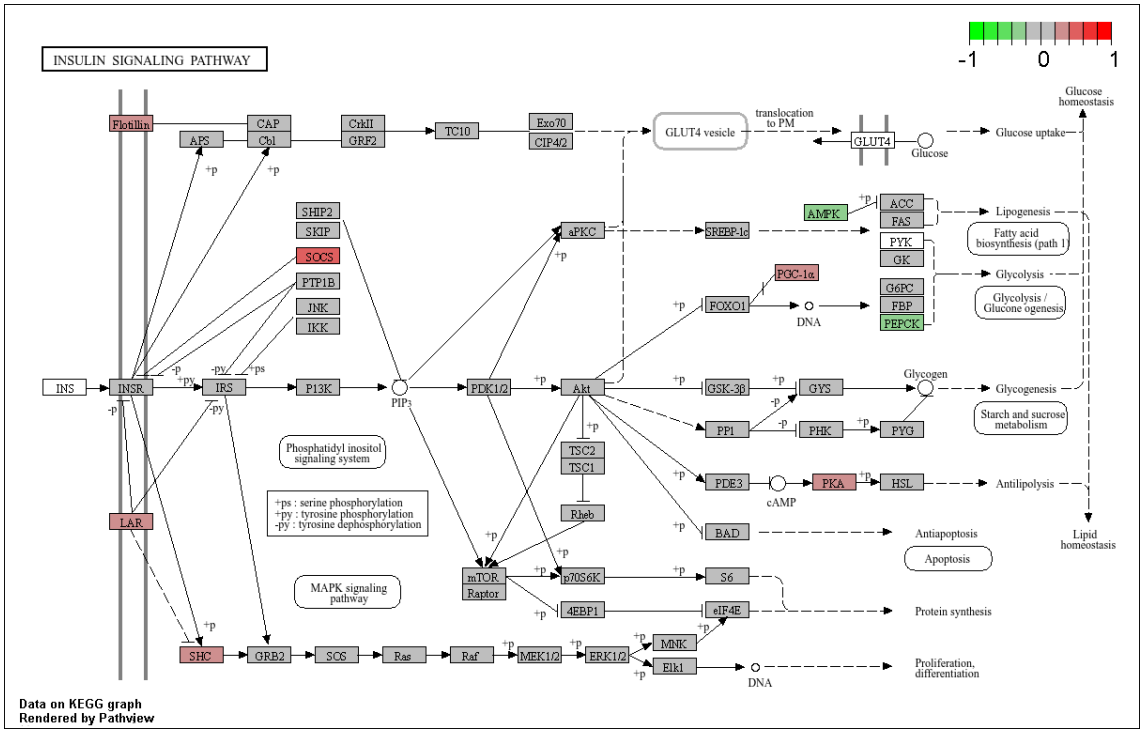


Figura 30: **Análisis GAGE de la vía de insulina en pacientes dislipémicos con KEGG.** Esquema de la vía de señalización de la insulina de KEGG (Kyoto Encyclopedia of Genes and Genomes) en el análisis de colesterol total alto vs bajo. En verde: genes subexpresados. Gris: sin diferencias significativas. Rojo: genes sobreexpresados.

Dentro de los genes diferencialmente expresados en la vía de señalización de las adipocitoquinas se encuentra NPY. Este gen codifica para el neuropéptido Y, una pequeña proteína orexigénica que se expresa en el sistema nervioso central y, periféricamente, en tejido adiposo, páncreas, hígado, músculo esquelético y osteoblastos. Se encontró que NPY se halla sobreexpresado en participantes con colesterol alto al compararlo con participantes con colesterol bajo.

La calcio-ATPasa de membrana plasmática (PMCA) es una bomba ubicua de las células eucariotas encargada de la regulación fina de la concentración de calcio citosólico. Su actividad es regulada de muchas maneras, pero al ser una enzima proteica de membrana la composición lipídica es importante, ya que se encuentra asociada a las balsas lipídicas de membrana, nanodominios enriquecidos en colesterol. Existe evidencia de que las proteínas de transporte de calcio están moduladas por su microambiente lipídico, por ejemplo, los fosfolípidos ácidos, como fosfatidil serina o ácidos grasos poliinsaturados, estimulan la PMCA (Krebs 2022)

y la eliminación del colesterol celular inhibe su actividad (Jiang 2007). Mutaciones en esta bomba se relacionaron con patologías relacionadas al estrés oxidativo como aterosclerosis, diabetes y enfermedades neurodegenerativas (Conrard 2019). En los sujetos con colesterol alto se observó una fuerte sobreexpresión de PMCA frente a los sujetos con colesterol bajo, apareciendo significativamente en la "vía de señalización del calcio" y en "secreción pancreática" en células acinares del páncreas de la Enciclopedia de Genes y Genomas de Kioto.

7.4.2.2. Análisis no supervisados en pacientes con dislipemia

Al realizar el análisis de componentes principales (Figura 31, arriba a la izquierda) con los genes diferencialmente expresados entre la condición de colesterol alto y bajo se logran separar dos grupos en el primer componente que explica el 83.7% de la varianza. La condición de colesterolemia alta no explica esta clusterización. Como se observa en la figura 31, los participantes tanto con colesterol alto (naranja), como con colesterol bajo (celeste) aparecen aleatoriamente distribuídos entre ambos clústeres. Para tratar de explicar el misterioso agrupamiento se coloreó en el análisis a los participantes con los datos clínicos conocidos, pero todos se ubicaron azarosamente entre ambos grupos. El sexo y la trigliceridemia se exponen de ejemplo en la figura 31, centro y derecha. También se constató edad, tabaquismo e índice de masa corporal.

Con el análisis de UMAP se puede trazar un paralelismo, ya que también logra agrupar a los participantes en dos clusters bien definidos que no responden a ninguna de las variables conocidas (Figura31 , abajo) ni siquiera la condición para la que se planteó el análisis.

7.4.2.3. Red neuronal

Se entrenó la red con 190 imágenes del estudio de sujetos sin enfermedad activa etiquetados dentro del grupo colesterol alto o colesterol bajo durante 25 épocas. A la izquierda de la figura 32 se observa el gráfico de la función de error para un entrenamiento de 30 épocas, a partir del cual se decidió recortar el entrenamiento a 25 épocas ya que no se observaron mejores resultados para el set de validación a partir de ese umbral. Para validar se utilizaron 47 imágenes, obteniéndose una exactitud de clasificación promedio de 0.634 ($\pm$0.088). A la derecha de la figura 32 se aprecia una de las matrices de confusión donde 30 sujetos fueron bien

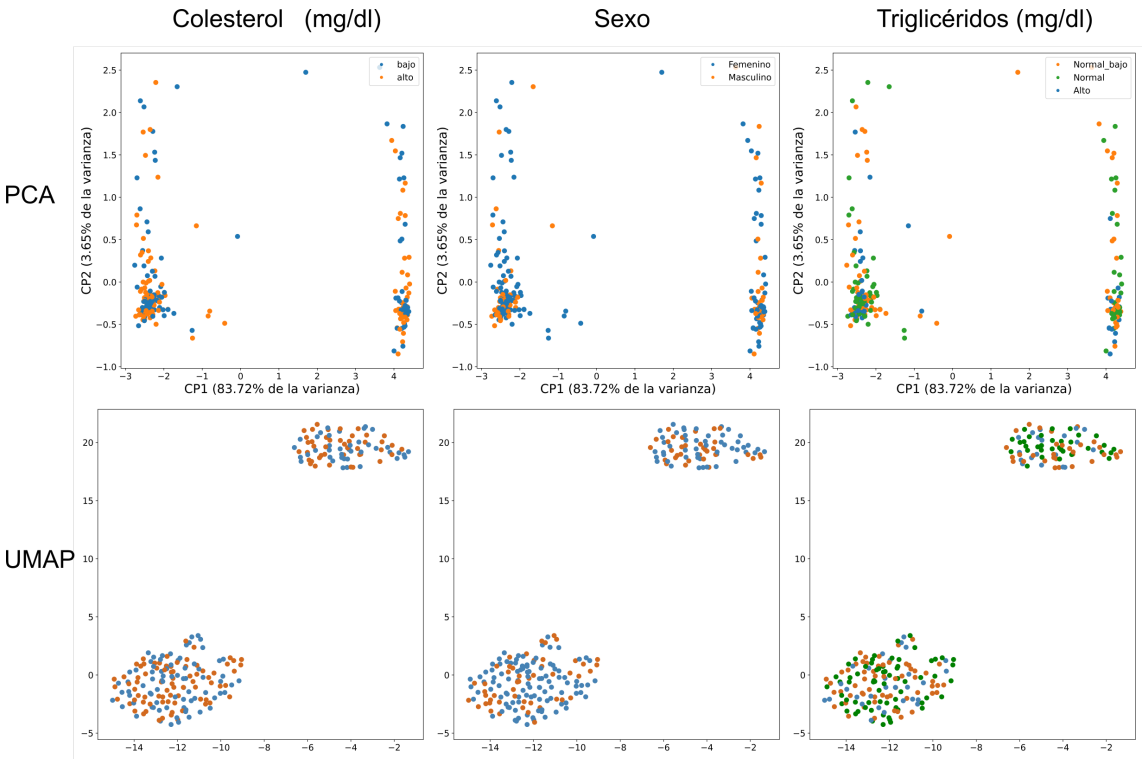


Figura 31: **Análisis no supervisados en pacientes con dislipemia.** Arriba: Gráfico de análisis de los componentes principales (PCA) entre colesterol alto (>= 207 mg/dl) y colesterol bajo (<= 183 mg/dl); CP: componente principal. Abajo: Uniform Manifold Approximation and Projection (UMAP) para participantes con colesterol alto vs bajo. Distancia mínima = 0.7, número de vecinos = 30, épocas = 500. Izq.: Condición en análisis, colesterol alto en naranja, colesterol bajo en celeste para ambos gráficos. Centro: En la condición de análisis se pintaron en celeste los participantes femeninos y en naranja los masculinos en ambos gráficos. Der.: En el análisis de colesterol alto vs. colesterol bajo se pintaron en azul los participantes con nivel de triglicéridos en sangre alto (>150 mg/dl); a los participantes con triglicéridos en valores normales se los dividió en normal (<=150 y >80 mg/dl) en verde y normal bajo (<80 mg/dl) en naranja.

clasificados.

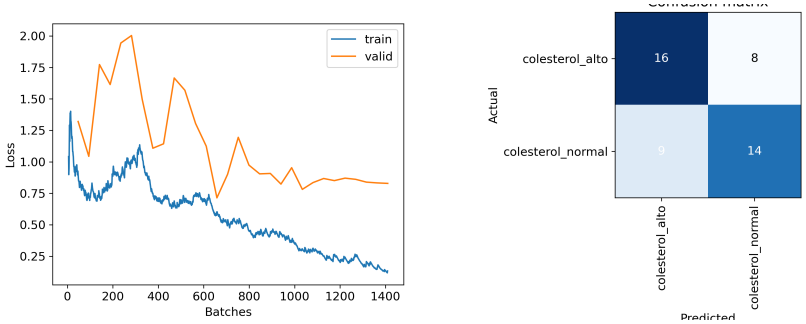


Figura 32: **Redes neuronales en pacientes con dislipemia.** Izq.: Gráfico de la función de pérdida o error por lote de 4 imágenes para 30 épocas en el entrenamiento para clasificar entre colesterol alto y colesterol bajo para el estudio de participantes sin enfermedad activa. Der.: Matriz de confusión resultado de una de las validaciones de ResNet50 para las 47 imágenes correspondientes a los transcriptomas de sujetos con colesterol alto y bajo.

7.4.3. Inflamación

Con la distribución de la proteína C reactiva en los sujetos sin enfermedad activa se formaron cuatro grupos:

- No detectable (0 mg/l)
- Baja (entre 0 y 1.3 mg/l)
- Media (igual o mayor a 1.3 y menor a 2.45 mg/l)
- Alta (mayor o igual a 2.45 mg/l)

Se descartaron las muestras con proteína C reactiva baja y media, obteniéndose 207 participantes luego del filtrado, de los cuales 123 no tenían proteína C reactiva detectable en sangre y 84 personas la tenían alta. Este grupo contó con 133 mujeres con un promedio de edad de 41 ($\pm$11) años y 74 participantes varones con un promedio de edad de 44 ($\pm$13) años.

7.4.3.1. Expresión diferencial de genes

Se realizó un análisis de expresión por medio del DEseq2 con un modelo lineal generalizado con distribución binomial negativa y se obtuvieron 106 genes diferencialmente expresados entre las personas con proteína C reactiva alta y las

personas control con un p ajustado igual a 0.01. A excepción de 8 genes, todos con un nivel de expresión dentro del rango de las 2 veces de cambio. En la figura 33 se muestra el mapa de calor con los genes listados en las filas y los participantes en las columnas. Se destaca que la mayoría de los genes diferencialmente expresados están sobreexpresados en el grupo del cuadrante superior izquierdo que coincide con los sujetos con proteína C reactiva alta (barra lila).

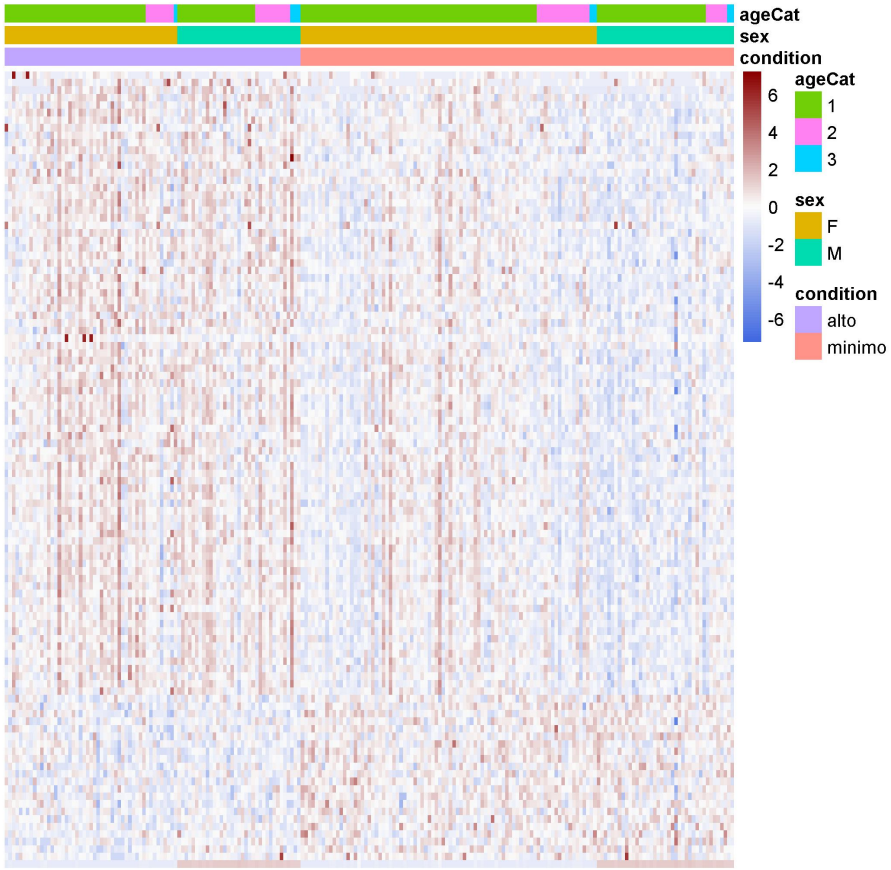


Figura 33: **Mapa de color de acuerdo a niveles de PCR.** Mapa de calor de la expresión diferencial de genes entre sujetos con proteína C reactiva alta (>= 2.45) en coral y controles (PCR = 0) en lila. Las participantes del sexo femenino (F) se marcan en mostaza y las columnas de sujetos de sexo masculino (M) en esmeralda. Los sujetos <50 años se distinguen en verde (1), 50 >= sujetos <= 65 en rosa (2) y los sujetos >65 en cian (3). Se muestran los 106 genes con p ajustado menor a 0.01 en las filas y los 207 participantes en las columnas. Expresión génica aumentada en rojo, disminuída en azul..

Se realizó el análisis de enriquecimiento génico y no se encontraron genes diferencialmente expresados con la notación de GO. Con la enciclopedia de genes y genomas de Kioto tampoco se encontraron resultados significativos (q menor a 0.05). En la figura 34 se observa la vía de señalización de los receptores tipo toll (q = 0.30) que juegan un rol central en la respuesta inmune desencadenada por inflamación (Sameer 2021). Se observa una subexpresión de citoquinas proin-

flamatorias en el grupo control respecto del grupo con proteína C reactiva alta. También se observaron otras vías de señalización como la de las MAPK (proteínas quinasas activadas por mitógeno) (q = 0.36) en la cual se observa una subexpresión de NFκB en el grupo control; la de la apoptosis (q = 0.61) donde TNFα aparece subexpresada en el grupo con proteína C reactiva indetectable; la de los receptores tipo NOD (q = 0.61) donde aparecen subexpresadas en el grupo control citoquinas proinflamatorias como IL-1β, TNFα y RANTES (quimioquina de regulación por activación expresada y secretada por los linfocitos T) entre otras. Aunque los pocos resultados hallados no alcanzaron un umbral satisfactorio para considerar que tienen una tasa de falsos positivos aceptable, van de la mano con lo esperado para un marcador de inflamación como la proteína C reactiva.

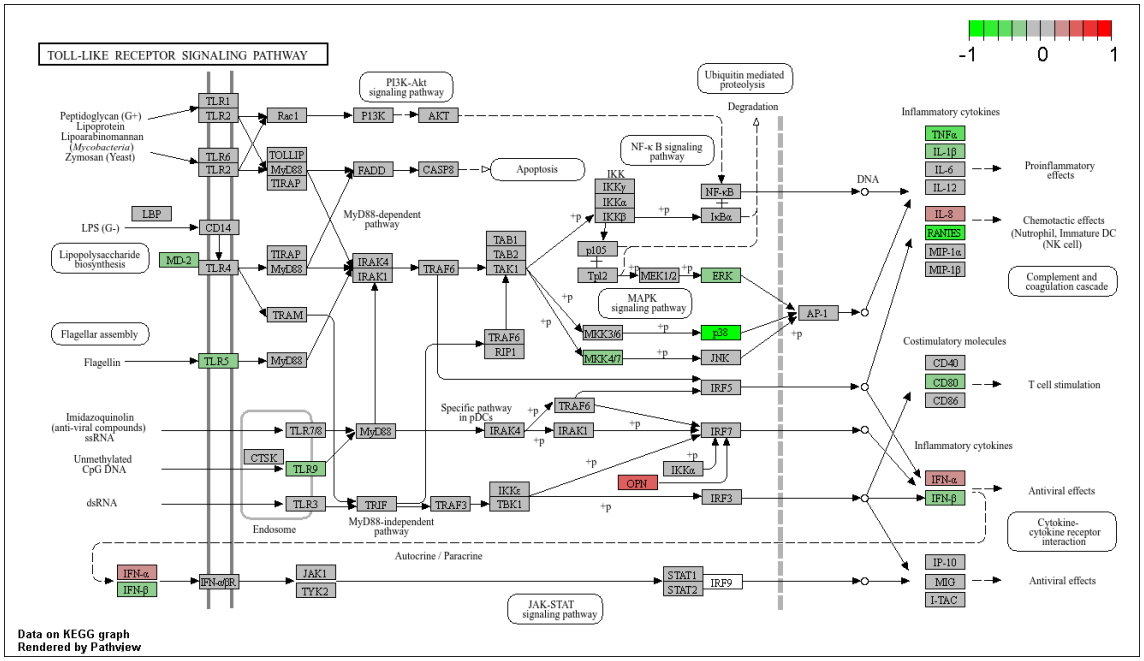


Figura 34: **Análisis GAGE en pacientes con diferentes niveles de PCR: via de MAPK.** Esquema de la vía de señalización de los receptores tipo toll de KEGG (Kyoto Encyclopedia of Genes and Genomes) en el análisis de sujetos con proteína C reactiva alta vs proteína C reactiva no detectable. En verde genes subexpresados en controles. Gris: sin diferencias significativas entre ambos grupos. Rojo: genes sobreexpresados en sujetos con proteína C reactiva no detectable. p = 0.006, q = 0.30.

7.4.3.2. Agrupamientos no supervisados

Se analizó la expresión de los genes diferencialmente expresados para los participantes del estudio de sujetos sin enfermedad activa que tuvieron la proteína C reactiva alta vs los controles mediante PCA y UMAP. En ambos casos se encontraron dos clusters bien separados. Como se detalla en la figura 35 a la izquierda

arriba, el primer componente principal hallado explica el 37.6% de la varianza, sin embargo, no responde a la condición en estudio sino al sexo. En la figura 35 a la izquierda abajo se observa el mismo análisis, pero aparecen los participantes coloreados por sexo, dando cuenta que el primer componente principal separa perfectamente femeninos de masculinos a partir de los transcriptos en estudio. El segundo componente principal no alcanza a dividir a los participantes en dos grupos, pero sí logra un ordenamiento que coincide con la condición en estudio. En los cuadrantes superiores del gráfico de PCA se acumulan los participantes con proteína C reactiva alta en naranja y en los cuadrantes inferiores los que tienen proteína C reactiva indetectable en azul. En el caso de UMAP sucede algo similar aunque no se pueden ubicar a los participantes en el gráfico, debido a la naturaleza estocástica del análisis. En el caso de la corrida de la figura 35 a la derecha arriba, también parece haber una acumulación de participantes con cada condición en cada nube, casualmente también aparecen los cuadrantes superiores enriquecidos en participantes con proteína C reactiva alta. En el gráfico que se encuentra a la derecha abajo en la figura 35 se observa el mismo análisis coloreando a los participantes por sexo y los resultados coinciden en dar crédito al sexo por la separación de los grupos como en el caso de PCA.

7.4.3.3. Red neuronal

A partir de 139 imágenes etiquetadas como nivel de proteína C reactiva alto y mínimo se entrenó la red por 20 épocas. En la figura 36 a la izquierda se observa el gráfico de la función de pérdida para uno de los entrenamientos en el que se puede apreciar un leve sobreajuste. A partir del lote 450 aproximadamente, el error de entrenamiento continúa disminuyendo mientras que el error de validación comienza a aumentar, por tal motivo se decidió no hacer entrenamientos más largos. Se validó el modelo con 35 imágenes y se obtuvo una exactitud promedio de clasificación de 0.771 ($\pm$0.089) para las 5 corridas. En la matriz de confusión de la figura 36 a la derecha se pueden ver los resultados de una validación en la que en el 50% de los casos la red no logró clasificar bien las imágenes presentadas.

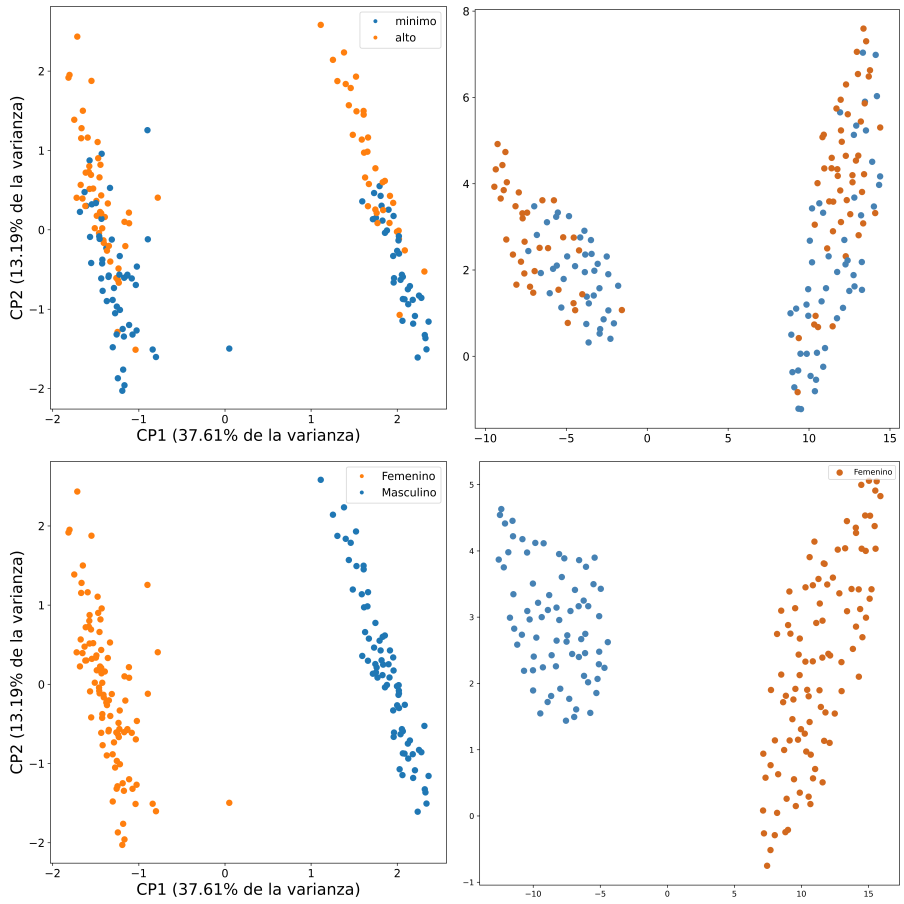


Figura 35: **Agrupamientos no supervisados en pacientes con diferentes niveles de PCR.** Arriba izquierda: Gráfico de análisis de componentes principales (PCA) en el estudio de sujetos sin enfermedad activa para los genes diferencialmente expresados en los participantes con proteína C reactiva alta (>= 2.45 mg/l) y no detectable (0 mg/l); CP: componente principal. Arriba derecha: Uniform Manifold Approximation and Projection (UMAP) para participantes con proteína C reactiva alta vs no detectable. Distancia mínima = 0.7, número de vecinos = 10, épocas = 500. En celeste proteína C reactiva no detectable, en naranja proteína C alta en ambos gráficos. Abajo izquierda: Análisis de componentes principales para el análisis de proteína C reactiva coloreando los participantes por sexo. Abajo derecha: UMAP del análisis de los genes diferencialmente expresados para proteína C reactiva coloreados por sexo. Femenino en naranja, masculino en celeste para ambos gráficos.

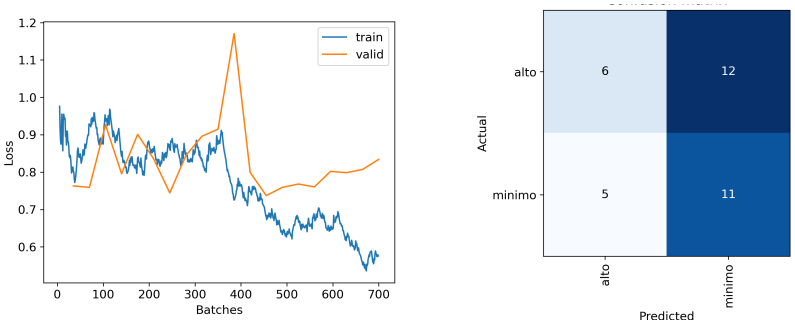


Figura 36: **Red neuronales en pacientes con diferentes niveles de PCR.** Izq.: Gráfico de la función de pérdida o error por lote de 4 imágenes para 20 épocas en el entrenamiento para clasificar entre proteína C reactiva alta y no detectable para el estudio de participantes sin enfermedad activa. Der.: Matriz de confusión resultado de una de las validaciones de ResNet50 para las imágenes correspondientes a los transcriptomas de sujetos con nivel de proteína C reactiva alto (alto) y no detectable (mínimo).

7.5. Estudio en pacientes con score de calcio

7.5.0.1. Características generales del estudio

El estudio reclutó a 200 participantes. Luego de aplicar el filtrado bioinformático, se obtuvieron 196 muestras válidas para el análisis final y 4 muestras se descartaron por problemas técnicos. La distribución de los participantes por sexo fue de 108 hombres (55.4%) y 87 mujeres (44.6%). La edad promedio de los participantes fue de 58 (+/- 8) años, con una media de 60 (+/- 7) años para las mujeres y 56 (+/- 9) años para los hombres, lo que responde a la restricción en los criterios de elegibilidad del estudio que establecían un mínimo de 40 años para los hombres y de 50 años para las mujeres. En la evaluación de los hábitos de fumar, se encontró que 29 participantes eran fumadores activos, 152 participantes (77.9%) no fumaban y 14 participantes habían dejado de fumar hacía más de 12 meses. Finalmente, mediante la tomografía computarizada, se diagnosticó esteatosis hepática en 56 de los 195 participantes, lo que representa el 28.7% de la muestra.

El análisis del score de calcio coronario (CAC) permitió clasificar a los participantes según el grado de aterosclerosis presente. Los resultados obtenidos fueron los siguientes:

- CAC 0 (Patología no identificable): 114 participantes
- CAC 1 a 99 (Patología leve): 46 participantes

- CAC 100 a 399 (Patología moderada): 19 participantes
- CAC 400 a 2785 (Patología severa): 16 participantes

		CAC (Puntuación)			
		0	**1 a 99**	**100 a 399**	**400 a 2785**
SIS (Puntuación)	**0**	99	0	0	0
	1 a 2	15	30	2	0
	3 a 4	0	13	6	2
	5 a 7	0	3	11	3
	>8	0	0	0	11

CAC - puntaje de calcio en las arterias coronarias.
SIS - puntaje de involucramiento de segmentos.

Tabla 6: Puntuaciones de CAC y SIS en la población del estudio.

Más de la mitad de los participantes (55.9%) no presentaron evidencia de calcificaciones coronarias (CAC 0). La prevalencia de aterosclerosis leve fue de 25.6%, mientras que los casos de aterosclerosis moderada y severa fueron menos comunes (10.3% y 8.2%, respectivamente). Estos resultados muestran la heterogeneidad en la progresión de la aterosclerosis entre la población de estudio. La evaluación del SIS muestra que los niveles más altos de CAC están asociados con un mayor número de segmentos coronarios afectados. Todos los participantes con CAC en el rango de 400 a 2785 y un SIS de 8 o más, indicando una enfermedad aterosclerótica extensa, se concentraron en este grupo de alta puntuación.

7.5.1. Análisis de la aterosclerosis coronaria

Para el análisis de la aterosclerosis en el estudio de calcio coronario se designó al grupo con aterosclerosis (67 participantes, media edad = 63 (±7) años) cuando el calcio coronario encontrado fue extenso, moderado o leve en las arterias coronarias y se eliminó del análisis a los pacientes cuando la lesión aterosclerótica se encontró en la aorta. El grupo control se asignó a los participantes con ausencia de calcio coronario (66 participantes, media edad = 52 (±6) años). La distribución del sexo fue de 55 mujeres con un promedio de edad de 59 (±7) años y 78 hombres con un promedio de edad de 56 (±10) años (Tabla 7).

	Controles	Casos	p-valor
Media de la edad (años)	52	63	<0.01
Sexo Femenino (%)	20	21	0.45
Diabetes (%)	12	12.5	1
Hipertensión (%)	29.0	46.9	<0.01
Dislipidemia (%)	23.0	36.5	0.01
Fumador Actual (%)	13	16.7	0.60
Obesidad (%)	39.0	37.5	0.95
Tratamiento con Estatinas (%)	15.0	31.3	0.01
Riesgo de ECV <10%	76.0	51.0	
Riesgo de ECV 10 a 19%	13.0	31.3	
Riesgo de ECV ≥20%	11.0	17.7	<0.01

Tabla 7: Características clínicas de la población del estudio de score de calcio coronario.

7.5.1.1. Análisis de expresión de genes

En la tabla 8 se detallan los 13 genes diferencialmente expresados estadísticamente significativos (p ajustado <0.05) que entregó como resultado el DEseq2 entre individuos con aterosclerosis y controles. El modelo fue ajustado por sexo y edad. Esta lista de genes no dió un resultado estadísticamente significativo al buscar asociaciones a procesos biológicos en GO. Sin embargo, algunos de los genes identificados podrían estar involucrados en la fisiopatología de la aterosclerosis. Por ejemplo, la alteración en la expresión de CNTNAP3B podría afectar la integridad de las uniones celulares y la estabilidad de la matriz extracelular. La expresión diferencial de RCCD1 sugiere que la dinámica del citoesqueleto podría estar comprometida en las células de los pacientes. Estos factores son importantes en el proceso de migración celular, clave en la progresión de la aterosclerosis. GPR15 codifica para un receptor de quimioquina. Las quimioquinas están implicadas en la respuesta inflamatoria y en la migración de células inmunitarias, procesos involucrados en la patogénesis de la aterosclerosis. CNOT3 codifica para una proteína que regula el mantenimiento de células madre. La expresión diferencial de CNOT3 podría influir en la capacidad de regeneración y reparación de tejidos en pacientes con aterosclerosis.

Se realizó un análisis de enriquecimiento génico utilizando el método GAGE con anotaciones de GO para comparar la expresión génica diferencial entre participantes con aterosclerosis y controles sin aterosclerosis. Los resultados del análisis

Nombre	$\log_2$(FC)	*p* ajust.	Tipo	Función
PPDPF	1.171	0.048	Codificante	Involucrado en diferenciación celular
GPR15	1.018	0.029	Codificante	Receptor de quimioquina
CNTNAP3B	1.950	0.1×10^{-4}	Codificante	Involucrado en adhesión celular
RCCD1	1.037	0.021	Codificante	Estabilidad de microtúbulos
ESPN	1.463	0.029	Codificante	Regulación en microvilli en células mecano y quimiosensoriales
RPL10P6	-4.201	2.99×10^{-7}	Pseudogen	Pseudogen de proteína ribosomal
LINC02641	0.472	0.048	lncRNA	Asociado a altura corporal en GWAS
LINC01362	1.171	0.048	lncRNA	Asociado a altura corporal en GWAS
RPL10P9	-1.803	3.11×10^{-5}	Pseudogen	Pseudogen de proteína ribosomal
RN7SL4P	-1.272	0.021	miscRNA	Facilita la translocación de proteínas a través de las membranas
KMT2B	0.948	0.048	Codificante	Metiltransferasa
CNOT3	1.454	3.11×10^{-5}	Codificante	Regulación del mantenimiento de células madre
LOC101928438	0.726	0.032	lncRNA	Sin datos

FC: Fold Change (cambio en la expresión); Codificante: codifica para proteína; lncRNA: ARN largo no codificante; GWAS: estudio de asociación del genoma completo; miscRNA: RNA misceláneo.

Tabla 8: Genes diferencialmente expresados identificados en el análisis de aterosclerosis coronaria.

del enriquecimiento (Figura 37) muestran que los participantes con aterosclerosis presentan una sobrerrepresentación de genes asociados con procesos inflamatorios, respuesta inmune y coagulación sanguínea. Estos hallazgos son consistentes con la patogénesis de la aterosclerosis, que se caracteriza por la inflamación crónica, la disfunción endotelial y la formación de placas ateroscleróticas. El enriquecimiento de genes relacionados con la respuesta a citoquinas y la cascada de I-κB kinasa/NF-κB muestra una activación de vías inflamatorias en los participantes con aterosclerosis. Además, la sobrerrepresentación de genes involucrados en la regeneración de tejidos y la regulación de la coagulación sanguínea muestra la importancia de estos procesos en la reparación y mantenimiento de la integridad vascular en el contexto de la aterosclerosis.

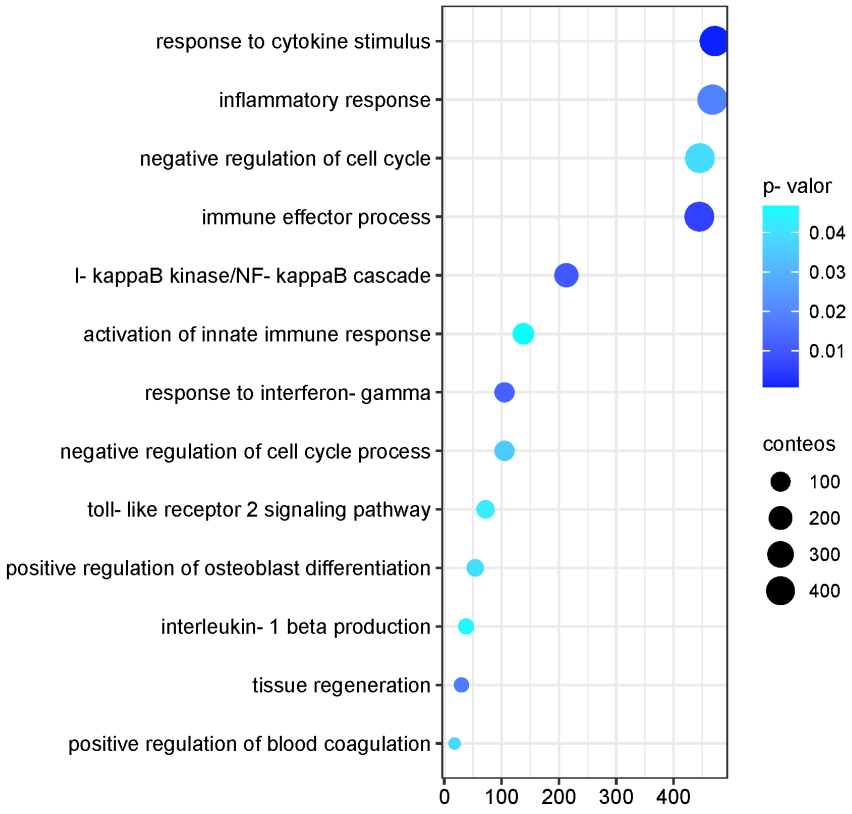


Figura 37: **Análisis GAGE en pacientes con aterosclerosis coronaria.** Gráfico de burbuja del análisis de enriquecimiento de genes con GAGE en procesos biológicos de GO para el estudio de calcio coronario para los sujetos con aterosclerosis vs controles sin aterosclerosis.

Al buscar contexto biológico de las vías alteradas en la aterosclerosis mediante GAGE con anotaciones de KEGG se encuentran algunas coincidencias. La vía de señalización de quimioquinas, crucial para la respuesta inmune y la inflamación

es la que aparece con mayor número de genes sobrerrepresentados. Pero también agrega dos ejes diferentes: metabolismo energético y detoxificación celular.

La desregulación del metabolismo del glutatión y de la vía de los peroxisomas, crucial para la defensa antioxidante y la detoxificación celular, refleja el aumento del estrés oxidativo en la aterosclerosis. La desregulación energética se vió reflejada en la vía del metabolismo del piruvato, las vías del metabolismo de hidratos de carbono (Fructosa y manosa, la vía de la glucólisis y la gluconeogénesis), de lípidos (Metabolismo del butirato y propionato) y de proteínas (Degradación de aminoácidos de cadena ramificada, metabolismo de glicina, serina y treonina).

Se ha visto en estudios de transcriptomas de placas ateroscleróticas que su composición no sólo depende de la progresión sino también de su ubicación. Se encontraron diferencias significativas entre placas de carótidas y femorales. A las femorales se las identificó con enriquecimiento de genes involucrados en osteogénesis y morfogénesis ósea, mientras que las placas carotídeas se vieron enriquecidas con genes de respuesta inmune y metabolismo de lípidos (Steenman 2018). En este caso, al tener el transcriptoma sanguíneo, se observa el amplio espectro de las diferencias de todas las placas ateroscleróticas, tanto las placas tendientes a microcalcificaciones como a las predispuestas a metaplasia osteoide.

7.5.1.2. Análisis no supervisados

Se realizó un análisis de componentes principales y se aplicó UMAP a las muestras de los participantes con aterosclerosis y sin aterosclerosis con los 13 genes encontrados como diferencialmente expresados por el DEseq2. Como se observa a la izquierda de la figura 38, se encontró un pequeño agrupamiento que no responde a ninguna variable medida en el estudio. Se analizó sexo, edad, tabaquismo, índice de masa corporal, presión sanguínea, efecto lote y zona de la lesión ateroscleróticа, sin embargo todas respondieron de manera aleatoria. Tampoco se encontró un ordenamiento de los sujetos con aterosclerosis (en naranja) y los controles (en azul en la figura) que haga sospechar que alguno de los componentes esté explicando la condición en estudio.

En el caso de UMAP se observó un comportamiento similar al de PCA, pero los participantes agrupados no coincidieron con los de PCA. Tampoco respondieron a ninguna de las variables estudiadas.

Estos resultados mostrarían que los 13 genes hallados son insuficientes para ex-

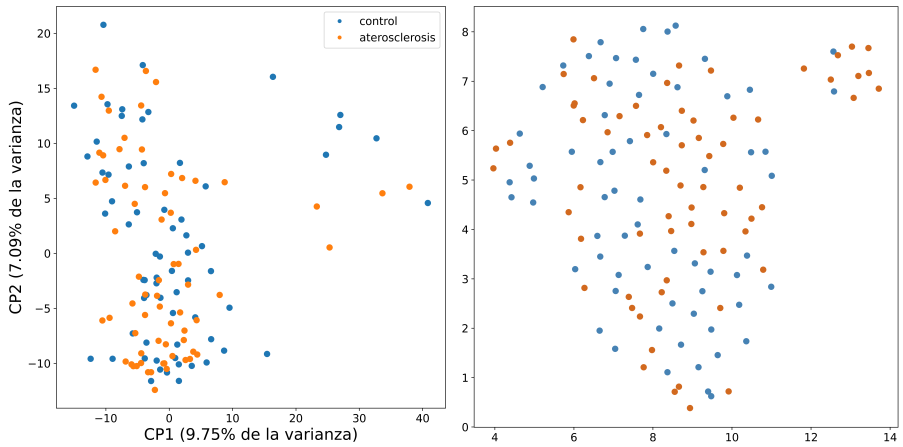


Figura 38: **Análisis no supervisados en pacientes con aterosclerosis coronaria.** Izq: Gráfico de análisis de los componentes principales (PCA) entre muestras con aterosclerosis y controles; CP: componente principal. Der: Uniform Manifold Approximation and Projection (UMAP): distancia mínima = 0.7, número de vecinos = 30, épocas = 500. En ambos gráficos se marcan los controles en azul y los participantes con aterosclerosis en naranja.

plicar la complejidad de la aterosclerosis a nivel de organismo. Los métodos utilizados no tienen la sensibilidad para encontrar la diversidad de genes responsables que, dada la naturaleza de la fisiopatología de la aterosclerosis, actúan sutilmente y en conjunto orquestado sostenido en el tiempo.

7.5.1.3. Red neuronal

Se entrenó la red con 111 imágenes representando la expresión de todos los transcriptos de los participantes del estudio de calcio coronario etiquetadas como aterosclerosis o control. El entrenamiento se realizó por 12 épocas. A la izquierda de la figura 39 se observa la función de pérdida para un entrenamiento de 100 épocas en la que se aprecia una mejora del error de entrenamiento, pero el de validación se mantiene prácticamente constante. Por lo tanto, se decidió entrenar hasta 12 épocas para evitar el sobreajuste. La exactitud de clasificación alcanzada promediando las 5 corridas fue de 0.733 ($\pm$0.066). Un punto interesante a destacar es que la red logra encontrar una señal en una patología complicada por su diversidad fisiológica y su sutileza en la diferencia de expresión en la que los algoritmos no supervisados no alcanzaron a encontrarla. En la matriz de confusión de la figura 38 se observan las cantidades de imágenes que corresponden a los participantes controles y con aterosclerosis verdaderos y los predichos por la red para una de las corridas. En 8 imágenes la red no logró diferenciar a los controles

y los predijo como ateroscleróticos y, a la inversa, predijo 4 controles cuando realmente padecían aterosclerosis. En este ejemplo, resultan un total de 12 imágenes mal clasificadas sobre 27 analizadas.

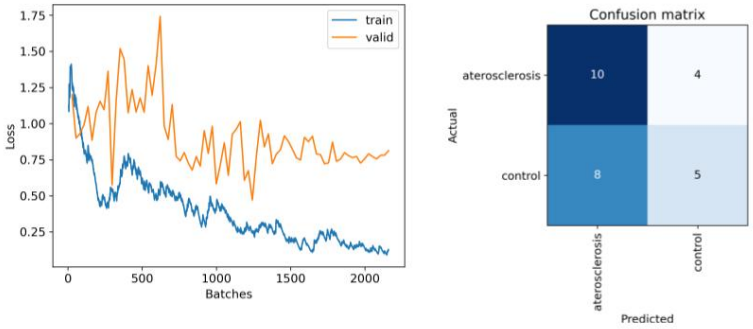


Figura 39: **Análisis de red neuronal en pacientes con aterosclerosis coronaria.** Izq.: Gráfico de la función de pérdida o error por lote de 4 imágenes para 100 épocas en el entrenamiento para clasificar entre participantes con aterosclerosis y controles para el estudio de calcio coronario. Der.: Matriz de confusión resultado de una de las validaciones de ResNet50 para las imágenes correspondientes a los transcriptomas de sujetos con aterosclerosis y controles.

7.6. Estudio en pacientes con insuficiencia cardíaca

Al finalizar el estudio se reclutaron un total de 108 pacientes con insuficiencia cardíaca, siendo 89 de ellos masculinos y 19 femeninos; por lo que la muestra representa el desbalance propio de la enfermedad hacia el sexo masculino. En la figura 40 A se puede observar la distribución de la edad sesgada hacia la izquierda, con una mayor concentración de pacientes en el rango de 50 a 70 años y una media de 62 ($\pm$12) años. Este rango de edad es consistente con la prevalencia de insuficiencia cardíaca en la población general, donde la incidencia aumenta con la edad. Las pacientes femeninas tienen un promedio de edad de 63 ($\pm$11) años y los masculinos 61 ($\pm$12) años.

Los valores de fracción de eyección del ventrículo izquierdo (FEVI) varían entre el 15 % y el 57 %, con una media aproximada de 35 %. La figura 40 B muestra la distribución de la FEVI entre los pacientes y confirma la insuficiencia cardíaca del tipo con fracción de eyección reducida.

La mayoría de los pacientes pertenecen a la clase funcional 1 y 2 de la New York Heart Association (47 pertenecen a la clase 1 y 48 a la clase 2), 12 se encuentran en la clase 3 y sólo 1 en la clase más severa 4. La figura 40 C presenta un histograma con la distribución de los pacientes según la clase funcional NYHA, mostrando la diversidad en la severidad de los síntomas entre los participantes. La etiología de la insuficiencia cardíaca en los pacientes del estudio es variada, con una predominancia de causas coronarias. Aproximadamente el 50% de los pacientes tienen insuficiencia cardíaca de origen coronario, seguida por etiologías idiopáticas y otras causas menos comunes como la valvulopatía, la hipertensión y la amiloidosis (Figura 40 D).

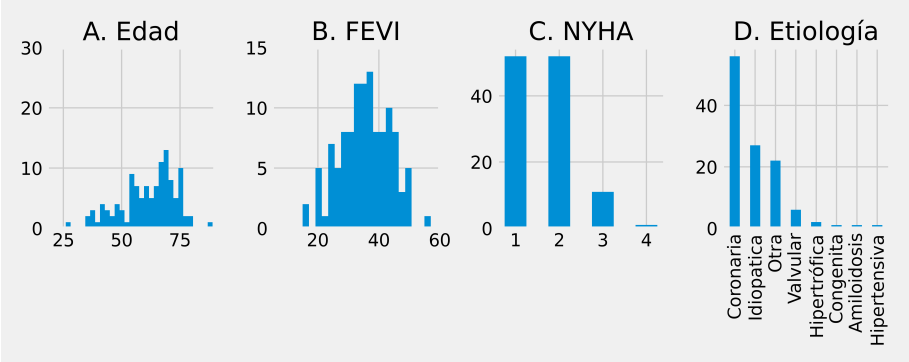


Figura 40: **Distribución de variables clínicas en pacientes con insuficiencia cardíaca.** Datos clínicos de los participantes del estudio de insuficiencia cardíaca. A: Distribución de edad. B: Distribución de la fracción de eyección del ventrículo izquierdo (FEVI). C: Histograma con el número de pacientes en cada una de las clases funcionales según la New York Heart Association (NYHA). D. Histograma con el número de pacientes según la causa de insuficiencia cardíaca diagnosticada.

Los resultados presentados hasta ahora, sumados a los que se especifican en la tabla 9, proporcionan en detalle el perfil clínico de los pacientes con insuficiencia cardíaca.

En este estudio sólo participaron pacientes con insuficiencia cardíaca, por lo tanto, los controles para los análisis posteriores fueron tomados del estudio de sujetos sin enfermedad aguda activa con algunos criterios específicos. Se tuvo en cuenta que los participantes utilizados como control tengan un nivel de riesgo cardiovascular bajo (o leve de no ser posible bajo) según las tablas de la Organización Mundial de la Salud para la zona sur de América del Sur (Argentina, Chile y Uruguay)(Figura 5). Además, se buscó un control de igual sexo y de similar edad para cada paciente. Por lo tanto, para los siguientes análisis el total de participantes asciende a 215. La distribución de pacientes con insuficiencia cardíaca y controles

Variable clínica	Media	Mediana	Mínimo	Máximo
Hemoglobina (g/dl)	14.0 (±1.5)	14.0	9.7	18.0
Índice de masa corporal	29 (±5)	28	18	45
Glucosa (mg/dl)	111 (±25)	104	70	216
Colesterol HDL (mg/dl)	43 (±12)	41	21	76
Colesterol LDL (mg/dl)	85 (±36)	81	5	180
Triglicéridos (mg/dl)	144 (±115)	115	54	1100
Na plasmático (mEq/l)	139 (±3)	140	126	145
Leucocitos (cél./mm^3)	7817 (±2250)	7200	4105	15650
Creatinina sérica (mg/dl)	1.11 (±0.33)	1.10	0.26	2.36
NT-pro-BNP (pg/ml)	1924 (±1892)	1405	42	4921
GOT-ASAT (UI/l)	22 (±9)	20	10	72
GPT-ALAT (UI/l)	23 (±15)	19	9	129
Presión Sistólica (mmHg)	117 (±17)	120	90	190
Presión Diastólica (mmHg)	72 (±10)	70	50	100
Frecuencia cardíaca (p/min)	69 (±10)	68	46	90

Tabla 9: Variables clínicas y sus estadísticos descriptivos.

queda determinada de la siguiente manera:

- 108 pacientes con una media de edad de 62 (±12) años
- 107 controles con una media de edad de 52 (±12) años
- 38 sujetos femeninas (62 (±11) años)
- 177 sujetos masculinos (56 (±13) años)

7.6.0.1. Expresión diferencial de genes

El resultado del DEseq2 arrojó 11971 genes diferencialmente expresados entre pacientes con insuficiencia cardíaca y controles (p ajustado ≤ 0.01). Este resultado dimensiona el desbalance metabólico de una enfermedad grave como la insuficiencia cardíaca con fracción de eyección reducida. En la figura 41 se observa un mapa de calor con los 2511 genes que aparecen diferencialmente expresados con un p ajustado menor a 0.01 y un $\log_2$(VC) mayor a 1 y menor a -1 (Desde 18 $\log_2$(VC) hasta -26). Se aprecia una mayor cantidad de genes mayormente expresados en los pacientes que en los controles.

Al observar los procesos biológicos del GO que resultaron significativos para los genes diferencialmente expresados entre pacientes con insuficiencia cardíaca y controles se encuentran procesos de inmunidad, respuesta a estímulos externos e internos y diferenciación celular mayormente. Se señalan algunos genes que se

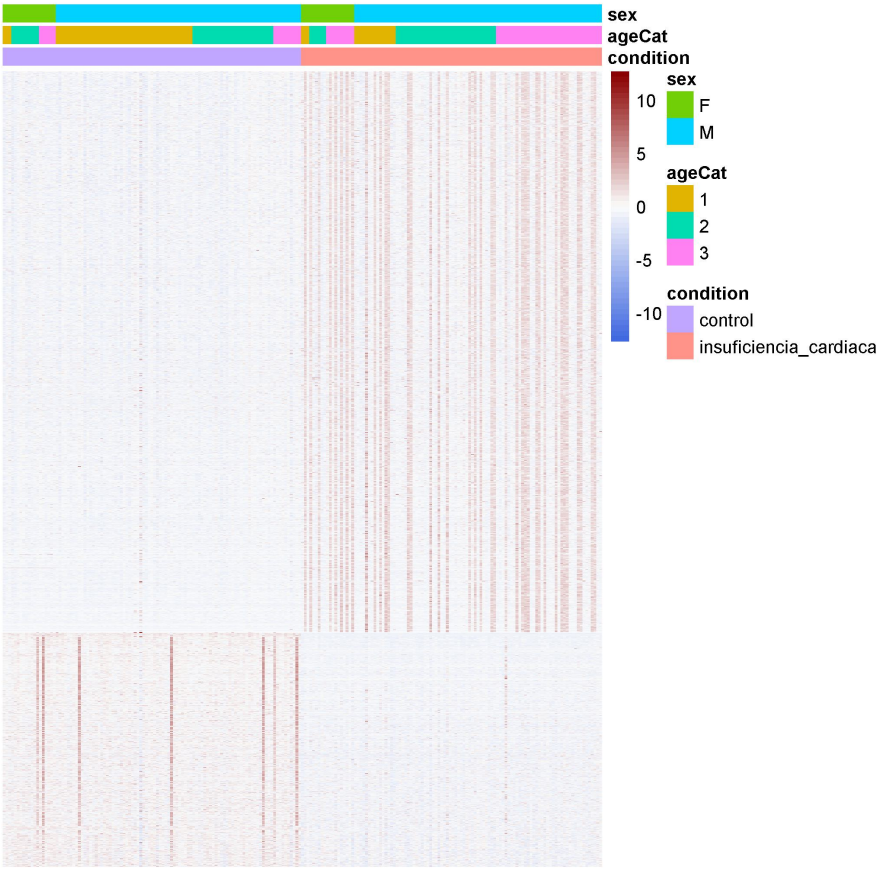


Figura 41: **Mapa de color de la expresión génica en pacientes con insuficiencia cardíaca.** Mapa de calor de la expresión diferencial de genes entre pacientes con insuficiencia cardíaca (barra coral) y controles (barra lila). Las participantes del sexo femenino (F) se marcan en verde y las columnas de sujetos de sexo masculino (M) en cian. Los sujetos <50 años se distinguen en mostaza (1), 50 >= sujetos <= 65 en esmeralda (2) y los sujetos >65 en rosa (3). Se muestran los 2511 genes con p ajustado menor a 0.01 y $\log_2$(VC) >1 y <-1 en las filas y los 215 participantes en las columnas. Expresión aumentada en rojo, expresión disminuida en azul.

encontraron en la vía de la producción de interleuquina 4 como ejemplo de respuesta inmune humoral: genes del complejo mayor de histocompatibilidad (HLA-E, HLA-DRB1), receptores de linfocitos como CD3E, CD28, CD83 y CD40LG. Otros genes que sus productos regulan las vías de inmunidad como GATA3, TNFSF4, FOXP3, LGALS9, CEBPB y ZFPM1. En la figura 42 se pueden apreciar algunos de los procesos biológicos que resultaron significativos y, dentro de ellos, se encuentra "respuesta celular a la hipoxia". Algunos genes relativos a este término que se encuentran diferencialmente expresados entre pacientes y controles son: genes que codifican proteínas inducidas por hipoxia como EPAS1, HILPDA, HIF1AN y endotelina 1. EPAS1 es parte del grupo de las proteínas de unión a ADN llamadas factores inducibles por hipoxia (HIF). Es un factor de transcripción que regula a la hormona eritropoyetina que llevará a aumentar la eritropoyesis para contrarrestar los bajos niveles de oxígeno (Kristan 2019). Induce la expresión de interleuquina 1 y NF-kappa B en adipocitos regulando el proceso inflamatorio. Alivia la resistencia a la insulina previniendo la activación del inflamasoma NLPR3. En el sistema cardiovascular es esencial para mantener la homeostasis de catecolamina y promueve la angiogénesis regulando positivamente la expresión de VEGF. Estudios muestran que tiene un rol importante en la patofisiología de la hipertensión pulmonar que puede llevar a insuficiencia cardíaca (Wang 2023).

En el enriquecimiento génico realizado mediante GAGE con anotaciones del GO 115 procesos biológicos dieron significativos con un q menor a 0.05. En general, se observaron alteraciones en procesos críticos de desarrollo y diferenciación celular, como el desarrollo y la morfogénesis embrionaria. También procesos de señalización y regulación, como la señalización de proteínas Rho y la cascada de MAPK. Se identificaron procesos relacionados con la homeostasis de iones de calcio y la respuesta a estímulos hormonales y mecánicos. Como se vio en la figura 7 la homeostasis de iones es esencial para la función cardíaca normal, por lo que las respuestas alteradas son esperables en la fisiología de la insuficiencia cardíaca. La coagulación sanguínea y la diferenciación de leucocitos también fueron procesos significativamente enriquecidos. La figura 43 muestra algunos de los procesos biológicos significativos ($q < 0.05$) hallados.

En el caso de la angiogénesis la media del estadístico es alta y positiva, indicando un aumento en la actividad de los genes relacionados con el desarrollo de vasos sanguíneos en los pacientes con insuficiencia cardíaca en comparación con los controles. Los factores de crecimiento, la migración celular, cambios en

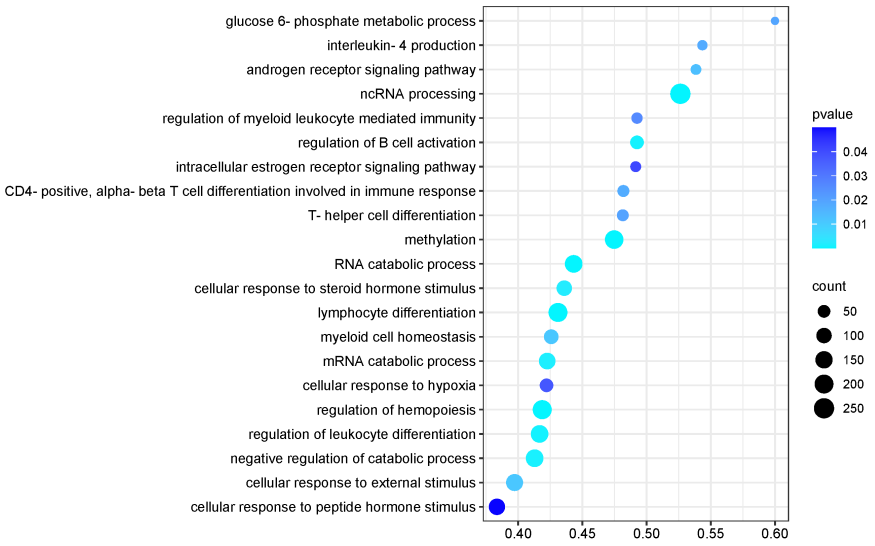


Figura 42: **Ontología génica en pacientes con insuficiencia cardíaca.** Gráfico de burbuja para el análisis de ontología génica de los genes diferencialmente expresados en pacientes con insuficiencia cardíaca frente a controles. Cada fila en el eje y describe un proceso biológico descrito en la ontología génica. En el eje x se expresa el enriquecimiento como la proporción de los genes de la lista de los diferencialmente expresados que se encuentran en esa vía sobre el total de genes de esa vía. El tamaño de la burbuja señala la cantidad de genes involucrados en el proceso. El color indica la significancia, p ajustado hasta 0.05.

la adhesión célula-célula, la formación de nuevos vasos sanguíneos son procesos importantes en la reparación y remodelación del tejido, y su respuesta alterada puede influir en la progresión de la enfermedad. El potencial de membrana es crucial para la excitabilidad celular y la función del músculo cardíaco por lo que su desequilibrio puede afectar la integridad del tejido cardíaco. A diferencia de los otros conjuntos de genes, los relacionados a la organización mitocondrial tienen una media del estadístico negativa, lo que indica una disminución en la organización mitocondrial. Las mitocondrias son esenciales para la producción de energía en las células cardíacas y su disfunción está bien documentada en la falla contráctil en la insuficiencia cardíaca (Zhou 2018).

El resultado del enriquecimiento mediante GAGE con las anotaciones de KEGG arrojó, con un q = 0.14 de significancia y -2.98 la media del estadístico, la vía del peroxisoma. Los peroxisomas son orgánulos celulares esenciales presentes en casi todas las células eucariotas, desempeñando funciones vitales tanto en el metabolismo como en la protección celular. Entre sus roles se encuentran la β-oxidación de ácidos grasos y la degradación del peróxido de hidrógeno. Además, los pero-

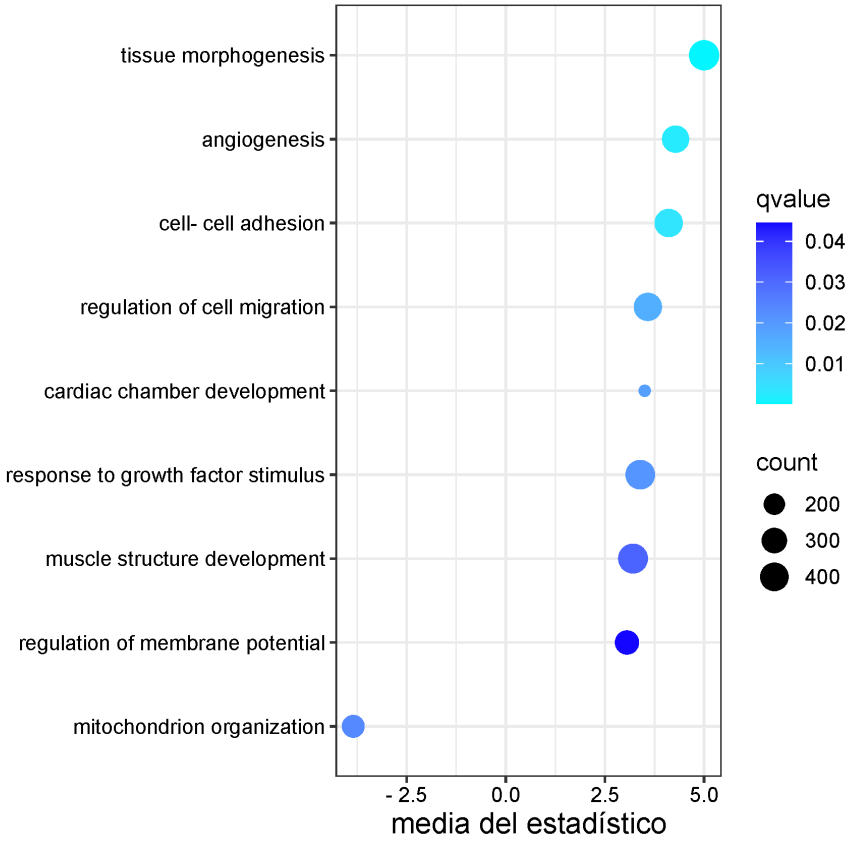


Figura 43: **Análisis GAGE en pacientes con insuficiencia cardíaca.** Gráfico de burbuja para el análisis de enriquecimiento génico mediante GAGE con anotaciones de GO en pacientes con insuficiencia cardíaca frente a controles. Cada fila en el eje y describe un proceso biológico descrito en la ontología génica. En el eje x se señala la media de los estadísticos individuales de las múltiples pruebas de conjuntos de genes; su valor absoluto mide la magnitud de los cambios a nivel del conjunto de genes y su signo indica la dirección de los cambios. El tamaño de la burbuja señala la cantidad de genes del conjunto. El color indica la significancia, q menor que 0.05.

xisomas participan en vías especializadas como la biosíntesis de ácidos biliares y lípidos éter. También están implicados en el metabolismo de radicales libres de oxígeno y óxido nítrico, así como en la señalización intra e intercelular. Su número y tamaño se incrementan cuando se añaden activadores del receptor PPARα, reguladores centrales del metabolismo de lípidos en el corazón (Página 30). Todas las proteínas de la matriz peroxisomal necesarias para que realice sus funciones deben ser importadas al interior desde el citoplasma. Existen dos tipos de señal para dirigir las proteínas al peroxisoma: PTS1 (reconocida por PEX5) y PTS2 (reconocida por PEX7) (Salceda 2008). En la figura 44 se aprecia un gráfico con los genes diferencialmente expresados entre pacientes con insuficiencia cardíaca y controles en las diferentes vías donde está involucrado el peroxisoma. La expresión diferencial de los genes involucrados en la importación de proteínas, tanto de matriz como de membrana, sugiere un desbalance en la composición de los peroxisomas entre pacientes y controles. Los participantes con insuficiencia cardíaca tienen sobreexpresada la vía PTS1-PEX5 y subexpresada la vía PTS2-PEX7. Además, se observa una subexpresión en los genes involucrados en la oxidación de ácidos grasos (alfa y beta oxidación) apoyando la explicación del cambio en el sustrato usado para la alta demanda energética cardíaca que se observa en la insuficiencia cardíaca (aumentando la dependencia de la glucosa y disminuyendo la de los ácidos grasos). Sin embargo, se observa una sobreexpresión de los genes involucrados en la oxidación de ácidos grasos insaturados. La subexpresión del sistema antioxidante del peroxisoma pone en evidencia la responsabilidad del aumento de las especies reactivas del oxígeno en la fisiopatología de la insuficiencia cardíaca como se mencionó anteriormente.

7.6.0.2. Agrupamientos no supervisados

En el análisis de componentes principales para las muestras de pacientes con insuficiencia cardíaca vs. controles, el primer componente logra explicar algo más del 29 % de la varianza (Figura 45 arriba izquierda). En este sentido se encuentran 2 *clusters*, uno más pequeño conteniendo solamente controles (en naranja) y uno mayor con pacientes (azul) y controles. En este grupo mayor pacientes y controles se ordenan de manera distinguible, aunque no se separan. El segundo componente principal explica el 8 % de la variabilidad, pero no coincidió con ninguna variable clínica medida en este estudio. Detalladamente se podría decir que

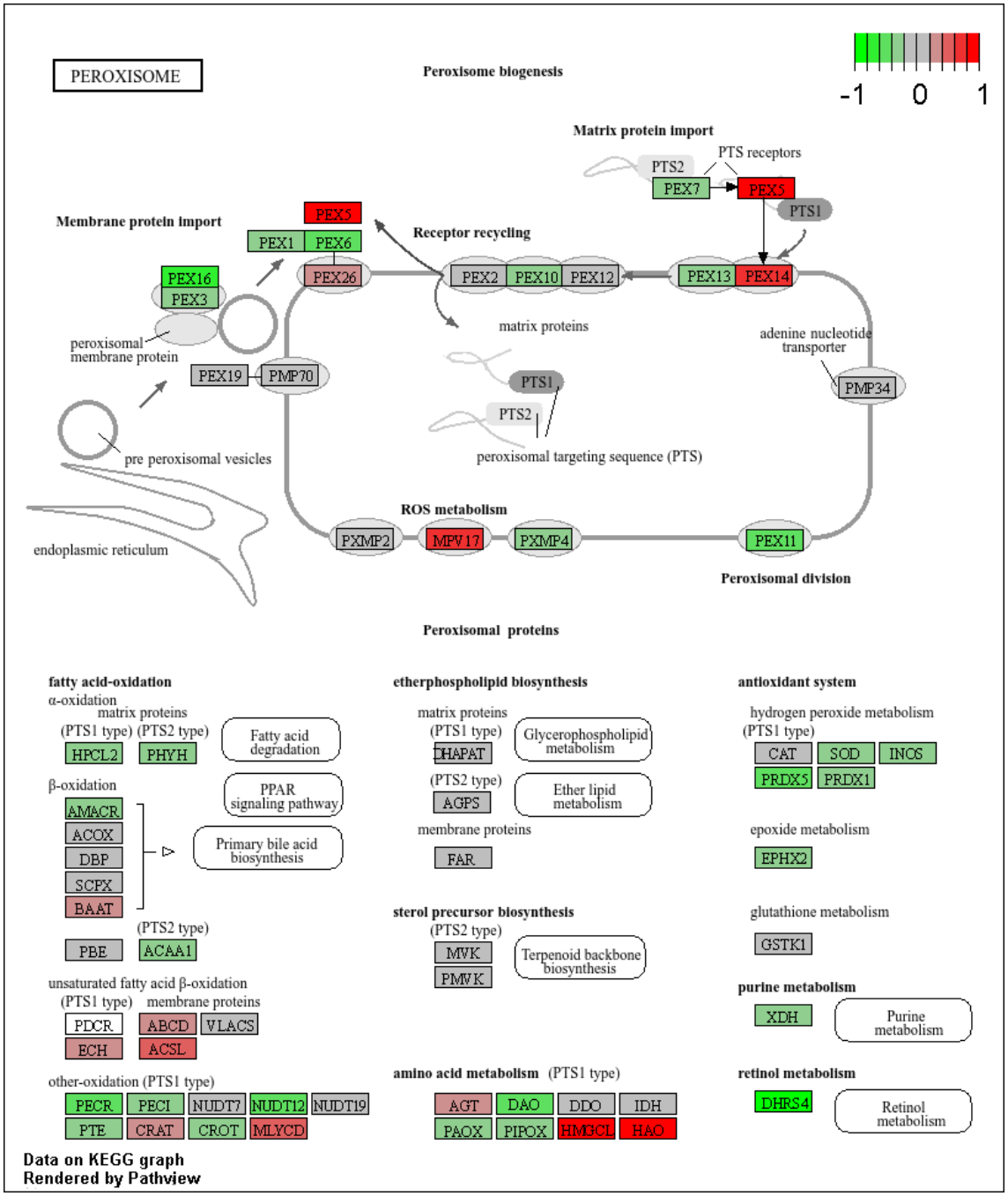


Figura 44: **Vía de peroxisomas en pacientes con insuficiencia cardíaca.** Gráfico de la vía de los peroxisomas con genes sobreexpresados (en rojo) y subexpresados (en verde) en el análisis de enriquecimiento de grupo de genes realizado con GAGE con anotaciones de la Enciclopedia de Genes y Genomas de Kioto para los pacientes con insuficiencia cardíaca frente a controles. q = 0.14. stat mean = -2.98.

se encuentra un grupo muy pequeño de pacientes que se separan en un tercer *cluster*. Al tratar de encontrar una variable conocida que explique la separación de estos participantes, ni edad, sexo, tabaquismo, consumo de oxígeno, IMC, ni la clasificación de la NYHA al que pertenece cada paciente lograron diferenciarse en los *clústeres* formados. De hecho, el único paciente en el grupo 4 de la NYHA, el de presentación más grave de la enfermedad, se encuentra en el centro de la nube mayor, en el límite entre controles y pacientes. También se descartó un efecto del lote debido a las tandas de secuenciación. Por lo tanto, no se logró encontrar el motivo por el cual esos grupos de pacientes o controles se diferenciaron en mayor medida de su grupo mayor de pacientes o controles respectivamente.

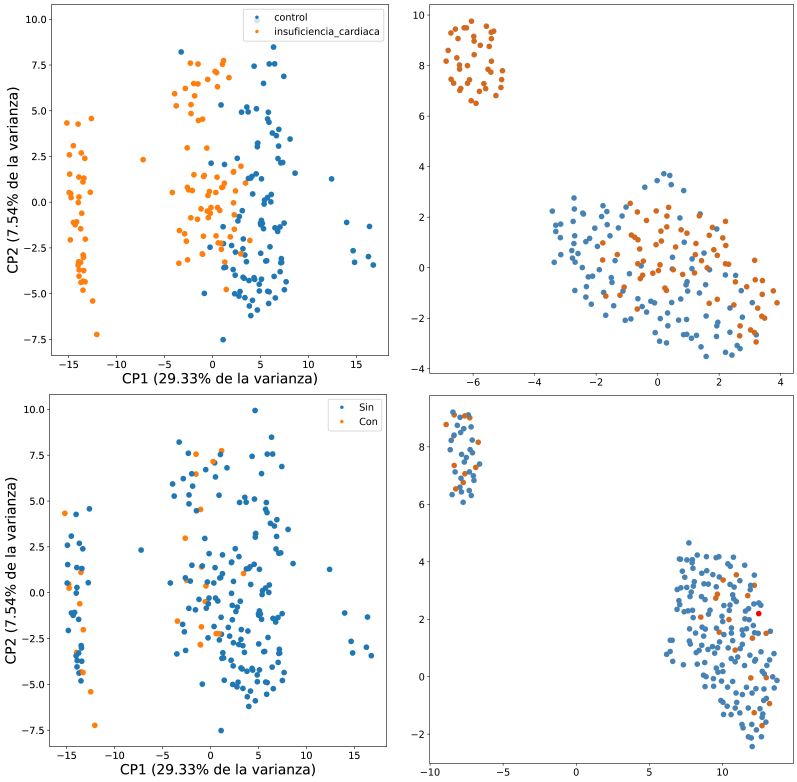


Figura 45: **Agrupamientos no supervisados en pacientes con insuficiencia cardíaca.** Izq: Gráfico de análisis de los componentes principales (PCA). Der: Uniform Manifold Approximation and Projection (UMAP) distancia mínima = 0.7, número de vecinos = 30, épocas = 500. Arriba: Gráficos para la condición en estudio, insuficiencia cardíaca vs control. En ambos gráficos se diferencian los controles en azul y los pacientes con insuficiencia cardíaca en naranja. Abajo: Gráficos de PCA y UMAP de insuficiencia cardíaca vs controles coloreados con datos de seguimiento. Con eventos cardíacos o muerte en los siguientes dos años de tomada la muestra en naranja y sin eventos en azul para ambos gráficos.

El gráfico de UMAP, en la parte derecha y arriba en la figura 45, muestra dos agrupamientos que tampoco pudieron ser explicados por las variables conocidas

antes mencionadas. Un *cluster* más pequeño formado sólo por muestras de personas con insuficiencia cardíaca (naranja) que son más semejantes entre sí que el *cluster* más grande formado por pacientes y controles (azul). Aunque en este grupo mayor se observan algo ordenadas, donde logra colocar las muestras de personas con insuficiencia cardíaca más próximas entre sí y las personas sin insuficiencia también más próximas entre sí.

Para la cohorte con insuficiencia cardíaca se contó con datos parciales de seguimiento (de dos años a partir de la toma de muestra) al momento de la realización de este análisis. Como se aprecia en la figura 45, en la línea de abajo, los eventos cardíacos y las muertes reportadas tampoco lograron explicar los grupos formados en los gráficos de PCA y UMAP.

7.6.0.3. Redes neuronales

El modelo se entrenó durante 30 épocas utilizando un conjunto de datos compuesto por 172 imágenes de transcriptomas de pacientes con insuficiencia cardíaca y controles. Para la validación se emplearon 43 imágenes adicionales. En la figura 46 a la izquierda, se muestra el gráfico de la función de pérdida o error por lote de 4 imágenes a lo largo de 100 épocas de entrenamiento. Este gráfico permite observar cómo la función de pérdida disminuye progresivamente a medida que avanza el entrenamiento, indicando que el modelo está aprendiendo a clasificar de manera más precisa las imágenes de los transcriptomas. La disminución constante de la pérdida sugiere que el modelo está mejorando su capacidad de generalización y reduciendo el error de clasificación. Se decidió entrenar hasta 30 épocas ya que en los modelos con una buena cantidad de imágenes se consiguen buenos resultados antes de un posible sobreajuste, lo que permite una generalización entre diferentes análisis.

A la derecha de la figura 46, se presenta la matriz de confusión resultante de una de las validaciones del modelo ResNet50. Esta matriz de confusión proporciona una visión del rendimiento del modelo al clasificar las imágenes de los transcriptomas en las categorías de sujetos con insuficiencia cardíaca y controles. La matriz muestra el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, lo que permite evaluar la precisión y la capacidad del modelo para distinguir entre las dos clases. El modelo alcanzó una exactitud de clasificación de 0.930 ($\pm$0.028) entre las 5 corridas realizadas. En un análisis

donde la señal es fuerte y diversa, el modelo alcanzó un alto nivel de precisión en la clasificación de las imágenes.

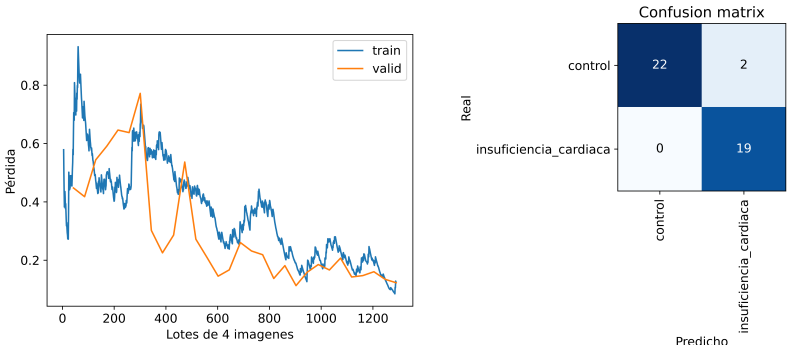


Figura 46: **Análisis por redes neuronales en pacientes con insuficiencia cardíaca.** Izq.: Gráfico de la función de pérdida o error por lote de 4 imágenes para 100 épocas en el entrenamiento para clasificar entre participantes con insuficiencia cardíaca y controles para el estudio de insuficiencia cardíaca y sujetos sin enfermedad activa. Der.: Matriz de confusión resultado de una de las validaciones de ResNet50 para las imágenes correspondientes a los transcriptomas de sujetos con insuficiencia cardíaca y controles.

Discusión

Durante el desarrollo de esta tesis se buscó la aplicación de diferentes técnicas analíticas de grandes datos sobre el transcriptoma de sangre entera, con particular interés en las redes neuronales. De tener éxito, el desarrollo de esta metodología puede conducir a métodos de diagnóstico superadores. El transcriptoma de un tejido presenta características únicas para el desarrollo de una tecnología diagnóstica ya que el ARN es una molécula fundamental en la conexión entre el genoma (nuestra herencia) y el medio ambiente que lo modula.

El trabajo de esta tesis se basó principalmente en los hallazgos iniciales de tres estudios clínicos, incluyendo uno en sujetos sanos, otro en sujetos sanos con posibilidad de sufrir enfermedad coronaria y un tercero en pacientes con insuficiencia cardíaca. Los trabajos fueron realizados en cumplimiento de los estándares internacionales para asegurar la calidad de los mismos y la seguridad de los sujetos participantes. Las muestras de sangre obtenidas fueron procesadas y secuenciadas con metodologías de última generación. Finalmente, el procesamiento bioinformático aplicado fue elaborado para obtener la expresión génica con el menor ruido posible. Todo este trabajo asegura la fiabilidad de los datos obtenidos.

A pesar de esta ventaja biológica, varios problemas se pueden mencionar como posibles dificultades para lograr nuestro objetivo. En primer lugar, existe un gran desconocimiento acerca de qué es exactamente el transcriptoma y cómo procesarlo. En los últimos años se han descrito nuevas formas de ARNs debido a un extenso post-procesado luego de su transcripción. Es más, muchos de estos cambios son ahora llamados epitranscriptoma en analogía al epigenoma y son, al momento, muy poco comprendidos. Lo único claro es que aumentan enormemente la variedad del transcriptoma y que los mismos tienen implicancias funcionales (Bove 2024, Cerneckis 2024). Pero esta variedad genera otro problema para los algoritmos de grandes datos: las bases de datos generadas presentan una enorme variedad entre muestra y muestra. Para estos algoritmos, el ruido para detectar señales puede conducir a pobres resultados. Una posibilidad a futuro es aumentar el

número de muestras para generar entrenamientos más robustos. Alternativamente, otra posibilidad es aumentar el post-procesado de las muestras para reducir el universo analítico. Ambas opciones, sin embargo, no están exentas de problemas.

Otro inconveniente de la metodología desarrollada es el ARN en sí mismo. Es bien conocido que el ARN es una molécula altamente inestable ante la presencia extendida y universal de enzimas que lo degradan, así como también la rápida degradación y síntesis de los mismos al presentar una tasa de recambio de sólo 20 minutos. Para evitar estos inconvenientes, se decidió aplicar en los estudios clínicos la técnica de extraer la sangre e inmediatamente colocarlas en un *buffer* estabilizador de ARN. La gran mayoría de los desarrollos de estudios de ARN en sangre periférica utilizan preparados celulares obtenidos por métodos de separación. De esta manera, al realizar los estudios de expresión génica sobre poblaciones celulares definidas, como por ejemplo monocitos, neutrófilos o simplemente la fracción mononuclear de la sangre periférica (conocida como PBMC), pueden obtener una mejor definición del transcriptoma de las mismas y reducir el ruido de la presencia de una mezcla de células. Sin embargo, consideramos que esta metodología no es aplicable en campo, ya que la traslación a la clínica de las metodologías de laboratorio conlleva una degradación de las capacidades y consistencias. Invariablemente, introducir deficiencias en la toma y el procesado de muestras conllevaría a una tasa de ruido mayor cuando estas técnicas sean aplicadas en el campo clínico.

En la tabla 10 se presentan los resultados de los entrenamientos de todos los estudios. En términos generales, los modelos de clasificación demostraron un rendimiento variable dependiendo de la tarea específica. Los modelos para clasificar entre sujetos femeninos y masculinos, así como entre jóvenes y adultos mayores, mostraron los mejores resultados, como era esperable, con funciones de pérdida bajas y altas exactitudes. La clasificación entre insuficiencia cardíaca y controles obtuvo resultados semejantes a sexo y edad lo que sugiere diferencias más marcadas y consistentes en los transcriptomas para esta patología.

El tercio medio de la tabla comprendido por el análisis del índice de masa corporal, hemoglobina glicosilada y proteína C reactiva obtuvieron resultados alrededor del 80 % de exactitud. Estos muy buenos resultados van en línea con el uso de estos biomarcadores en el uso clínico habitual. Tienen una señal metabólica clara que se refleja en los transcriptomas.

El tamaño y la calidad del conjunto de datos de entrenamiento y validación jugaron un papel importante en el rendimiento de los modelos. Los modelos en-

	FPE	**FPV**	**Exactitud**	**Épocas**	**Tiempo (ms)**
Femenino vs. masculino	0.120 (± 0.022)	0.040 (± 0.028)	0.990 (± 0.007)	30	1:54 (± 0)
Jóven vs. adulto mayor	0.181 (± 0.094)	0.094 (± 0.015)	0.980 (± 0.011)	30	0:37 (± 0)
Insuficiencia cardíaca vs. control	0.244 (± 0.113)	0.228 (± 0.024)	0.930 (± 0.028)	30	0:40 (± 0)
Índice de masa corporal (Obesidad)	0.269 (± 0.074)	0.701 (± 0.122)	0.793 (± 0.033)	30	0:50 (± 0.0003)
BBK - hemoglobina glicosilada (Prediabetes)	0.706 (± 0.061)	0.499 (± 0.281)	0.880 (± 0.045)	8	0:10 (± 0)
BBK - proteína C reactiva (Inflamación)	0.218 (± 0.055)	0.558 (± 0.157)	0.771 (± 0.089)	20	0:33 (± 0)
Aterosclerosis vs. control	0.502 (± 0.260)	0.838 (± 0.217)	0.733 (± 0.066)	12	0:25 (± 0)
BBK - colesterol total (Dislipemia)	0.440 (± 0.244)	0.900 (± 0.153)	0.634 (± 0.088)	25	0:44 (± 0)

FPE: Función de Pérdida de Entrenamiento
FPV: Función de Pérdida de Validación

Tabla 10: Métricas de rendimiento para diferentes clasificaciones

trenados con un mayor número de imágenes generalmente mostraron un mejor rendimiento, lo que destaca la importancia de disponer de grandes conjuntos de datos para entrenar modelos de clasificación robustos. La hemoglobina glicosilada fue una excepción, con muy pocos datos (sólo 10 imágenes para validar) no es un resultado robusto y deberían realizarse mayor cantidad de estudios, pero los resultados preliminares son prometedores.

En las líneas finales de la tabla se encuentran los análisis que presentaron mayores desafíos, aterosclerosis y colesterol total. Estos modelos mostraron funciones de pérdida más altas y exactitudes más bajas, indicando que estas tareas son más complejas y que los modelos tienen más dificultades para distinguir entre las diferentes clases. Esto puede deberse a la naturaleza más sutil y variada de las diferencias en los transcriptomas asociados con la aterosclerosis. En el caso de la dislipemia refleja el uso coordinado de diversos marcadores como las lipoproteínas de alta y baja densidad y los triglicéridos para un diagnóstico de dislipemia. Por lo que un sólo marcador tampoco parece reflejar una señal clara del metabolismo de lípidos en el transcriptoma.

La exactitud de validación y la función de pérdida en validación son indicadores clave de la capacidad de generalización de los modelos. Los modelos que mostraron una menor diferencia entre la función de pérdida en entrenamiento y validación tuvieron una mejor capacidad de generalización, lo que es un requisito crítico para su aplicación clínica real.

El tiempo de entrenamiento por época varió entre los diferentes modelos, reflejando la complejidad de las tareas y el tamaño del set de entrenamiento. Los modelos con tiempos de entrenamiento más largos generalmente correspondieron a tareas más complejas y conjuntos de datos más grandes. Sin embargo, es importante equilibrar el tiempo de entrenamiento con la precisión del modelo para asegurar que los recursos computacionales se utilicen de manera eficiente. El modelo ResNet50 se comportó bien en este aspecto, ResNet101 conseguía valores similares de exactitud con mayores tiempos de entrenamiento. A su vez, ResNet18 conseguía entrenamientos más rápidos, pero que no lograban alcanzar los valores de exactitud de los modelos con mayor número de capas.

El análisis de componentes principales es utilizado de rutina en el análisis de datos transcriptómicos para identificar outliers y tener un panorama general inicial resumiendo la enorme cantidad de datos en 2 o 3 dimensiones. UMAP, aunque más usado en transcriptómicas de célula única, también es muy utilizado para reducir

la información de transcriptomas en masa gráficamente. Al realizar los gráficos con los genes diferencialmente expresados según el DEseq2 para cada condición, en general no se lograron agrupamientos bien definidos, a excepción del sexo. En la figura 47 se muestran los gráficos del análisis de componentes principales de todos los análisis realizados. Como era esperable, la señal transcriptómica más fuerte fue la del dimorfismo sexual, entreviéndose aún cuando los genes diferencialmente expresados seleccionados fueron para otra condición. Como fue el caso del índice de masa corporal y la proteína C reactiva. El gráfico del análisis de insuficiencia cardíaca se asemeja más al de comparación entre la población joven frente a los adultos mayores. Parecería que la señal es suficientemente fuerte para generar un ordenamiento en el primer componente, pero no tanto como para lograr una verdadera clusterización. Este también es el caso del gráfico de la proteína C reactiva que, a pesar de clusterizarse por sexo, logra un ordenamiento de los participantes en el segundo componente según tengan el marcador de inflamación alto o no detectable. Otro es el caso del colesterol total que separa a los participantes en dos grupos debido al primer componente con el 84% de la varianza, pero no por la condición de colesterol total alto o bajo, no pudiéndose encontrar ninguna de las variables clínicas medidas como responsable. Por último, en los gráficos de prediabetes y aterosclerosis los participantes aparecen sin ningún tipo de agrupamiento ni ordenamiento en el primer y segundo componente.

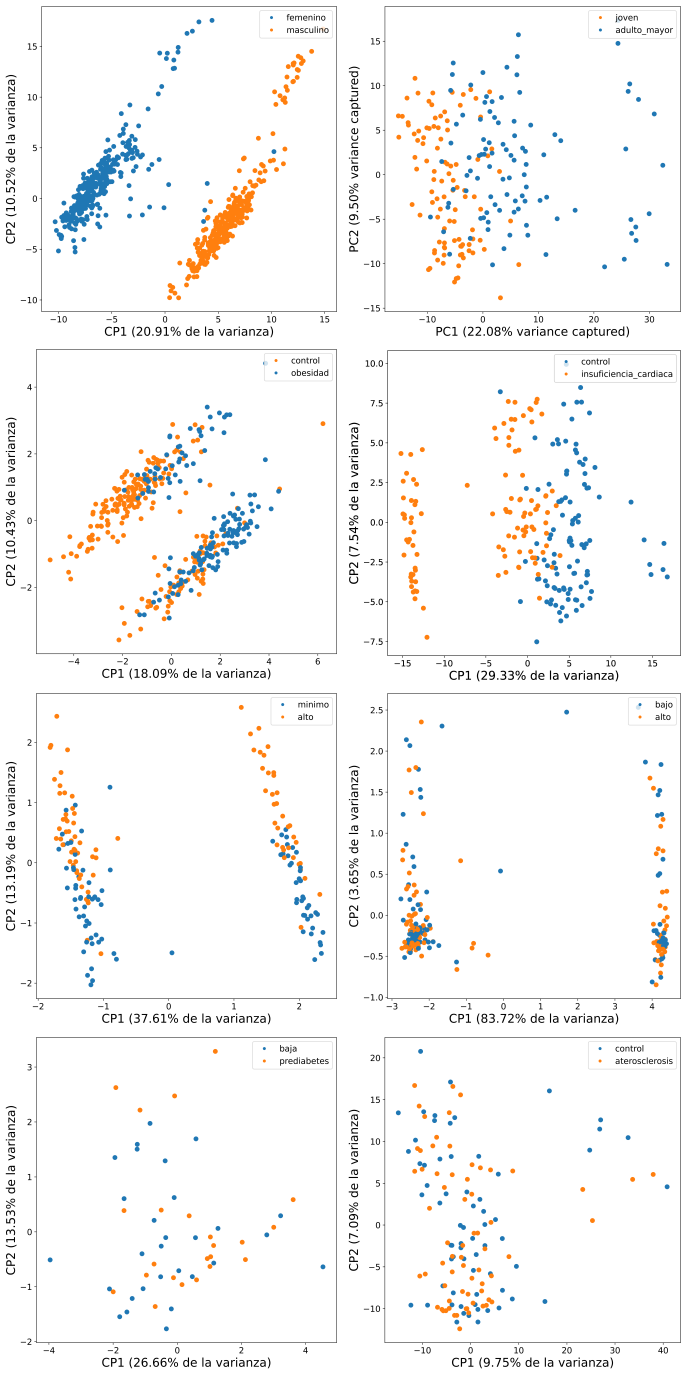


Figura 47: **Análisis de componentes principales de todos los análisis realizados.**

Tratando de paralelizar el desempeño de ambos tipos de modelos, tanto supervisado como no supervisados, siguiendo un razonamiento de captación de señal biológica relevante se encuentra al sexo como la señal más fuerte y, por lo tanto, el mejor desempeño para ambos tipos de modelos. Seguido, en ambos casos, por los análisis de edad e insuficiencia cardíaca. También parecen coincidir en el caso del colesterol total como la señal más confusa con la peor performance en ambos tipos de modelos, con la peor exactitud de clasificación en la red neuronal

y una clusterización muy fuerte sin respetar la condición de estudio en PCA. En el caso de la hemoglobina glicosilada los modelos parecen diferir dado que los no supervisados no encontraron una estructura latente, sin embargo, la red logró una muy buena exactitud de clasificación. Aunque se debe tener en cuenta que son muy pocas imágenes de validación y el resultado no es tan robusto como los otros análisis planteados. Para el análisis de aterosclerosis vs control la red alcanzó un buen desempeño, sin embargo los modelos no supervisados no lograron agrupar a los participantes con aterosclerosis en un grupo diferente de los controles, lo que muestra limitaciones en su capacidad para discernir señales más sutiles.

8.1. Uso de base de datos

Generar una base de datos propia para el entrenamiento de algoritmos de aprendizaje automático fue una gran ventaja, tanto en el caso de los transcriptomas como en el de las micrografías de cultivos celulares.

Los datos a los que se pueden acceder a partir de repositorios públicos de datos de expresión, por ejemplo el Gene Expression Omnibus (GEO), son muy heterogéneos debido a que cada estudio suele tener un número reducido de muestras. La estrategia de muchos trabajos que analizan estos datos es agrupar estudios lo que genera un gran efecto lote. Esta variabilidad se debe a que las bases de datos pueden reunir transcriptomas obtenidos a partir de diferentes metodologías y cada tipo de técnica, ya sea secuenciación masiva o microarreglos, tiene su variabilidad asociada. Además, no contar aún con un protocolo bioinformático estandarizado agrega mayor variación al conteo de expresión génica aunque provengan del mismo tipo de metodología. Tampoco hay un fenotipado clínico común entre países y entre estudios, ya que los criterios de elegibilidad para pacientes y controles pueden ser diferentes, lo que puede llevar a incongruencias en la asignación de etiquetas a las muestras cuando se agrupan los datos.

Una base de datos muy utilizada es Adult Genotype Tissue Expression (GTEx) (GTEx consortium 2013). Es un recurso público que estudia en diversos tejidos humanos la expresión génica y su regulación. El proyecto colectó muestras hasta de 54 tejidos sanos de 946 personas fallecidas, sumando 19788 muestras. Se accedió a los datos de expresión génica de cada tejido mediante una secuenciación de ARN a granel. En la figura 48 se puede observar el análisis de componentes

principales que arrojan los datos de expresión en sangre entera para la base de datos GTEx. A la izquierda de la figura los datos están coloreados según la escala Hardy que clasifica el tipo de muerte de los participantes a los cuales se les extrajo la muestra de sangre para la base de datos. Esta escala agrupa los tipos de muerte de la siguiente manera:

0. Respirador: Todos los casos en los que se utiliza un respirador inmediatamente antes de la muerte.

1. Muerte violenta rápida: Debida a un accidente, suicidio o el impacto de un golpe no penetrante.

2. Muerte natural rápida: Súbita y no esperada de personas razonablemente sanas con una fase terminal menor a 1 hora. El infarto de miocardio es un ejemplo modelo de esta categoría.

3. Intermedia: Luego de una fase terminal entre 1 y 24 horas que no puedan clasificarse como 2 o 4. Son pacientes enfermos, pero su muerte no es esperada.

4. Muerte lenta: Luego de una larga enfermedad, con fase terminal mayor a 24 horas, son muertes esperables, como el caso del cáncer o enfermedades respiratorias.

A la derecha de la figura 48 se observa el mismo análisis destacando con diferentes colores los diferentes tiempos en que el tejido dejó de recibir oxígeno y nutrientes antes de tomar la muestra. Tanto en el gráfico de escala Hardy como en el de isquemia, se observan clústeres que separan a los participantes del estudio, explicando en el primer componente la varianza que, ni el sexo, ni la edad logran explicar (Esto se puede observar en la fila inferior de la figura 48). Estudios indican que en la muerte súbita hay genes diferencialmente expresados que, al realizar los análisis de enriquecimiento, se asociaron a aterosclerosis, enfermedad cardiovascular, enfermedad renal e infarto de miocardio (Zhou 2022). Por lo tanto, evitar las diferencias de expresión que trae aparejada la isquemia de la toma de muestra post mortem, como así también el tipo de deceso, es una ventaja importante al analizar los datos de una base propia y, más aún, en el estudio de enfermedad cardiovascular dada la conocida diferencia de expresión génica en la muerte súbita. En nuestra metodología de toma de muestra y procesamiento la sangre entra

en contacto inmediatamente con el buffer en el tubo de extracción. Este buffer lisa las células y vesículas sanguíneas, y estabiliza el RNA intracelular evitando la degradación o modificación del mismo. Además, se evita la síntesis de nuevos ARNs. De esta manera se evitan los problemas encontrados en GTEx. Otros estudios con muestras de sangre realizan una separación y selección de los elementos celulares o utilizan plasma o suero. Cualquiera de estas opciones significa tiempos de procesamiento largos y complejos. Además, debido a la alta inestabilidad del ARN, proporcionan ventanas en las cuales el transcriptoma se puede alterar significativamente.

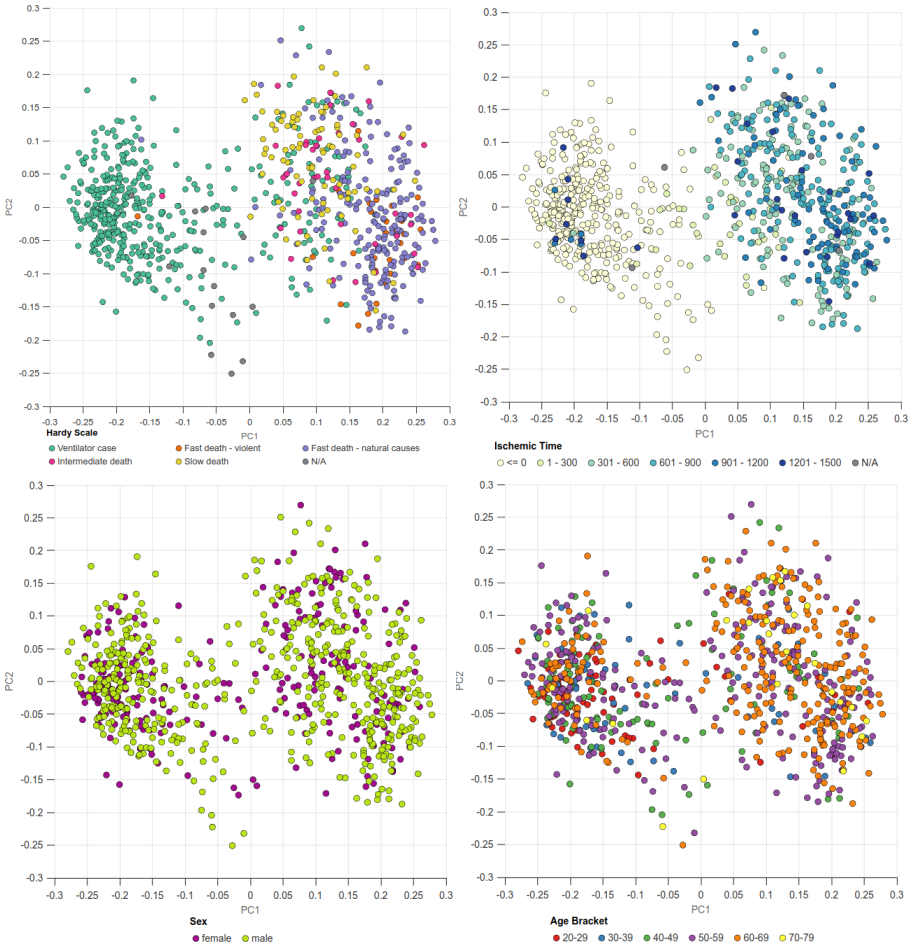


Figura 48: **Análisis de componentes principales de los datos de GTEx de sangre entera.** Gráficos de análisis de componentes principales de los datos de GTEx de sangre entera. Izquierda arriba: Escala de Hardy (0: Respirador, 1: Muerte violenta rápida, 2: Muerte natural rápida, 3: Intermedia, 4: Muerte lenta). Derecha arriba: Tiempo de isquemia en minutos. Izquierda abajo: Sexo. Derecha abajo: Edad categorizada en decenas.

Otra ventaja de generar una base de datos propia recae en el desempeño de las ecuaciones de riesgo poligénico ya que se ve afectado por las características de la población en la cual se utiliza. Múltiples estudios señalan la necesidad de cali-

brar el cálculo de riesgo según la cohorte y la ancestría genética (Cook 2016). En el mismo camino, una proporción significativa de la variabilidad en la expresión es heredable (Skelly 2009) entonces, es valioso contar con datos de expresión de nuestra población. Es importante destacar que la población argentina tiene poca representatividad en los datos públicos disponibles. En los países desarrollados se destina una gran cantidad de fondos de los Estados como un esfuerzo para acelerar las estrategias de medicina de precisión para la prevención, el diagnóstico y el tratamiento de las enfermedades. Tal es el caso del Programa de Investigación Todos Nosotros (All of Us Research Program Genomics Investigators 2024) en Estados Unidos, en el cual los Institutos Nacionales de Salud (NIH) invirtieron 130 millones de dólares para reclutar a más de 1 millón de personas que den su consentimiento para recabar información digital de su salud, datos biométricos y biológicos. Además, en el reclutamiento pusieron especial atención en la diversidad, incluyendo poblaciones históricamente subrepresentadas, para entender las disparidades en el manejo de la salud y que los resultados sean aplicables a todas las poblaciones de manera más personalizada y equitativa. Sin embargo, en los países en vías de desarrollo como la Argentina los fondos invertidos en medicina de precisión son mucho más escasos y al momento no existen iniciativas similares de tal magnitud. En este trabajo no se tuvo especial cuidado en la ancestría de los participantes de los estudios clínicos, ni pretendió ser representativa del país en su totalidad (Luisi 2020), ya que los participantes fueron reclutados exclusivamente en la ciudad y la provincia de Buenos Aires (AMBA zona norte). Aunque, sin duda, es una mejor aproximación a la representatividad de la población argentina que las bases de datos internacionales asequibles al momento.

Una dificultad titánica que afrontan las enfermedades poligénicas es despejar su variabilidad inherente de la señal biológica para lograr un modelo que permita predecir. Los resultados de los algoritmos de aprendizaje automático se limitan a la información disponible en los datos de entrenamiento, por lo tanto, no solo se debe priorizar la cantidad sino la calidad de los datos para su entrenamiento. La información debe ser abarcativa, representativa y con un fenotipado estandarizado para sacar el máximo provecho de estas técnicas. Entonces, los estudios clínicos permitieron lograr datos controlados desde su obtención, su procesamiento y su análisis ayudando a disminuir la variabilidad no deseada. En este trabajo se tuvo en cuenta un criterio unificado de fenotipado para un objetivo común a través de los tres estudios clínicos para lograr abarcar la enfermedad cardiovascular en todo

su espectro clínico.

Este somero compendio resalta la importancia de recabar información propia para el entrenamiento de una red neuronal.

Una mención aparte merece el origen de la muestra de los estudios clínicos. Por un lado, la sangre es un tejido accesible y su obtención es mínimamente invasiva, lo que facilita su uso en la práctica clínica y permite la recolección de muestras de manera repetida para el monitoreo de enfermedades. Por otro, gracias a trabajos previos que validaron la exactitud diagnóstica de estudios de expresión génica para enfermedad obstructiva coronaria (Rosenberg 2010) y a los resultados obtenidos en este trabajo, se puede afirmar que la sangre es un sustrato ideal para el sondeo poblacional, ya que refleja el estado fisiológico general del organismo y puede proporcionar información sobre procesos patológicos en diversos órganos y sistemas. Además, el uso de muestras de sangre facilita la estandarización de protocolos en la aplicación clínica. Por lo que podría plantearse la expansión de la red a otras patologías de interés con las cuales pueda entrenarse.

8.2. Agrupamientos no supervisados

Los resultados de los agrupamientos no supervisados de los participantes no fueron satisfactorios para los análisis planteados. Los métodos utilizados de reducción de la dimensionalidad no lograron encontrar una estructuración tal en los datos transcriptómicos que logre separar a los participantes con las condiciones planteadas de los controles. Estos modelos tuvieron limitaciones en la capacidad de discernir cuando las señales eran sutiles.

Al correlacionar los datos clínicos disponibles tampoco se pudieron explicar los agrupamientos hallados, a excepción del sexo que sí explicó en algunos casos los cluster que no respondían a la condición en estudio.

Los análisis de ontología génica de los genes expresados diferencialmente resultaron más enriquecedores. En general respondieron a las vías biológicas involucradas directa o indirectamente con la patología o factores de riesgo planteados en los análisis. El GO se desempeñó mejor, logrando encontrar vías significativas con mayor relevancia biológica, cuando la cantidad de genes diferencialmente expresados fue mayor. Por el contrario, cuando el DEseq2 hallaba pocos genes (decenas) diferencialmente expresados, la estrategia de enriquecimiento de genes de GAGE

logró contextualizar biológicamente mejor, con vías más apropiadas según la patología o factor de riesgo planteado. En el rango de los cientos de genes ambas estrategias lograron complementarse con varias coincidencias.

Aún con los avances en los métodos que se utilizan para dar un contexto biológico a la inmensa información transcriptómica todavía es un proceso dificultoso y dependiente de la herramienta utilizada. La regulación génica se ejerce mediante una red de genes que interactúan (Ding 2020), las relaciones entre estos genes es compleja y dista mucho de ser lineal, esta falta de linealidad imposibilita a la mayoría de los métodos disponibles para estudiar la expresión génica basados en regresión o correlación. Las listas de genes, los clústeres o los análisis de enriquecimiento de genes no abarcan la totalidad de la complejidad (Diaz 2020). Además, los genes que interactúan con genes desregulados sin estar ellos mismos expresados diferencialmente no son visibles en los estudios de expresión diferencial tradicionales (Magnusson 2022).

8.3. Redes Neuronales

En los estudios clínicos presentados se trató de cubrir el amplio espectro de la enfermedad cardiovascular para poder aplicar la plataforma planteada en el objetivo principal en cada una de las etapas en curso de la enfermedad. Es claro que el desbalance metabólico va a ser mayor en una persona con insuficiencia cardíaca que en una que esté en los primeros estadíos de la enfermedad cardiovascular, por lo tanto, es importante, y muy alentador, que la clasificación tenga una buena exactitud de clasificación a etapas tempranas, cuando no hay síntomas de enfermedad cardiovascular. En la etapa de insuficiencia cardíaca las redes pueden plantearse como una herramienta para la prognosis, una subclasificación dentro de la patología más que una clasificación frente a controles. Para esto se utilizarán los resultados del seguimiento a 5 años que se está realizando a la cohorte del estudio de insuficiencia cardiaca. Con los datos preliminares de 2 años de seguimiento no es posible aún tener un número suficiente de pacientes con eventos cardiovasculares como para lograr un entrenamiento de la red. Con estos resultados solamente se identificaron los pacientes en los modelos no supervisados utilizados (PCA y UMAP), pero no se logró ninguna clasificación relevante, como se vio en los resultados.

En el desarrollo de la plataforma también se probaron redes neuronales lineales o totalmente conectadas a partir de los datos de expresión brindados directamente a partir de la matriz de expresión génica (No se muestran los datos). Se probaron diferentes arquitecturas con una variedad de cantidad de capas con tamaños diferentes, pero no se logró convergencia al mantener la cantidad de genes totales originales. La función de pérdida siempre se disparó a miles en estos casos. Como se vio en trabajos anteriores (Miao 2024) para la utilización de redes lineales existe un filtrado de genes previo que puede sesgar el análisis.

Es importante recalcar que al resumir la información de la expresión de todos los genes en la imágen planteada en esta tesis, la red convolucional fue capaz de resolver diferentes clústeres de datos sin necesidad de un filtrado previo de genes, que es la manera en la que usualmente se afronta el problema de tener mucha menor cantidad de muestras que el número de genes (Kakati 2022). Además se logró tener evidencia que la estrategia de resumir la información transcriptómica en imágenes también es válida para clasificar pacientes con enfermedades cardiovasculares a partir de una muestra de sangre periférica, ya que los trabajos hasta el momento fueron realizados sobre muestras de tumores de pacientes (Lyu 2018, Ma 2018), en los cuales las diferencias en la transcripción son mucho mayores al comparar tejidos diferentes. La particularidad de lograr entrenar a la red con un resumen de toda la expresión génica más que con unos pocos genes preseleccionados conlleva la ventaja de aportar la información del sexo y la edad fisiológica, más que estrictamente cronológica, del individuo que es sabido que correlacionan con la enfermedad cardiovascular. Sumado esto a la información de su sistema inmune y los matices de expresión resultantes de la presión que ejerce el ambiente al que está expuesto el individuo que se deja traslucir en una muestra de sangre entera.

El mayor desafío para la incorporación de las redes neuronales para el cálculo del riesgo cardiovascular en la población es el escepticismo de la comunidad científica frente a una metodología de caja negra. Los conceptos de interpretabilidad y explicabilidad, usados comúnmente como sinónimos, tienen un matiz en el uso de los modelos de aprendizaje automático. La interpretabilidad está ligada a la capacidad para explicar a un ser humano los resultados de un modelo, mientras que la explicabilidad está asociada con la comprensión de la lógica interna del algoritmo, es la medida en que la mecánica interna de un sistema de aprendizaje automático se puede explicar en términos humanos (Lage 2019). En los modelos

muy complejos, como las redes neuronales con muchos nodos, la explicabilidad es aún un reto no resuelto. Sin embargo, se han desarrollado muchas técnicas para mejorar la interpretabilidad. Igualmente, la comunidad médica podría no querer utilizar una herramienta sin comprender el mecanismo exacto de funcionamiento.

Otro desafío para la aplicabilidad de las redes neuronales artificiales es que estas pueden ser entrenadas para adaptarse a variados escenarios, pero, luego del entrenamiento, pueden no generalizar bien en escenarios desconocidos (Cheng 2024). Para un profesional de la salud no familiarizado con la optimización de modelos de aprendizaje automático la herramienta puede entregar peores resultados que las ecuaciones tradicionales de riesgo. Para atacar este problema se debe alcanzar un entrenamiento amplio y representativo, con una gran cantidad de datos y una interfaz amigable para el operador. En la actualidad se está llevando a cabo otro estudio con 800 participantes para la evaluación de calcio coronario con el cual se podrá testear la plataforma y ampliar el entrenamiento con mayor cantidad de muestras de la población local. Paradójicamente, una forma de mejorar el rendimiento del modelo sería la propia implementación. Cuando una red neuronal se implementa en una plataforma que permite la incorporación dinámica de nuevos datos operativos, se establece un ciclo de retroalimentación que potencia su capacidad predictiva de forma iterativa. Este proceso, conocido como aprendizaje incremental adaptativo, permite que el modelo evolucione y se adapte a patrones emergentes.

8.4. Contexto biológico de los hallazgos

Dentro de los objetivos planteados en esta tesis se encuentra correlacionar los hallazgos de la clasificación con las características clínicas y la evolución de los pacientes. Las vías metabólicas que se encontraron desreguladas, en general, coincidieron con lo esperado por los antecedentes bibliográficos. Sin embargo, los cambios en la expresión génica observados son efectos de correlación con la enfermedad cardiovascular que pueden ser debido tanto a causas de la patología como a efectos de respuesta a ella, reflejando un riesgo general debido a la enfermedad y a la actividad inflamatoria de cada participante del estudio. Entonces, los hallazgos de los genes diferencialmente expresados deben ser validados, abriendo líneas de investigación interesantes para ensayos funcionales de proteínas y análisis de

mecanismos de regulación génica dentro de la enfermedad cardiovascular.

Con la mirada puesta en la traslación de la plataforma a la práctica clínica existen, al menos, dos caminos para plantear un score de riesgo coronario. Un método más indirecto es la clasificación de las personas según su score de calcio coronario. Como vimos, el calcio correlaciona muy bien con el riesgo a 10 años de padecer un evento cardiovascular. Clasificando el score de riesgo coronario a partir del agregado de la transcriptómica al análisis de sangre de rutina, se obtendría indirectamente el riesgo de evento cardiovascular. Un testeo de la expresión génica en sangre periférica tiene mayores ventajas clínicas respecto a otros ensayos no invasivos, ya que requiere solo de una extracción sanguínea estándar, sin necesidad de radiaciones ionizantes, contrastes intravenosos ni estresores farmacológicos ni psicológicos (Rosenberg 2010). Sin embargo, desde una perspectiva molecular, otros tejidos involucrados como el músculo liso, el endotelio o el hígado, por mencionar algunos, podrían brindar información complementaria de productos inflamatorios que no sean detectados en la sangre. Además este enfoque no responde la pregunta de por qué algunas personas con varios factores de riesgo y un estilo de vida desfavorable cardiovascularmente hablando llegan a edades avanzadas. O cuál es el motivo por el cual personas con estilos de vida saludables y sin factores de riesgo se enfrentan a eventos cardiovasculares tempranos.

El otro método propuesto debe seguir su estudio en un futuro cercano. Con los datos de seguimiento a 5 años de los pacientes se podrá investigar si la plataforma puede clasificar a los participantes según su evolución. El planteo más interesante sería replantear los entrenamientos enfrentando a las personas que tuvieron eventos cardiovasculares dentro de los 5 años de tomada la muestra de sangre y a las que no los tuvieron. De esta manera analizar si la plataforma es capaz de detectar cambios tempranos en la expresión de los genes que puedan predecir los eventos cardiovasculares con una ventana de tiempo que le permita al médico la intervención temprana del tratamiento. Esta estrategia encontraría directamente el riesgo de padecer el evento independientemente de la medición previa de los factores de riesgo. Mientras para algunos individuos la calcificación puede ser la causa principal de su elevada susceptibilidad, para otros puede ser la inflamación crónica o una falta de homeostasis de lípidos hereditaria (Biros 2008). El estilo de vida y todas las variables clínicas tradicionales estarían ya contempladas y pesadas por el valor de su responsabilidad en el desarrollo del evento. Igualmente comparte la desventaja de la falta de información complementaria de productos inflamatorios

en otros tejidos que no sean detectados en la sangre ya que, en ambos casos, la muestra parte del tejido sanguíneo.

Hallar a las personas que van a sufrir un evento cardiovascular 5 años antes del suceso con mayor exactitud que los métodos tradicionales sería, sin duda, un arma poderosa para el médico en la lucha en el acompañamiento clínico de sus pacientes. Además de los beneficios personalizados, epidemiológicamente se podrían redireccionar recursos económicos y humanos para un manejo poblacional de la enfermedad cardiovascular más eficiente.

Las herramientas existentes para la evaluación de riesgo cardiovascular no son universales debido a diferencias genéticas, culturales y socioeconómicas de cada población. En la población argentina no se han realizado estudios de cohortes prospectivas para la validación de los métodos de evaluación de riesgo cardiovascular que se utilizan en países desarrollados. En entornos de atención sanitaria de bajos recursos, cuando algunos de los factores de riesgo no se encuentran disponibles, se pueden utilizar los cuadros de la Organización Mundial de la Salud (Figura 5, página 23). Los cuadros categorizan incorrectamente a muchos individuos en el grupo de bajo riesgo, lo que lleva al subtratamiento de la población, a subsecuentes complicaciones y, finalmente, a un mayor gasto del sistema de salud. Por el contrario, las ecuaciones implementadas en países con alta inversión sanitaria requieren un mayor desembolso inicial para screenings, exámenes y un sistema informático apropiado para una evaluación de riesgo más precisa (Badawy 2022). La continua baja en el costo de secuenciación abre una oportunidad para desarrollar una herramienta de evaluación de riesgo cardiovascular que contemple al transcriptoma y tenga una buena relación costo-efectividad para nuestro país.

8.5. Perspectivas a futuro

La implementación de las tecnologías ómicas en la medicina de precisión es un gran desafío por diversas causas (Babu 2023), pero el esfuerzo de su integración a diferentes niveles de complejidad es sabido que rendirá sus frutos en un futuro algo más lejano. Cada tecnología ómica por separado se ha utilizado con diferentes grados de éxito, pero una gran meta de la medicina es poder implementar una combinación que brinde información superadora. Por ejemplo, combinar la transcriptómica y la proteómica agrega información funcional que no puede capturar

la genómica. Estas combinaciones permiten deshojar la complejidad molecular de las enfermedades mediante una visión holística. En la figura 49 se ejemplifica un flujo de trabajo de tres tecnologías ómicas integradas mediante aprendizaje automático. Se puede observar la formación de un cuerpo tridimensional en el cual cada cara es una imagen que representa los datos ómicos de una persona. Con el objetivo de una integración sinérgica de varias tecnologías de punta en el campo de la cardiología nace el concepto de gemelo digital, tomado de la ingeniería, donde representaciones in silico de un sistema físico, como podría ser un motor, son usadas para optimizar procesos. En sanidad, el gemelo digital denota una herramienta virtual que integra coherente y dinámicamente los datos clínicos del paciente que fueron adquiridos a través del tiempo usando modelos estadísticos, mecanísticos y simulaciones (Corral-Acero 2020). Los modelos mecanísticos engloban, principalmente, el conocimiento sobre fisiología. Por ejemplo, las ecuaciones de Navier-Stokes para el modelado del flujo sanguíneo humano o el modelo de bidominio para la actividad eléctrica del corazón (Leslie 1978). Los modelos estadísticos encapsulan el conocimiento y las relaciones provenientes de los datos. El objetivo es brindar al médico conocimiento individualizado sobre la salud cardiovascular general del paciente. Un beneficio importante para el médico es la capacidad de probar en el gemelo digital terapias cardiovasculares para un paciente en particular y así poder evaluar el desempeño sin poner en riesgo al paciente real (Singh 2024). Ya se obtuvo evidencia del escalado de este concepto en ensayos clínicos in silico (Faris 2017). Un robusto modelado matemático ayudó a reducir los requerimientos de los estudios para la aprobación por la FDA de un marcapasos seguro en el contexto de resonancias magnéticas. La integración mediante inteligencia artificial de todos los datos ómicos, clínicos y de dispositivos de uso diario (*Wearables*) sumado al gran desempeño de las computadoras cuánticas podrán brindar al médico y al científico en el futuro una grandiosa herramienta para la lucha contra las enfermedades.

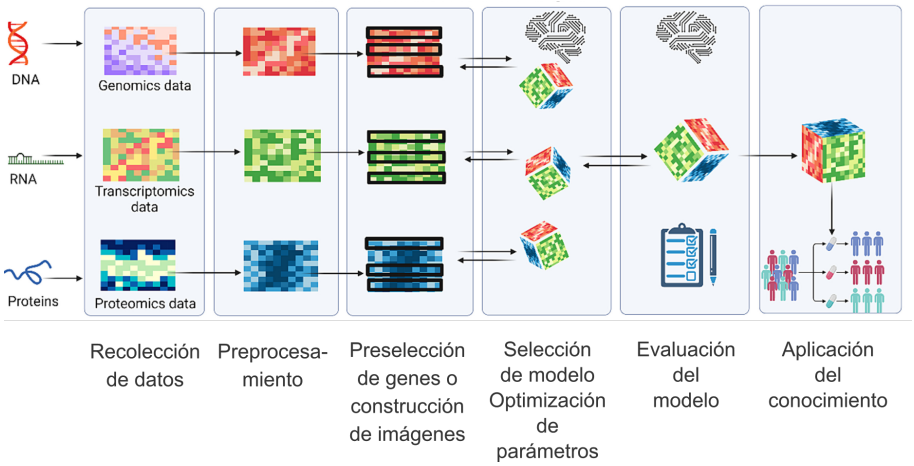
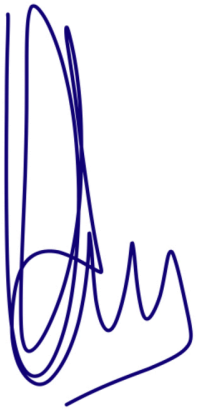


Figura 49: **Esquema del flujo de trabajo de la integración de diferentes tecnologías ómicas.** Esquema del flujo de trabajo de genómica, transcriptómica y proteómica mediante inteligencia artificial. Gráfico tomado de Sopić 2023.

Conclusión

En este trabajo se desarrolló un flujo de trabajo que incluyó un modelo de red neuronal artificial para clasificar datos transcriptómicos de sangre entera de pacientes y controles convertidos en imágenes RGB. La red convolucional residual de 50 capas fue capaz de extraer rasgos de los datos que diferenciaron las características fenotípicas propuestas con diferente grado de exactitud. La evidencia presentada mediante datos propios tomados de estudios clínicos realizados avalan una posible aplicación clínica. Los datos de seguimiento serán cruciales para definir el alcance del test no invasivo que pueda ser implementado. Actualmente el aporte suma evidencia a la posibilidad del uso de la transcriptómica y las redes neuronales en la medicina de precisión.



Miriuka Santiago

Pérez María Nelba

Anexo

Anexo I: Aplicación de redes neuronales en un modelo de muerte celular

Introducción

Parte de la inteligencia artificial desarrollada para esta tesis se gesta en trabajos previos en los cuales se lograron clasificar cultivos celulares, separando los que iban a morir de los que sobrevivían, a partir de imágenes de microscopio invertido. Este trabajo anterior devino en una herramienta de uso libre llamada celldeath accesible en `https://github.com/miriukaLab/celldeath`.

La muerte celular es un evento complejo y muy estudiado que ocurre en procesos fisiológicos y patológicos (D'Arcy, 2019). Es un mecanismo ampliamente utilizado en investigación básica (Kabore 2004; Merino 2018), por lo que se han desarrollado múltiples técnicas para analizar la muerte celular. Todas ellas involucran el estudio de rasgos particulares de la célula en los diferentes estadíos hacia la muerte celular final, incluyendo la fragmentación del ADN, las modificaciones proteicas o la inversión de proteínas de membrana, entre otros (Elmore 2007; Majtnerová 2018; Kay 2019). Estos estudios moleculares se pueden realizar de muchas maneras, e incluyen ensayos basados en microscopía, citometría de flujo o western-blot. Todos ellos, eventualmente, demandan tiempo y dinero.

La camptotecina es un inhibidor de la topoisomerasa I que induce rápidamente una señal en células madre embrionarias humanas que deriva en apoptosis (García 2014) y es un quimioterapéutico ampliamente utilizado en investigación para generar apoptosis en células tumorales (Cheng-Wu 2012). La inhibición de la topoisomerasa I genera cortes en las dos cadenas del ADN (DSB) (Strumberg 2000) que lleva a la fosforilación de H2AX (γH2AX) y a la activación de la proteína supresora de tumores p53 (Sedelnikova 2003; Sordet 2009).

La propuesta fue clasificar cultivos celulares que ya habían definido su vía hacia la muerte celular de los cultivos sanos mediante micrografías de campo claro, con las ventajas que esto trae aparejado: Para obtenerlas no se pierde el cultivo y no se utilizan reactivos ni equipos costosos, sólo un microscopio de uso habitual en cualquier laboratorio, siendo una técnica rápida y económica. Clasificar estas fotografías es imposible para el ojo humano, aún para el científico más entrenado en cultivo celular. Por este motivo se entrenó una red neuronal residual (Kaiming 2015) con micrografías de 7 cultivos celulares a los cuales se indujo la muerte celular o se los mantuvo como control.

Objetivo

Desarrollar un flujo de trabajo con redes neuronales que pueda determinar, por medio de la utilización de redes neuronales, la existencia de muerte celular en cultivos celulares en los estadíos iniciales del proceso.

Materiales y Métodos

Se generaron las etiquetas necesarias para clasificar -muerte celular/control- incubando 7 líneas celulares con camptotecina (CPT) o su vehículo: dimetilsulfóxido (DMSO). Se determinó la concentración óptima de CPT para cada línea celular y se incubaron las líneas celulares control con el volumen correspondiente de DMSO. Se fotografiaron a la hora en un experimento, a las dos horas en otro y a tres en otro para ambas condiciones, control y muerte celular.

Se confirmó en cada caso que las líneas entraron en apoptosis mediante inmunofluorescencia, observando a H2AX fosforilada y la acumulación de p53; y mediante citometría de flujo, identificando la exposición en la membrana plasmática de la fosfatidilserina mediante su interacción con anexina V.

Se utilizaron 4 líneas de células tumorales y 3 líneas de células madre pluripotentes inducidas que se mantuvieron en una atmósfera humidificada y filtrada a 37 grados centígrados y 5 % de CO2. En la tabla A1 se detallan los medios de cultivo utilizados para cada una. Todas las células fueron removidas del plato de cultivo mediante TrypLETM Select 1X (ref. A1217702; Thermo Fisher Scientific, United States) cada 4 ó 5 días dependiendo de su densidad.

Línea	**Medio**
U2OS (Osteosarcoma); MCF7 (Mama epitelio luminal)	Dulbecco's Modified Eagle Medium (ref. 12430054, DMEM; Thermo Fisher Scientific, United States) suplementado con 10 % de suero fetal bovino (NTC-500, FBS; Natocor, Argentina) y 1 % de penicilina/estreptomicina (ref. 15140–122, Pen/Strep; Thermo Fisher Scientific, United States)
PC3 (Próstata); T47D (Mama epitelio luminal)	Roswell Park Memorial Institute medium (ref. 22400089, RPMI; Thermo Fisher Scientific, United States) suplementado con 10 % de suero fetal bovino y 1 % de penicilina/estreptomicina
IPS1, IPS2 (Células madre pluripotentes inducidas); H9 (Célula madre embrionaria)	Se utilizaron platos de cultivos cubiertos en GeltrexTM (ref. A1413302; Thermo Fisher Scientific, United States) y medio definido Essential 8 flex (ref. A2858501, E8 flex; Thermo Fisher Scientific, United States), reemplazado todos los días.

Tabla A1: Detalle de los medios de cultivos utilizados en las 7 líneas celulares mantenidas para obtener las micrografías.

Para inducir la muerte celular, se sembraron aproximadamente $3x10^5$ células en los 4 pocillos centrales de un plato de cultivo de 12 pocillos (ref. 3513; CORNING Inc., United States) y, al día siguiente, se les retiró el suero a las células tumorales por 24 h. Luego se trataron 2 pocillos con camptotecina 1–10μM (ref. C9911, CPT; Sigma-Merck, Argentina) y a los restantes pocillos de control con la misma cantidad de vehículo (DMSO) (ref. D2660, dimethyl sulfoxide; Sigma-Merck, Argentina) por 1, 2 y 3 horas.

Las imágenes fueron tomadas antes del tratamiento, a 1 h, 2 h y 3 h posteriores al tratamiento.

Para analizar el daño al ADN causado por la CPT se realizaron imágenes de microscopía de inmunofluorescencia. Las células se fijaron con 4 % de formaldehído por 30 minutos a temperatura ambiente y se las lavó 3 veces con PBS, luego se las permeabilizó con 0.1 % de BSA/PBS y 0.1 % de solución de Tritón X-100 por 1 h, seguido de un bloqueo con 10 % de suero de cabra/PBS y 0.1 % de solución Tween20. La incubación con anticuerpos primarios anti-γH2AX (rabbit IgG, ref. ab2893; Abcam, United States) y p53 (mouse IgG, ref. ab1101; Abcam, United States) se realizó durante toda la noche a 4°Cen dilución 1:100 en solución bloqueante y la incubación con el anticuerpo secundario con Alexa Fluor 594 (anti-mouse, ref. R37121; Thermo Fisher Scientific, United States) y Alexa Fluor 488 (anti-rabbit, ref. A11034; Thermo Fisher Scientific, United States) se realizó en cuarto oscuro a temperatura ambiente por 1 h junto con DAPI. Las células fueron

lavadas y fotografiadas con un microscopio de fluorescencia EVOS (Thermo Fisher Scientific, United States). Se evaluó la unión no específica de anticuerpo secundario en ausencia de anticuerpo primario. Se tomaron imágenes de 4 campos de 3 réplicas independientes y se analizaron con el software ImageJ para determinar el promedio de la intensidad de fluorescencia por núcleo y la significancia estadística entre cultivos tratados y controles se evaluó con un t-test de Welch de dos muestras utilizando el paquete estadístico R.

Se realizó además un ensayo de anexina en las células, el cual mide la muerte celular en estadios tempranos. Para detectar la translocación de los residuos de fosfatidilserina (PS) en células apoptóticas se utilizaron los kits comerciales de Anexina V-FITC (ref. 556547; BD Pharmingen, United States) y Anexina V-PE (ref. 559763; BD Pharmingen, United States). Las células fueron colectadas incluyendo el sobrenadante e incubadas con los reactivos provistos por el fabricante del kit y finalmente se las hizo correr por el citómetro BD Accuri Flow. Los resultados de 3 réplicas independientes se analizaron utilizando el software FlowJo (v7.6) y la significancia estadística entre tratadas y controles del tercer cuadrante se evaluó con un t-test de Welch de dos muestras utilizando el paquete estadístico R.

Para capturar las imágenes para el aprendizaje automático se utilizó un microscopio EVOS con un objetivo 20x y una intensidad constante de luz de 40 %. Se tomaron entre 30 y 50 micrografías al azar a través de los 4 pocillos centrales, 2 tratados y 2 controles, en 4 réplicas biológicas independientes de las 7 líneas celulares evitando la superposición de campos y las zonas con pocas células. Se guardaron en archivos .png. El tamaño original de estas imágenes fue de 960 x 1280 píxeles y fueron cortadas para obtener de cada una 4 imágenes de 480 x 640 x 3, produciendo un total de 58596 micrografías considerando todos los tiempos (15224 imágenes pertenecientes a 1h, 15312 a 2h y 15032 al tratamiento de 3h).

Para entrenar la red neuronal y validar los resultados se utilizó fast.ai (v1.0.60) un frontend de PyTorch (v1.4). Se eligió la ResNet50 entre las diferentes arquitecturas probadas (ResNet34, ResNet50, ResNet101 y DenseNet121) debido a su excelente resultado. Se dejó una réplica de las cuatro para testear el modelo. De las tres réplicas restantes se dejó un 70 % de las imágenes para entrenamiento y un 30 % para validación. Se puede acceder a una rutina de Python con los detalles de los hiperparámetros utilizados en los entrenamientos en celldeath: a deep learning-based tool for classification of cell death.

Resultados y Discusión

En primer lugar, buscamos confirmar que en los tiempos estipulados se producía el fenómeno de muerte celular en las líneas celulares utilizadas. En la figura A1 se observa una inmunofluorescencia representativa de una de las líneas tumorales, MCF7, luego de 6h de tratamiento con 10μM de CPT y su control (DMSO). Consistentemente con las consecuencias del daño al DNA (DSB) el cultivo muestra un incremento en la señal nuclear de H2AX fosforilada y la acumulación de p53 significativamente dependiente de CPT. Se observan resultados similares para todas las líneas tumorales entre 3 y 6h de tratamiento. Las líneas pluripotentes inducidas de nuestro laboratorio (IPS1 e IPS2) y las embrionarias H9 también mostraron diferencias significativas en el aumento de γH2AX y p53, pero en general se mostraron más sensibles a la droga, observándose la aparición de apoptosis a concentraciones y tiempos de exposición menores que las líneas tumorales. En la figura A2 se puede observar la cuantificación de la intensidad de fluorescencia de las inmunofluorescencias de los cultivos de IPS1 sometidos a CPT 1μM por 1.5h mostrando diferencias significativas entre tratamiento y control.

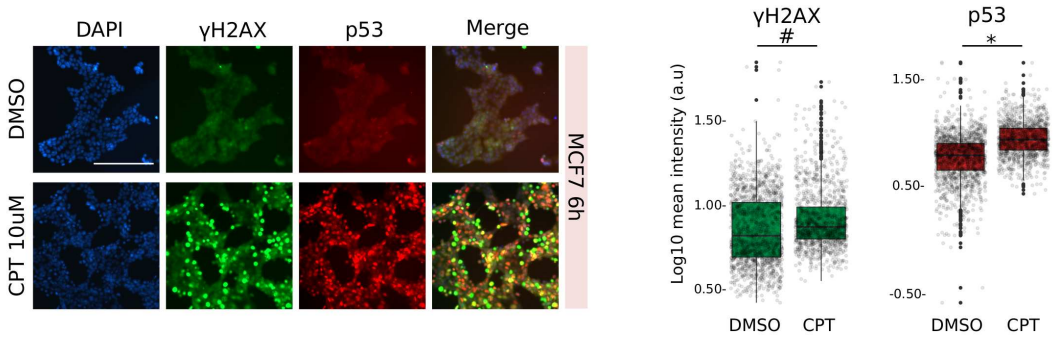


Figura A1: **Evaluación de muerte celular por medio de inmunofluorescencia en células tumorales MCF7** Izq: Inmunofluorescencia representativa con anti-γH2AX y anti-p53 de la línea tumoral MCF7 tratada con 10μM de CPT y su control (DMSO) por 6h (n=3). Barra de escala = 200μm. Der: Distribución del promedio de intensidad de fluorescencia por núcleo de todos los campos de la inmunofluorescencia que vemos a la izquierda medida en unidades arbitrarias (log10 u.a.) para γH2AX a la izquierda y p53 a la derecha.

Como segundo ensayo utilizamos la detección de muerte celular con Anexina-V/7-AAD. Se detectó un incremento significativo y temprano de muerte celular (Figura A3) en los cultivos tratados con CPT comparados con los cultivos control.

Con las imágenes obtenidas se formó un conjunto de datos con las etiquetas DMSO y CPT en las que el objetivo fue clasificar el tratamiento independiente-

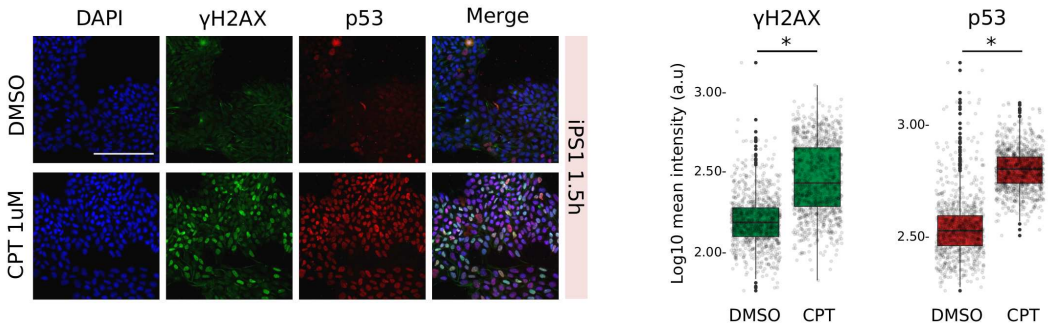


Figura A2: Izq: Inmunofluorescencia representativa con anti-γH2AX y anti-p53 de la línea iPS1 tratada con 1μM de CPT y su control (DMSO) por 1.5h (n=3). Barra de escala = 200μm. Der: Distribución del promedio de intensidad de fluorescencia por núcleo de todos los campos de la inmunofluorescencia que vemos a la izquierda medida en unidades arbitrarias (log10 media de intensidad [u.a.]) para γH2AX a la izquierda y p53 a la derecha.

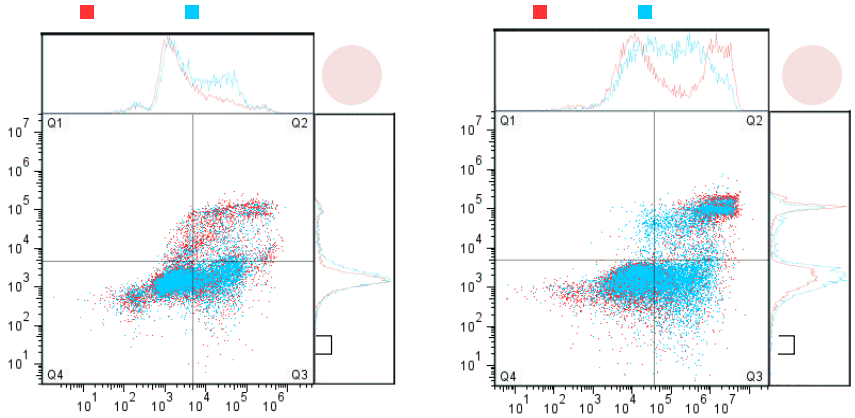


Figura A3: **Análisis de las citometrías de flujo para annexina V/7-AAD** Análisis de las citometrías de flujo entre células tratadas (celeste) y células control (rojo) discriminando células en apoptosis temprana (Q3) de células muertas (Q2). La exposición de fosfatidilserina en la membrana plasmática de la célula es un marcador de apoptosis temprana. Se pueden diferenciar células apoptóticas ya que la fosfatidilserina interacciona con la proteína anexina V, pero al estar intactas no están exponiendo el núcleo por lo que serán 7-AAD negativas. Las células en apoptosis tardía o necróticas son anexina V positivas y 7-AAD positivas, las células sanas no se tiñen. Izq.: Citometría de las células de la línea MCF7 tratadas con CPT 10μM por 6h, análisis con anexinaV-FITC y 7-AAD; n=3. Der.: Citometría de las células de la línea IPS1 tratadas con CPT 1μM por 3h, análisis con anexinaV-PE y 7-AAD; n=3.

mente de la línea celular utilizada. Otro experimento que se realizó fue "todos vs todos", tenía el objetivo de clasificar tratamiento y línea celular, lo que resultó en 14 etiquetas (DMSO-MCF7, CPT-MCF7, DMSO-H9, CPT-H9, DMSO-PC3, ...). Por último, se entrenó cada línea celular por separado para predecir tratamiento o control (CPT o DMSO). La clasificación a 1h de aplicado el tratamiento o el vehículo ("CPT vs DMSO", tabla A2) fue exitosa con una exactitud promedio de 5 corridas de 98.18 ± 0.33 % en el grupo de validación y un 96.58 ± 0.24 % en el grupo de testeo cuando se comparan las imágenes sin identificar la línea celular. Comenzar el entrenamiento con parámetros pertenecientes a un modelo entrenado con imágenes pertenecientes a la base de datos ImageNet (Russakovsky, 2015) no modificó la exactitud. La matriz de confusión (Figura A4, Izq) muestra pocos eventos mal clasificados: de 4188 imágenes, 65 fueron falsos positivos (predijo CPT cuando en realidad era DMSO) y 52 falsos negativos (predijo DMSO, siendo CPT en realidad). El caso de todos vs todos también arrojó muy buenos valores de exactitud y mejoró con los parámetros preentrenados ("Todos vs todos", tabla A2). La matriz de confusión (Figura A4, derecha) muestra pocas clasificaciones erradas, pero la mayor cantidad de confusiones son generadas entre líneas pluripotentes inducidas (IPS1-CPT/IPS2-CPT e IPS1-DMSO/IPS2-DMSO) y no entre tratamientos, probablemente por su parecido fenotípico. La alta densidad de eventos en la diagonal indica que la red neuronal fue capaz de identificar rasgos específicos de la muerte celular que le permite clasificar las etiquetas eficazmente. Cada línea por separado mostró muy buenos resultados, con valores de exactitud algo mayores en las líneas pluripotentes, confirmando que el modelo puede ser utilizado en experimentos de línea única o múltiples líneas en paralelo.

Para mayor evidencia de que la red advierte rasgos específicos de muerte celular, sorprendentemente, en algunas líneas, el modelo pudo discriminar entre tratadas y control aún sin haber sido entrenada con esa línea específica (Tabla A3). Entrenamos la red con todas las líneas menos una que apartamos para testear. La red no pudo discriminar tratadas de control en las líneas tumorales PC3 (53 %) y U2OS (64 %) como se esperaba, pero tuvo una buena performance con el resto de líneas tumorales y una excelente con las líneas pluripotentes, alcanzando la exactitud del grupo de validación.

La generalización es un objetivo de los modelos matemáticos, pero en el aprendizaje profundo, si queremos clasificar a una imagen en el grupo de imágenes de automóviles, se debe entrenar a la red con imágenes de automóviles. En este caso,

Condición	FPE	FPV	Exactitud (validación)	Exactitud (test)
CPT vs DMSO	0.068	0.045	0.9837	0.9723
CPT vs DMSO*	0.055	0.051	0.9825	0.9790
Todos vs todos	0.068	0.330	0.9979	0.8271
Todos vs todos*	0.029	0.035	0.9900	0.8658
PC3	0.138	0.041	0.9860	0.9550
MCF7	0.081	0.146	0.9528	0.9234
T47D	0.204	0.054	0.9746	0.8667
U2OS	0.141	0.002	1.000	0.9444
IPS1	0.379	0.056	0.998	0.970
IPS2	0.091	0.0007	1.000	0.948
H9	0.007	0.002	1.000	0.996

FPE: Función de Pérdida de Entrenamiento
FPV: Función de Pérdida de Validación

Tabla A2: Tabla de valores de la función de pérdida para el conjunto de entrenamiento y validación de los valores de exactitud más altos logrados para los conjuntos de validación y entrenamiento de ResNet50 con las imágenes tomadas a 1h del tratamiento. Se incluyen los resultados de comenzar el entrenamiento con los parámetros preentrenados (*) con imágenes de ImageNet.

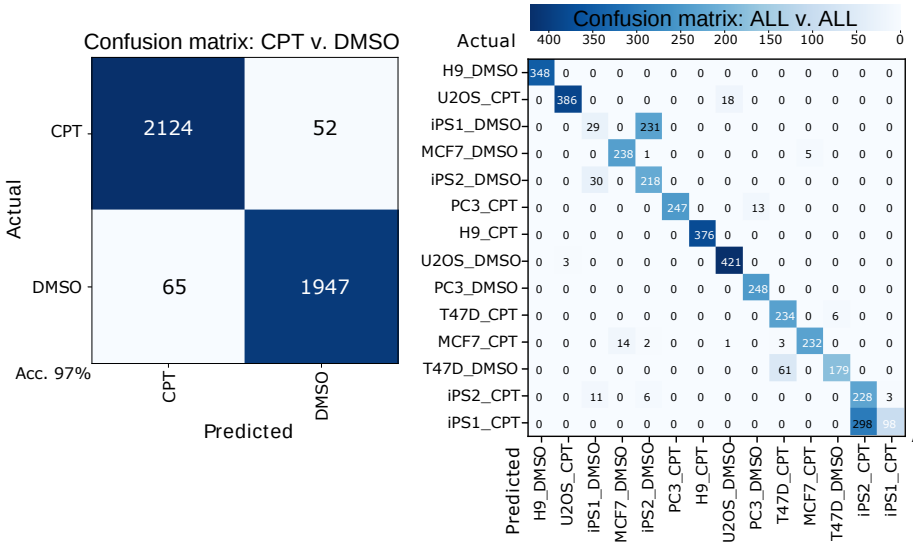


Figura A4: **Matriz de confusión para CPT vs DMSO.** Matriz de confusión para el conjunto de testeo de CPT vs DMSO a 1h del tratamiento a la izquierda y para Todos vs todos a la derecha.

Línea de testeo	FPE	FPV	Exactitud (validación)	Exactitud (test)
PC3	0.053	0.032	0.9872	0.5283
MCF7	0.054	0.038	0.9901	0.8688
T47D	0.071	0.047	0.9858	0.7734
U2OS	0.043	0.059	0.9800	0.6363
IPS1	0.063	0.052	0.9820	0.9871
IPS2	0.046	0.056	0.9826	0.9708
H9	0.076	0.058	0.9822	0.9752

FPE: Función de Pérdida de Entrenamiento
FPV: Función de Pérdida de Validación

Tabla A3: Mayores valores de exactitud alcanzados por el conjunto de imágenes de testeo, la línea de testeo no era conocida por la red ya que fue totalmente excluida del entrenamiento. También se muestran los valores correspondientes a la función de pérdida para el entrenamiento y la validación de cada corrida.

la red pudo clasificar las imágenes de líneas celulares con las que no fue entrenada. Este inesperado resultado sugiere que la red fue capaz de extraer rasgos asociados a la muerte celular de las líneas con las que fue entrenada y extrapolarlos a imágenes desconocidas aunque relacionadas. Aunque esto no deja de ser una curiosidad. Para un buen desempeño en la generalización la red debe ser entrenada con las líneas que quieren ser clasificadas. Esto se pudo comprobar con las líneas tumorales en las cuales el desempeño de la red fue mucho menor.

Para un laboratorio que mida muerte celular rutinariamente utilizar el celldeath puede ser una buena inversión de tiempo inicial. Con una simple imagen tomada en un microscopio disponible en cualquier laboratorio, y sin matar el cultivo, se puede saber si las células entraron en proceso de muerte o no mucho antes de observar cambios morfológicos a simple vista. Esto repercute en evitar gastos de reactivos y tiempo, además de la versatilidad del diseño experimental al no tener que ponerle fin al cultivo.

Conclusión

Se creó y compartió una herramienta (celldeath) para clasificar micrografías de cultivos celulares en proceso de muerte celular o control. En un principio es necesario contar con una gran cantidad de imágenes de los cultivos celulares anhelados para entrenar la red convolucional residual. Pero, luego de la puesta a punto, cada micrografía puede predecir si el cultivo continuará su desarrollo normalmente o

entrará en un camino de muerte celular luego de algún detonante experimental.

Dedicatoria

A Gonzalo y nuestros hijos, Facundo y Candelaria.

Agradecimientos

Mi mayor agradecimiento es para Santiago Miriuka. Por prestarme su querido proyecto y permitirme, con gran paciencia, hacerlo mío para aprender. Porque no sólo demostró sus capacidades desde lo académico sino ser una buena persona desde el primer día, dándome libertad para equivocarme y brindándome lo más importante: su tiempo. Su esfuerzo y compromiso constante hicieron posible este proyecto de nivel internacional.

Otro enorme gracias es para Alejandro La Greca. La persona que estuvo día a día enseñándome desde el más simple procedimiento hasta los conceptos más abstractos. Codo a codo, con paciencia y alegría, logró que disfrute de aprender en cada momento.

Por supuesto a las instituciones que le dieron marco formal a esta idea: La Facultad de Ciencias Exactas y Naturales, que con sus magistrales docentes dejan una huella en el corazón de todos sus egresados. A CONICET por su apoyo económico. A FLENI y a todos sus empleados por recibirme cada día.

A todos los amigos y amigas que hice en el proceso, que me acompañaron desde los momentos de cursar en la facultad hasta hoy: A “Las Biólogas” por la fuerza espiritual, a “Los BIA” por todas las cervezas compartidas, porque su ejemplo y ayuda me dio fuerzas para llegar a la meta. Y a mis compañeros y compañeras de FLENI por su generosidad para compartir su conocimiento, por su paciencia con la “bioinformática” trabajando en la mesada y su buena onda tan necesaria en el trabajo diario.

A Marcelo Martí por abrirme las puertas hacia la genómica clínica. A Esteban Mocskos por hacerme llorar enseñándome a programar. A Jonathan Zaiat por regalarme independencia con su frase: “¡Googleá el error!”.

A Gustavo Lado por desvelar el mundo del aprendizaje profundo con buena onda y brindar su tiempo con sus consejos y correcciones para esta tesis.

A Verónica Lía por compartir su amor a la genética y dedicar su tiempo a este trabajo con sus aportes.

A Ernesto Pérez, mi papá, por investigar mis preguntas raras mostrándome el mundo de los libros.

A Gonzalo Fuentes y Arballo, mi marido, por ser mi pilar y mantener la ilusión de que algún día “voy a ganar plata”. A mis hijos que crecieron junto a mí durante estos años.

Y finalmente a los jurados que accedieron al desafío de evaluar esta tesis: ¡Llegaron al final!.

Muchas, muchísimas gracias a todos.

Bibliografía

Abouelwafa M., Georrge J J. 2020. Transcriptomics databases. Recent Trends in Science and Technology-2020, pp. 155-161. Rajkot, Gujarat, India: Christ Publications. ISBN: 9788192952154. https://doi.org/10.6084/m9.figshare.13491453 Available at SSRN:
https://ssrn.com/abstract=3722472
www.publications.christcollegerajkot.edu.in

Agatston A.S., Janowitz W.R., Hildner F.J., Zusmer N.R., Viamonte M., Detrano R. Quantification of coronary artery calcium using ultrafast computed tomography. Journal of the American College of Cardiology. 1990;15(4):827-832 ISSN 0735-1097.
https://doi.org/10.1016/0735-1097(90)90282-T.

Alcidi G., Goffredo G., Correale M., Brunetti N.D., Iacoviello M. Brain Natriuretic Peptide Biomarkers in Current Clinical and Therapeutic Scenarios of Heart Failure. J Clin Med. 2022;11(11):3192.
https://doi.org/10.3390/jcm11113192.

All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. Nature. 2024;627(8003):340-346.
https://doi.org/10.1038/s41586-023-06957-x

Anders S., Pyl P.T., Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-169.
https://doi.org/10.1093/bioinformatics/btu638

Ang M.Y., Takeuchi F., Kato N. Deciphering the genetic landscape of obesity: a data-driven approach to identifying plausible causal genes and therapeutic targets. J Hum Genet. 2023;68:823–833.
https://doi.org/10.1038/s10038-023-01189-3

Aragam K.G., Dobbyn A., Judy R., et al. Limitations of Contemporary Guidelines for Managing Patients at High Genetic Risk of Coronary Artery Disease. J Am Coll Cardiol. 2020;75(22):2769-2780.
https://doi.org/10.1016/j.jacc.2020.04.027

Aragam K.G., Jiang T., Goel A., et al. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. Nat Genet. 2022;54(12):1803-

1815.
https://doi.org/10.1038/s41588-022-01233-6

Aras M., Tchang B.G., Pape J. Obesity and Diabetes. Nurs Clin North Am. 2021;56(4):527-541.
https://doi.org/10.1016/j.cnur.2021.07.008

Arber D.A., Orazi A., Hasserjian R.P., et al. International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. Blood. 2022;140(11):1200-1228.
https://doi.org/10.1182/blood.2022015850

Arnett D.K., Blumenthal R.S., et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. American Heart Association Circulation. 2019;140(11):563-595.
https://doi.org/10.1161/CIR.0000000000000677

Asakura M., Kitakaze M. Global gene expression profiling in the failing myocardium. Circ J. 2009;73(9):1568-1576.
https://doi.org/10.1253/circj.cj-09-0465

Ashburner M., Ball C.A., Blake J.A., et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-29.

Azevedo T., Dimitri G.M., Lió P. et al. Multilayer modeling of the human transcriptome and biological mechanisms of complex diseases and traits. npj Syst Biol Appl. 2021;7:24.
https://doi.org/10.1038/s41540-021-00186-6

Babu M., Snyder M. Multi-Omics Profiling for Health. Mol Cell Proteomics. 2023;22(6):100561.
doi: 10.1016/j.mcpro.2023.100561.

Badawy M.A.E.M.D., Naing L., Johar S., Ong S., Rahman H.A., Tengah D.S.N.A.P., Chong C.L., Tuah N.A.A. Evaluation of cardiovascular diseases risk calculators for CVDs prevention and management: scoping review. BMC Public Health. 2022;22(1):1742.
doi: 10.1186/s12889-022-13944-w.

Balasubramanian S, Aggarwal P, Sharma S. Lipoprotein Lipase Deficiency. [Updated 2023 Jul 3]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-.
www.ncbi.nlm.nih.gov/books/NBK560795/

Bellman, R. Adaptive control processes. Princeton, NJ: Princeton University Press. 1961.

Bild D.E., Bluemke D.A., Burke G.L., et al. Multi-Ethnic Study of Atherosclerosis: objectives and design. Am J Epidemiol. 2002;156(9):871-881.
https://doi.org/10.1093/aje/kwf113

Biros E., Karan M., Golledge J. Genetic variation and atherosclerosis. Curr Genomics. 2008;9(1):29-42.
doi: 10.2174/138920208783884856.

Bove G., Del Gaudio N., Altucci, L. Epitranscriptomics and epigenetics: two sides of the same coin?. Clin Epigenet. 2024;16:121. https://doi.org/10.1186/s13148-024-01729-4

Bozkurt B., Coats A. J. S., Tsutsui H.; Abdelhamid C. M., Adamopoulos S., Albert N., Anker S. D., Atherton J., Butler J.Universal definition and classification of heart failure: A report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal Definition of Heart Failure. European Journal of Heart Failure. 2021;23(3):352-380. ISSN 1879-0844.
https://doi.org/10.1002/ejhf.2115

Butt J.H., Petrie M.C., Jhund P.S., et al. Anthropometric measures and adverse outcomes in heart failure with reduced ejection fraction: revisiting the obesity paradox. Eur Heart J. 2023;44(13):1136-1153.
https://doi.org/10.1093/eurheartj/ehad083

Budoff M.J., Hokanson J.E., Nasir K. Progression of coronary artery calcium predicts all-cause mortality. JACC: Cardiovascular Imaging. 2010;3:1229–1236.

Camaré C., Pucelle M., Nègre-Salvayre A., Salvayre R. Angiogenesis in the atherosclerotic plaque. Redox Biology. 2017;12:18-34.
https://doi.org/10.1016/j.redox.2017.01.007

Campisi J., Kapahi P., Lithgow G.J., Melov S., Newman J.C., Verdin E. From discoveries in aging research to therapeutics for healthy aging. Nature. 2019;571(7764):183-192.
https://doi.org/10.1038/s41586-019-1365-2

Cano-Espinosa C., González G., Washko G.R., Cazorla M., Estépar R.S.J. Automated Agatston Score Computation in non-ECG Gated CT Scans Using Deep Learning. Proc SPIE Int Soc Opt Eng. 2018;10574:105742K.
https://doi.org/10.1117/12.2293681

Casamassimi A., Federico A., Rienzo M., Esposito S., Ciccodicola A. Transcriptome Profiling in Human Diseases: New Advances and Perspectives. Int J Mol Sci. 2017;18(8):1652. Published 2017 Jul 29.
https://doi.org/10.3390/ijms18081652

CDC. Division of Nutrition, Physical Activity, and Obesity, National Center for Chronic Disease Prevention and Health Promotion. How good is BMI as an indicator of body fatness? [internet]. Last Reviewed: June 3, 2022.
www.cdc.gov/healthyweight/assessing/bmi/adultbmi/index.html

Cerneckis J., Ming G., Song H., He C., Shi Y. The rise of epitranscriptomics: recent developments and future directions. Cell - Trends in pharmacological Sciences. 2024;45(1):24-38.
https://www.cell.com/trends/pharmacological-sciences/fulltext/S0165-6147(23)00254-7

Chao H., Shan H., Homayounieh F. et al. Deep learning predicts cardiovascular disease risks from lung cancer screening low dose computed tomography. Nat Commun. 2021;12:2963.
https://doi.org/10.1038/s41467-021-23235-4

Cheng Y., Xu S., Santucci K., Lindner G., Janitz M. Machine learning and related approaches in transcriptomics. Biochemical and Biophysical Research Communications. 2024;724:150225.
https://doi.org/10.1016/j.bbrc.2024.150225

Cheng-Wu Zeng, Xing-Ju Zhang, Kang-Yu Lin, Hua Ye, Shu-Ying Feng, Hua Zhang and Yue-Qin Chen. miR-125b-Mediated Mitochondrial Pathways by Camptothecin. Molecular Pharmacology Abril 2012;81(4)578-586;
https://doi.org/10.1124/mol.111.076794

Chellapilla K., Puri S., Simard P. 2006. High Performance Convolutional Neural Networks for Document Processing. In Guy Lorette, editor, Tenth International Workshop on Frontiers in Handwriting Recognition, Université de Rennes 1, Oct 2006, La Baule (France). ffinria-00112631

Chiorescu R.M., Mocan M., Inceu A.I., Buda A.P., Blendea D., Vlaicu S.I. Vulnerable Atherosclerotic Plaque: Is There a Molecular Signature?. Int J Mol Sci. 2022;23(21):13638. Published 2022 Nov 7.
https://doi.org/10.3390/ijms232113638

Handwriting Recognition, La Baule (France). Université deRennes 1, Suvisoft.
www.suvisoft.com

Choquet H., Meyre D. Genetics of Obesity: What have we Learned?.
Curr Genomics. 2011;12(3):169-179.
https://doi.org/10.2174/138920211795677895

Conrard L., Tyteca D. Regulation of Membrane Calcium Transport Proteins by the Surrounding Lipid Environment. Biomolecules. 2019;9(10):513. Published 2019 Sep 20.
https://doi.org/10.3390/biom9100513

Cook N.R., Ridker P.M. Calibration of the Pooled Cohort Equations for Atherosclerotic Cardio-

vascular Disease: An Update. Ann Intern Med. 2016;165(11):786-794.
https://doi.org/10.7326/M16-1739

Corral-Acero J, Margara F, Marciniak M, Rodero C, Loncaric F, Feng Y, Gilbert A, Fernandes JF, Bukhari HA, Wajdan A, Martinez MV, Santos MS, Shamohammdi M, Luo H, Westphal P, Leeson P, DiAchille P, Gurev V, Mayr M, Geris L, Pathmanathan P, Morrison T, Cornelussen R, Prinzen F, Delhaas T, Doltra A, Sitges M, Vigmond EJ, Zacur E, Grau V, Rodriguez B, Remme EW, Niederer S, Mortier P, McLeod K, Potse M, Pueyo E, Bueno-Orovio A, Lamata P. The 'Digital Twin' to enable the vision of precision cardiology. Eur Heart J. 2020;41(48):4556-4564.
doi: 10.1093/eurheartj/ehaa159.

Crous-Bou M., Harrington L.B., Kabrhel C. Environmental and Genetic Risk Factors Associated with Venous Thromboembolism. Semin Thromb Hemost. 2016;42(8):808-820.
https://doi.org/10.1055/s-0036-1592333

Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9(386):eaal5209.

Curtis K.M., Mohllajee A.P., Martins S.L., Peterson H.B. Combined oral contraceptive use among women with hypertension: a systematic review. Contraception. 2006;73(2):179-188.
https://doi.org/10.1016/j.contraception.2005.08.005

D'Agostino R.B. Sr., Vasan R.S., Pencina M.J., et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117(6):743-753.
https://doi.org/10.1161/CIRCULATIONAHA.107.699579

D'Arcy MS. Cell death: a review of the major forms of apoptosis, necrosis and autophagy. Cell Biol Int. 2019;43(6):582–592.

Da Dalt L., Cabodevilla A.G., Goldberg I.J., Norata G.D. Cardiac lipid metabolism, mitochondrial function, and heart failure. Cardiovasc Res. 2023;119(10):1905-1914.
https://doi.org/10.1093/cvr/cvad100

Daan N.M.P., Louwers Y.V., Koster M.P.H., Eijkemans M.J.C., de Rijke Y. B., Lentjes E.W.G., Fauser B.C.J.M., Laven J.S.E. Cardiovascular and metabolic profiles amongst different polycystic ovary syndrome phenotypes: who is really at risk? Fertility and Sterility. 2014;102(5):1444-1451.e3. ISSN 0015-0282.
https://doi.org/10.1016/j.fertnstert.2014.08.001

De Genst E., Foo K.S., Xiao Y. Blocking phospholamban with VHH intrabodies enhances contractility and relaxation in heart failure. Nat Commun. 2022;13(1):3018. Published 2022 May 31.
https://doi.org/10.1038/s41467-022-29703-9

De Vos B. D., Wolterink J. M., Leiner T., De Jong P. A., Lessmann N., Išgum I. Direct Automatic Coronary Calcium Scoring in Cardiac and Chest CT. IEEE Transactions on Medical Imaging. 2019;38(9):2127-2138. Doi: 10.1109/TMI.2019.2899534

Demer L.L., Tintut Y. Vascular calcification: pathobiology of a multifaceted disease. Circulation. 2008;117(22):2938-2948.
https://doi.org/10.1161/CIRCULATIONAHA.107.743161

Diaz J., Ahsen M.E., Schaffter T., Chen X., Realubit R.B., Karan Ch., Califano A., Losic B., Stolovitzky G., Valencia A., Cheah K.S.E., Asmund F. The transcriptomic response of cells to a drug combination is more than the sum of the responses to the monotherapies. eLife. 2020;9:1–62.
https://doi.org/10.7554/eLife.52707

Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. A review of UMAP in population genetics. J Hum Genet. 2021;66(1):85-91.
https://doi.org/10.1038/s10038-020-00851-4

Ding J., Bar-Joseph Z. Analysis of time-series regulatory networks. Curr. Opinion Sys. Biol. 2020;21:16–24. ISSN 2452-3100.
https://doi.org/10.1016/j.coisb.2020.07.005

Dobin A., Davis C.A., Schlesinger F., et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
https://doi.org/10.1093/bioinformatics/bts635

Dodt M., Roehr J.T., Ahmed R., Dieterich C. FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. Biology (Basel). 2012;1(3):895-905.
https://doi.org/10.3390/biology1030895

Doenst T., Nguyen T.D., Abel E.D. Cardiac metabolism in heart failure: implications beyond ATP production. Circ Res. 2013;113(6):709-724.
https://doi.org/10.1161/CIRCRESAHA.113.300376

Dolgin M., New York Heart Association. 1994. Nomenclature and Criteria for Diagnosis of Diseases of the Heart and Great Vessels. 9th ed. Boston: Little Brown.

Doust J.A., Pietrzak E., Dobson A., Glasziou P. How well does B-type natriuretic peptide predict death and cardiac events in patients with heart failure: systematic review. BMJ. 2005;330(7492):625.
https://doi.org/10.1136/bmj.330.7492.625

Dzau V., Braunwald E. Resolved and unresolved issues in the prevention and treatment of coronary artery disease: a workshop consensus statement. Am Heart J. 1991;121(4 Pt 1):1244-1263.

https://doi.org/10.1016/0002-8703(91)90694-d

Dzau V.J., Antman E.M., Black H.R. The cardiovascular disease continuum validated: clinical evidence of improved patient outcomes: part I: Pathophysiology and clinical trial evidence (risk factors through stable coronary artery disease). Circulation. 2006;114(25):2850-2870.
https://doi.org/10.1161/CIRCULATIONAHA.106.655688

Elliott J., Bodinier B., Bond T.A., et al. Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. JAMA. 2020;323(7):636-645.
https://doi.org/10.1001/jama.2019.22241

Elmore S. Apoptosis: a review of programmed cell death. Toxicol Pathol. 2007;35(4):495–516.

Emelia J.B., Muntner P. Heart Disease and Stroke Statistics—2019 Update A Report From the American Heart Association. AHA Journal. Circulation. 2019;139:e442.
https://doi.org/10.1161/CIR.0000000000000659

Eraslan G., Simon L.M., Mircea M., Mueller N.S., Theis F.J. Single-cell RNA-seq denoising using a deep count autoencoder. Nature Communications. 2019;10:390.
https://doi.org/10.1038/s41467-018-07931-2

Faris O., Shuren J. An FDA Viewpoint on Unique Considerations for Medical-Device Clinical Trials. N Engl J Med. 2017;376(14):1350-1357.
https://doi.org/10.1056/NEJMra1512592

Faulkner J.L. Obesity-associated cardiovascular risk in women: hypertension and heart failure. Clin Sci (Lond). 2021;135(12):1523-1544.
https://doi.org/10.1042/CS20210384

Frésard L., Smail C., Ferraro N.M. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat Med. 2019;25(6):911-919.
https://doi.org/10.1038/s41591-019-0457-8

Frisk M., Le C., Shen X. Etiology-Dependent Impairment of Diastolic Cardiomyocyte Calcium Homeostasis in Heart Failure With Preserved Ejection Fraction. J Am Coll Cardiol. 2021;77(4):405-419.
https://doi.org/10.1016/j.jacc.2020.11.044

Fonarow G.C., Srikanthan P., Costanzo M.R., Cintron G.B., Lopatin M.; ADHERE Scientific Advisory Committee and Investigators. An obesity paradox in acute heart failure: analysis of body mass index and inhospital mortality for 108,927 patients in the Acute Decompensated Heart Failure National Registry. Am Heart J. 2007;153(1):74-81.

https://doi.org/10.1016/j.ahj.2006.09.007

García CP, Videla Richardson GA, Romorini L, Miriuka SG, Sevlever GE, Scassa ME. Topoisomerase I inhibitor, camptothecin, induces apoptogenic signaling in human embryonic stem cells. Stem Cell Res. 2014;12(2):400–14.

Garrow J.S., Webster J. Quetelet's index (W/H2) as a measure of fatness.
Int J Obes. 1985;9(2):147-153.

Gautier-Stein A., Soty M., Chilloux J., Zitoun C., Rajas F., Mithieux G. Glucotoxicity induces glucose-6-phosphatase catalytic unit expression by acting on the interaction of HIF-1α with CREB-binding protein. Diabetes. 2012;61(10):2451-2460.
https://doi.org/10.2337/db11-0986

Ge Z., Li A., McNamara J., Dos Remedios C., Lal S. Pathogenesis and pathophysiology of heart failure with reduced ejection fraction: translation to human studies. Heart Fail Rev. 2019;24(5):743-758.
https://doi.org/10.1007/s10741-019-09806-0

Geraghty L., Figtree G.A., Schutte A.E., Patel S., Woodward M., Arnott C. Cardiovascular Disease in Women: From Pathophysiology to Novel and Emerging Risk Factors. Heart Lung Circ. 2021;30(1):9-17.
https://doi.org/10.1016/j.hlc.2020.05.108

Gershoni M., Pietrokovski S. The landscape of sex-differential transcriptome and its consequent selection in human adults. BMC Biol. 2017;15(1):7. Published 2017 Feb 7.
https://doi.org/10.1186/s12915-017-0352-z

Gonorazky H.D., Naumenko S., Ramani A.K. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease [published correction appears in Am J Hum Genet. 2019 May 2;104(5):1007]. Am J Hum Genet. 2019;104(3):466-483.
https://doi.org/10.1016/j.ajhg.2019.01.012

Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press [on line]. 2016. Available from:
www.deeplearningbook.org

Graham I.M., Di Angelantonio E., Visseren F., et al. Systematic Coronary Risk Evaluation (SCORE): JACC Focus Seminar 4/8. J Am Coll Cardiol. 2021;77(24):3046-3057.
https://doi.org/10.1016/j.jacc.2021.04.052

Greenland P., Blaha M.J., Budoff M.J., Erbel R., Watson K.E. Coronary Calcium Score and Cardiovascular Risk. Journal of the American College of Cardiology. 2018;72(4):434-447.

https://doi.org/10.1016/j.jacc.2018.05.027.

Grønbech C.H., Vording M.F., Timshel P.N., Sønderby C.K., Pers T.H., Winther O. scVAE: variational auto-encoders for single-cell gene expression data. Bioinformatics. 2020;36(16):4415–4422. https://doi.org/10.1093/bioinformatics/btaa293

GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-585.
https://doi.org/10.1038/ng.2653

Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang G, Cai J, et al. Recent advances in convolutional neural networks. Pattern Recogn. 2018;77:354–77

Gul B, Lansky A, Budoff MJ, et al. The Clinical Utility of a Precision Medicine Blood Test Incorporating Age, Sex, and Gene Expression for Evaluating Women with Stable Symptoms Suggestive of Obstructive Coronary Artery Disease: Analysis from the PRESET Registry. J Womens Health (Larchmt). 2019;28(5):728-735.
https://doi.org/10.1089/jwh.2018.7203

Gurgoglione F.L., Solinas E., Pfleiderer B., Vezzani A., Niccol G. Coronary atherosclerotic plaque phenotype and physiopathologic mechanisms: Is there an influence of sex? Insights from intracoronary imaging. atherosclerosis. 2023;384:117273 DOI:
https://doi.org/10.1016/j..117273

Haring B., Wissel S., Manson J.E. Somatic Mutations and Clonal Hematopoiesis as Drivers of Age-Related Cardiovascular Risk. Curr Cardiol Rep. 2022;24(8):1049-1058.
doi: 10.1007/s11886-022-01724-2.

He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. CoRR.2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 2016;770-778. 1512.03385.
doi: 10.1109/CVPR.2016.90.
http://arxiv.org/abs/1512.03385

Herrera B.M., Keildson S., Lindgren CM. Genetics and epigenetics of obesity.
Maturitas. 2011;69(1):41-49.
https://doi.org/10.1016/j.maturitas.2011.02.018

Hoffmann U., Massaro J.M., Fox C.S., Manders E., O'Donnell C.J. Defining normal distributions of coronary artery calcium in women and men (from the Framingham Heart Study). Am J Cardiol. 2008;102(9):1136-1141.e1.
https://doi.org/10.1016/j.amjcard.2008.06.038

Holzscheck N, Falckenhayn C, Söhle J, et al. Modeling transcriptomic age using knowledge-primed artificial neural networks. NPJ Aging Mech Dis. 2021;7(1):15. Published 2021 Jun 1. https://doi.org/10.1038/s41514-021-00068-5

Hopfield J.J. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences. 1982;79(8):2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Hou Z., Lu B., Gao Y. Prognostic value of coronary CT angiography and calcium score for major adverse cardiac events in outpatients. JACC Cardiovasc Imaging. 2012;5:990–999.

Howe K., Ross J.M., Loiselle D.S., Han J.C., Crossman D.J. Right-sided heart failure is also associated with transverse tubule remodeling in the left ventricle. Am J Physiol Heart Circ Physiol. 2021;321(5):H940-H947. https://doi.org/10.1152/ajpheart.00298.2021

Inouye M., Abraham G., Nelson C.P., et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. J Am Coll Cardiol. 2018;72(16):1883-1893. https://doi.org/10.1016/j.jacc.2018.07.079

Iravanian S., Dudley S.C. Jr. The renin-angiotensin-aldosterone system (RAAS) and cardiac arrhythmias [published correction appears in Heart Rhythm. 2008;5(10):1499]. Heart Rhythm. 2008;5(6 Suppl):S12-S17. https://doi.org/10.1016/j.hrthm.2008.02.025

Jaiswal S., Fontanillas P., Flannick J., et al. Age-related clonal hematopoiesis associated with adverse outcomes. N Engl J Med. 2014;371(26):2488-2498. https://doi.org/10.1056/NEJMoa1408617

Jaiswal S., Natarajan P., Silver A.J., et al. Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. N Engl J Med. 2017;377(2):111-121. https://doi.org/10.1056/NEJMoa1701719

Jiang L., Fernandes D., Mehta N., Bean J.L., Michaelis M.L., Zaidi A. Partitioning of the plasma membrane Ca2+-ATPase into lipid rafts in primary neurons: effects of cholesterol depletion. J Neurochem. 2007;102(2):378-388. https://doi.org/10.1111/j.1471-4159.2007.04480.x

Jensen M.D., Ryan D.H. Management of Overweight and Obesity in Adults: Guidelines From the Expert Panel. 2013;1:5. www.nhlbi.nih.gov/sites/default/files/media/docs/obesity-evidence-review.pdf

Jin S., Zeng X, Xia F., Huang W., Liu X. Application of deep learning methods in biological networks. Briefings in Bioinformatics. 2021;22(2):1902–1917.
https://doi.org/10.1093/bib/bbaa043

Jokela M., Laakasuo M. Obesity as a causal risk factor for depression: Systematic review and meta-analysis of Mendelian Randomization studies and implications for population mental health. J Psychiatr Res. 2023;163:86-92.
https://doi.org/10.1016/j.jpsychires.2023.05.034

Jolliffe I.T., Cadima J. Principal component analysis: a review and recent developments Phil. Trans. R. Soc. A. 2016;374:20150202
http://doi.org/10.1098/rsta.2015.0202

Kabore AF, Johnston JB, Gibson SB. Changes in the apoptotic and survival signaling in cancer cells and their potential therapeutic implications. Curr Cancer Drug Targets. 2004;4(2):147–63.

Kaiming H., Xiangyu Z., Shaoqing R., Jian S. Deep Residual Learning for Image Recognition. Cornell University Arxiv. 2015. doi.org/10.48550/arXiv.1512.03385

Kakati T., Bhattacharyya D.K., Kalita J.K. et al. DEGnext: classification of differentially expressed genes from RNA-seq data using a convolutional neural network with transfer learning. BMC Bioinformatics. 2022;23:7.
https://doi.org/10.1186/s12859-021-04527-4

Kanehisa M., Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30.
https://doi.org/10.1093/nar/28.1.27

Kanzi A.M., San J.E., Chimukangara B., Wilkinson E., Fish M., Ramsuran V.. Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. Front Genet. 2020;11: 544162.

Kay JG, Fairn GD. Distribution, dynamics and functional roles of phosphatidylserine within the cell. Cell Commun Signal. 2019;17(1):126.

Kenchaiah S., Evans J.C., Levy D., et al. Obesity and the risk of heart failure. N Engl J Med. 2002;347(5):305-313.
https://doi.org/10.1056/NEJMoa020245

Kernohan K.D., Frésard L., Zappala Z. Whole-transcriptome sequencing in blood provides a diagnosis of spinal muscular atrophy with progressive myoclonic epilepsy.
Hum Mutat. 2017;38(6):611-614.

https://doi.org/10.1002/humu.23211

Khan J, Wei JS, Ringnér M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001;7(6):673-679.
https://doi.org/10.1038/89044

Khan S.S., Post W.S., Guo X., et al. Coronary Artery Calcium Score and Polygenic Risk Score for the Prediction of Coronary Heart Disease Events. JAMA. 2023;329(20):1768-1777.
https://doi.org/10.1001/jama.2023.7575

Khan S.S., Matsushita K., Sang Y., et al. Development and Validation of the American Heart Association's PREVENT Equations. Circulation. 2024;149(6):430-449.
https://doi.org/10.1161/CIRCULATIONAHA.123.067626

Kho C., Lee A., Jeong D.SUMO1-dependent modulation of SERCA2a in heart failure. Nature. 2011;477(7366):601-605. Published 2011 Sep 7.
https://doi.org/10.1038/nature10407

Kho C. Targeting calcium regulators as therapy for heart failure: focus on the sarcoplasmic reticulum Ca-ATPase pump. Front Cardiovasc Med. 2023;10:1185261.
https://doi.org/10.3389/fcvm.2023.1185261

Khot U.N., Khot M.B., Bajzer C.T., Sapp S.K., Ohman E.M., Brener S.J., Ellis S.G., Lincoff A.M., Topol E.J. Prevalence of conventional risk factors in patients with coronary heart disease. JAMA. 2003;290(7):898-904.
doi: 10.1001/jama.290.7.898.

Kolbusz J., Rozycki P., Wilamowski B.M. 2017. The Study of Architecture MLP with Linear Neurons in Order to Eliminate the "vanishing Gradient" Problem. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L., Zurada, J. (eds) Artificial Intelligence and Soft Computing. ICAISC 2017. Lecture Notes in Computer Science, vol 10245. Springer, Cham.
https://doi.org/10.1007/978-3-319-59063-9_9

Koks G., Pfaff A.L., Bubb V.J., Quinn J.P., Koks S. At the dawn of the transcriptomic medicine. Exp Biol Med (Maywood). 2021;246(3):286-292.
https://doi.org/10.1177/1535370220954788

Kõks S., Keermann M., Reimann E. Psoriasis-Specific RNA Isoforms Identified by RNA-Seq Analysis of 173,446 Transcripts. Front Med (Lausanne). 2016;3:46. Published 2016 Oct 7.
https://doi.org/10.3389/fmed.2016.00046

Krebs J. Structure, Function and Regulation of the Plasma Membrane Calcium Pump in Health and Disease. Int J Mol Sci. 2022;23(3):1027. Published 2022 Jan 18.

https://doi.org/10.3390/ijms23031027

Kristan A., Debeljak N., Kunej T. Genetic variability of hypoxia-inducible factor alpha (HIFA) genes in familial erythrocytosis: Analysis of the literature and genome databases. Eur J Haematol. 2019;103(4):287-299.
https://doi.org/10.1111/ejh.13304

Kuksin M., Morel D., Aglave M., Danlos F-X, Marabelle A., Zinovyev A., Gautheret D., Verlingue L. Applications of single-cell and bulk RNA sequencing in onco-immunology. European Journal of Cancer. 2021;149:193-210. DOI:
https://doi.org/10.1016/j.ejca.2021.03.005

Lachmann A., Torre D., Keenan A.B, Jagodnik K.M., Lee H.J., Wang L., Silverstein M.C., Ma'ayan A.. Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun. 2018;9:1366.
https://doi.org/10.1038/s41467-018-03751-6

Lee S.H., Park S.Y., Choi C.S. Insulin Resistance: From Mechanisms to Therapeutic Strategies. Diabetes Metab J. 2022;46(1):15-37.
https://doi.org/10.4093/dmj.2021.0280

Lage I., Chen E., He J., Narayanan M., Kim B., Gershman S., Doshi-Velez F. An Evaluation of the Human-Interpretability of Explanation. Cornell arXiv:1902.00006. 2019.
https://doi.org/10.48550/arXiv.1902.00006

Lee-Liu D., Almonacid L.I., Faunes F., Melo F., Larrain J. Transcriptomics using next generation sequencing technologies. Methods Mol Biol. 2012;917:293-317.
https://doi.org/10.1007/978-1-61779-992-1_18

Leslie T. A bi-domain model for describing ischemic myocardial d-c potentials. Thesis. Ph.D. Massachusetts Institute of Technology. Dept. of Electrical Engineering and Computer Science. 1978.
https://dspace.mit.edu/handle/1721.1/16177

Lessmann N, van Ginneken B, Zreik M, et al. Automatic Calcium Scoring in Low-Dose Chest CT Using Deep Neural Networks With Dilated Convolutions. IEEE Trans Med Imaging. 2018;37(2):615-625.
https://doi.org/10.1109/TMI.2017.2769839

Li Q., Zhao Q., Zhang J., Zhou L., Zhang W, Chua B., Chen Y., Xu L., Li P. The Protein Phosphatase 1 Complex Is a Direct Target of AKT that Links Insulin Signaling to Hepatic Glycogen Deposition. Cell Reports. 2019;28(13):3406-3422.e7 ISSN 2211-1247.
https://doi.org/10.1016/j.celrep.2019.08.066.

Li L, Chen Z, von Scheidt M, et al. Transcriptome-wide association study of coronary artery disease identifies novel susceptibility genes [published correction appears in Basic Res Cardiol. 2022 Apr 5;117(1):19. doi: 10.1007/s00395-022-00923-w]. Basic Res Cardiol. 2022;117(1):6. Published 2022 Feb 17.

Liu W., Zhang Y., Yu C.M. Current understanding of coronary artery calcification. J Geriatr Cardiol. 2015;12(6):668-675.
https://doi.org/10.11909/j.issn.1671-5411.2015.06.012

Lloyd-Jones D.M., Nam B.H., D'Agostino R.B. Sr, et al. Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: a prospective study of parents and offspring. JAMA. 2004;291(18):2204-2211.
https://doi.org/10.1001/jama.291.18.2204

Lopaschuk GD, Karwi QG, Tian R, Wende AR, Abel ED. Cardiac Energy Metabolism in Heart Failure. Circ Res. 2021;128(10):1487-1513.
https://doi.org/10.1161/CIRCRESAHA.121.318241

López R., Regier C., Jordan M. I., Yosef N. Deep generative modeling for single-cell transcriptomics. Nature Methods. 2018;15:1053-1058.
https://doi.org/10.1038/s41592-018-0229-2

Love M.I., Huber W., Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15:550. 10.1186/s13059-014-0550-8

Love, M.I. Anders S., Huber W. Analyzing RNA-seq data with DESeq2. [internet]. Bioconductor. 2024. Available from:
www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Lowe R., Shirley N., Bleackley M., Dolan S., Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017;13(5):e1005457. Published 2017 May 18.
https://doi.org/10.1371/journal.pcbi.1005457

Lu X., Liu Z., Cui Q., et al. A polygenic risk score improves risk stratification of coronary artery disease: a large-scale prospective Chinese cohort study. Eur Heart J. 2022;43(18):1702-1711.
https://doi.org/10.1093/eurheartj/ehac093

Luecken M.D., Theis F.J. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15(6):e8746. Published 2019 Jun 19.
https://doi.org/10.15252/msb.20188746

Luisi P., García A., Berros J.M., et al. Fine-scale genomic analyses of admixed individuals re-

veal unrecognized genetic ancestry components in Argentina. PLoS One. 2020;15(7):e0233808. Published 2020 Jul 16.
https://doi.org/10.1371/journal.pone.0233808

Luo W., Friedman M.S., Shedden K., Hankenson K.D., Woolf P.J. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics. 2009;10:161. Published 2009 May 27.
https://doi.org/10.1186/1471-2105-10-161

Lyu B., Haque A. Deep Learning Based Tumor Type Classification Using Gene Expression Data. Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2018;8:89-96.
https://doi.org/10.1145/3233547.3233588

Ma S., Zhen Z. OmicsMapNet: Transforming omics data to take advantage of Deep Convolutional Neural Network for discovery. ArXiv abs/1804.05283 2018: n. pag.
https://arxiv.org/pdf/1804.05283

Maekawa S., Suzuki A., Sugano S., Suzuki Y. RNA Sequencing: From Sample Preparation to Analysis. Methods in Molecular Biology. Humana Press, New York, NY. 2014;1164:51-65.
https://doi.org/10.1007/978-1-4939-0805-9_6

Magnusson R., Tegnér J.N., Gustafsson M. Deep neural network prediction of genome-wide transcriptome signatures – beyond the Black-box. npj Syst Biol Appl. 2022;8,(9).
https://doi.org/10.1038/s41540-022-00218-9

Majtnerová P, Roušar T. An overview of apoptosis assays detecting DNA fragmentation. Mol Biol Rep. 2018;45(5):1469–1478.

Manolio T.A., Collins F.S., Cox N.J., Goldstein D.B., Hindorff L.A., Hunter D.J. Finding the missing heritability of complex diseases. Nature. 2009;461: 747–753.

Mardis E.R. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008;9:387–402.

Marenberg M.E., Risch N., Berkman L.F., Floderus B., de Faire U. Genetic susceptibility to death from coronary heart disease in a study of twins. N Engl J Med. 1994;330(15):1041-1046.
https://doi.org/10.1056/NEJM199404143301503

Mars N., Koskela J.T., Ripatti P., et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat Med. 2020;26(4):549-557.
https://doi.org/10.1038/s41591-020-0800-0

Martinez A. M., Kak A. C. PCA versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021;23(2):228–233.
https://doi.org/10.1109/34.908974

McInnes L., Healy J., Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2020 Cornell University arXiv:1802.03426v3[stat.ML]

Melé M. et al. The human transcriptome across tissues and individuals. Science. 2015;348:660-665 .DOI:10.1126/science.aaa0355

Merino D., Kelly G.L., Lessene G., Wei A.H., Roberts A.W., Strasser A. BH3-Mimetic Drugs: Blazing the Trail for New Cancer Medicines. Cancer Cell. 2018;34(6):879–891.

Metzker M.L. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11: 31–46.

Mi H., Muruganujan A., Casagrande J.T., Thomas P.D. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013;8(8):1551-1566.
https://doi.org/10.1038/nprot.2013.092

Miao L., Qin Y.A., Yang Z.J., et al. Identification of potential therapeutic targets for plaque vulnerability based on an integrated analysis. Nutr Metab Cardiovasc Dis. 2024;34(7):1649-1659.
https://doi.org/10.1016/j.numecd.2024.02.005

Minor L.B. Harnessing the power of data in health. Stanford Medicine 2017 Health Trends Report. 2017;2-6.
https://med.stanford.edu/content/dam/sm/sm-news/documents/StanfordMedicineHealthTrendsWhitePaper2017.pd
Accessed March 20 2020. [Ref list]

Moon J., Posada-Quintero H.F., Chon K.H. A literature embedding model for cardiovascular disease prediction using risk factors, symptoms, and genotype information. Expert Systems with Applications. 2023;213A:118930.
https://doi.org/10.1016/j.eswa.2022.118930

Mora M.T., Zaza A., Trenor B. Insights from an electro-mechanical heart failure cell model: Role of SERCA enhancement on arrhythmogenesis and myocyte contraction. Comput Methods Programs Biomed. 2023;230:107350.

Morrow G.B., Whyte C.S., Mutch N.J. Functional plasminogen activator inhibitor 1 is retained on the activated platelet membrane following platelet activation. Haematologica 2020;105(12):2824-2833.
https://doi.org/10.3324/haematol.2019.230367.

Mokry M., Boltjes A., Pasterkamp G. et al. Transcriptomic-based clustering of advanced atherosclerotic plaques identifies subgroups of plaques with differential underlying biology that associate with clinical presentation. medRxiv 2021.11.25.21266855; doi:
https://doi.org/10.1101/2021.11.25.21266855

Moonira T., Chachra S.S., Ford B.E., et al. Metformin lowers glucose 6-phosphate in hepatocytes by activation of glycolysis downstream of glucose phosphorylation.
J Biol Chem. 2020;295(10):3330-3346.
https://doi.org/10.1074/jbc.RA120.012533

Mosca L., Appel L.J., Benjamin E.J., et al. Evidence-based guidelines for cardiovascular disease prevention in women. American Heart Association scientific statement. Arterioscler Thromb Vasc Biol. 2004;24(3):e29-e50.
https://doi.org/10.1161/01.ATV.0000114834.85476.81

Newgard C.B., Brady M.J., O'Doherty R.M., Saltiel A.R. Organizing glucose disposal: emerging roles of the glycogen targeting subunits of protein phosphatase-1. Diabetes. 2000;49(12):1967-1977.
https://doi.org/10.2337/diabetes.49.12.1967

Newman S., Nakitandwe J., Kesserwan C.A., et al. Genomes for Kids: The Scope of Pathogenic Mutations in Pediatric Cancer Revealed by Comprehensive DNA and RNA Sequencing. Cancer Discov. 2021;11(12):3008-3027.
https://doi.org/10.1158/2159-8290.CD-20-1631

Novack M.L., Zubair M. Natriuretic Peptide B Type Test. [Updated 2023 Apr 23]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from:
www.ncbi.nlm.nih.gov/books/NBK556136/

O'Sullivan J.W., Raghavan S., Marquez-Luna C., et al. Polygenic Risk Scores for Cardiovascular Disease: A Scientific Statement From the American Heart Association. Circulation. 2022;146(8):e93-e118.
https://doi.org/10.1161/CIR.0000000000001077

Oei H.H., Vliegenthart R., Hak A.E., et al. The association between coronary calcification assessed by electron beam computed tomography and measures of extracoronary atherosclerosis: the Rotterdam Coronary Calcification Study. J Am Coll Cardiol. 2002;39(11):1745-1751.
https://doi.org/10.1016/s0735-1097(02)01853-3

Örd T., Lönnberg T., Nurminen V., et al. Dissecting the polygenic basis of atherosclerosis via disease-associated cell state signatures. Am J Hum Genet. 2023;110(5):722-740.
https://doi.org/10.1016/j.ajhg.2023.03.013

Pacher P., Beckman J.S., Liaudet L. Nitric oxide and peroxynitrite in health and disease. Physiol Rev. 2007;87(1):315-424.
https://doi.org/10.1152/physrev.00029.2006

Panahiazar M., Taslimitehrani V., Pereira N., Pathak J. Using EHRs and Machine Learning for Heart Failure Survival Analysis. Stud Health Technol Inform. 2015;216:40-44.

Pandey D., Onkara P.P. Improved downstream functional analysis of single-cell RNA-sequence data using DGAN. Scientific Reports. 2023;13(1):1618.
https://doi.org/10.1038/s41598-023-28952-y

Park J.-H., Gail M.H., Weinberg C.R., Carroll R.J., Chung C.C., Wang Z. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci U S A. 2011;108: 18026–18031.

Park J., Kim H., Kim J., Cheon M. A practical application of generative adversarial networks for RNA-seq analysis to predict the molecular progress of Alzheimer's disease. PLoS Comput Biol. 2020;16(7):e1008099. Published 2020 Jul 24.
https://doi.org/10.1371/journal.pcbi.1008099

Park J.J. Epidemiology, Pathophysiology, Diagnosis and Treatment of Heart Failure in Diabetes [published correction appears in Diabetes Metab J. 2021;45(5):796]. Diabetes Metab J. 2021;45(2):146-157.
https://doi.org/10.4093/dmj.2020.0282

Patel J., Pallazola V.A., Dudum R., et al. Assessment of Coronary Artery Calcium Scoring to Guide Statin Therapy Allocation According to Risk-Enhancing Factors: The Multi-Ethnic Study of Atherosclerosis. JAMA Cardiol. 2021;6(10):1161-1170.
https://doi.org/10.1001/jamacardio.2021.2321

Pedroso J.A.B., Ramos-Lobo A.M., Donato J. Jr. SOCS3 as a future target to treat metabolic disorders. Hormones (Athens). 2019;18(2):127-136.
https://doi.org/10.1007/s42000-018-0078-5

Peng M., Hou F., Cheng Z. et al. Prediction of cardiovascular disease risk based on major contributing features. Sci Rep. 2023;13:4778
https://doi.org/10.1038/s41598-023-31870-8

Peng S., Wang M., Zhang S. Hydrogen sulfide regulates SERCA2a SUMOylation by S-Sulfhydration of SENP1 to ameliorate cardiac systole-diastole function in diabetic cardiomyopathy. Biomed Pharmacother. 2023;160:114200.
https://doi.org/10.1016/j.biopha.2022.114200

Peters M.J., Joehanes R., Pilling L.C., et al. The transcriptional landscape of age in human peripheral blood. Nat Commun. 2015;6:8570. Published 2015 Oct 22.
https://doi.org/10.1038/ncomms9570

Peymani F., Farzeen A., Prokisch H. RNA sequencing role and application in clinical diagnostic. Pediatr Investig. 2022;6(1):29-35. Published 2022 Mar 5.
https://doi.org/10.1002/ped4.12314

Pinali C, Malik N, Davenport JB, et al. Post-Myocardial Infarction T-tubules Form Enlarged Branched Structures With Dysregulation of Junctophilin-2 and Bridging Integrator 1 (BIN-1). J Am Heart Assoc. 2017;6(5):e004834. Published 2017 May 4.

Powe C.E., Levine R.J., Karumanchi S.A. Preeclampsia, a disease of the maternal endothelium: the role of antiangiogenic factors and implications for later cardiovascular disease. Circulation. 2011;123(24):2856-2869.
https://doi.org/10.1161/CIRCULATIONAHA.109.853127

Poznyak A.V., Zhang D., Orekhova V., Grechko A.V., Wetzker R., Orekhov A.N. A brief overview of currently used atherosclerosis treatment approaches targeting lipid metabolism alterations. Am J Cardiovasc Dis. 2020;10(2):62-71. Published 2020 Jun 15.

Poznyak A.V., Sukhorukov V.N., Guo S., Postnov A.Y., Orekhov A.N. Sex Differences Define the Vulnerability to Atherosclerosis. Clin Med Insights Cardiol. 2023;17:11795468231189044. Published 2023 Jul 29.
https://doi.org/10.1177/11795468231189044

Questa M, Romorini L, Blüguermann C, Solari CM, Neiman G, Luzzani C, et al. Generation of iPSC line iPSC-FH2.1 in hypoxic conditions from human foreskin fibroblasts. Stem Cell Res. 2016;16(2):300–3.

Ramirez Flores R.O., Lancer j.D., Holland C.H., Leuschner F., Most P., Schultz J-H., Levinson R.T., Saez-Rodriguez J. Consensus Transcriptional Landscape of Human End-Stage Heart Failure. Journal of the American Heart Association. 2021;10(7):e019667 .
https://doi.org/10.1161/JAHA.120.019667

Ridker P.M. Clinical application of C-reactive protein for cardiovascular disease detection and prevention. Circulation. 2003;107(3):363-369.
https://doi.org/10.1161/01.cir.0000053730.47739.3c

Riveros-Mckay F., Weale M.E., Moore R., et al. Integrated Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction [published correction appears in Circ Genom Precis Med. 2021 Aug;14(4):e000085.

doi: 10.1161/HCG.0000000000000085]. Circ Genom Precis Med. 2021;14(2):e003304.
https://doi.org/10.1161/CIRCGEN.120.003304

Romero-Corral A., Somers V.K., Sierra-Johnson J. Accuracy of body mass index in diagnosing obesity in the adult general population. Int J Obes (Lond). 2008;32(6):959-966.
https://doi.org/10.1038/ijo.2008.11

Rosenberg S., Elashoff M.R., Beineke P., et al. Multicenter validation of the diagnostic accuracy of a blood-based gene expression test for assessing obstructive coronary artery disease in nondiabetic patients. Ann Intern Med. 2010;153(7):425-434.
https://doi.org/10.7326/0003-4819-153-7-201010050-00005

Rowe G.C., Jiang A., Arany Z. PGC-1 coactivators in cardiac development and disease. Circ Res. 2010;107(7):825-838.
https://doi.org/10.1161/CIRCRESAHA.110.223818

Ruderman NB, Carling D, Prentki M, Cacicedo JM. AMPK, insulin resistance, and the metabolic syndrome. J Clin Invest. 2013;123(7):2764-2772.
https://doi.org/10.1172/JCI67227

Rumelhart D., Hinton G., Williams, R. Learning representations by back-propagating errors. Nature. 1986;323:533-536.
https://doi.org/10.1038/323533a0

Russakovsky, O., Deng, J., Su, H. et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis. 2015;115:211–252.
https://doi.org/10.1007/s11263-015-0816-y

Saad Shaukat, MH, Stys, P, Sjovold, A, Hajek, C, Petrasko, P, Pham, M, Stys, V, Petrasko, M, Rynders, B, Singh, K, et al. Concordance of high polygenic CAD Risk score with high coronary artery calcium score & low polygenic CAD risk score with low coronary artery calcium score: a real world experience. Circulation. 2021;144(suppl):A12277–A12277

Saidon P., Andrade N. S., Johnson E., Granados G. G., Pascual M. A., Woollen S., Aguilar M., Ruiz C. M., Arabetti C. Buenas prácticas clínicas: Documento de las Américas. 2005. Conferencia panamericana para la armonización de la reglamentación farmacéutica [on line]. Available from: www.ms.gba.gov.ar/ssps/investigacion/DocTecnicos/BuenasPracticas-DocAmericas.pdf

Salceda, R. Peroxisomas: Organelos polifacéticos.
Revista de Educación Bioquímica. 2008;27(3):85-92

Salminen A, Kaarniranta K, Kauppinen A. Insulin/IGF-1 signaling promotes immunosuppression via the STAT3 pathway: impact on the aging process and age-related diseases. Inflamm Res.

2021;70(10-12):1043-1061.
https://doi.org/10.1007/s00011-021-01498-3

Sameer AS, Nissar S. Toll-Like Receptors (TLRs): Structure, Functions, Signaling, and Role of Their Polymorphisms in Colorectal Cancer Susceptibility. Biomed Res Int. 2021;2021:1157023. Published 2021 Sep 12.
https://doi.org/10.1155/2021/1157023

Saremi A., Bahn G., Reaven P.D. Progression of vascular calcification is increased with statin use in the veterans affairs diabetes trial (VADT) Diabetes Care. 2012;35:2390–2392.

Savji N., Rockman C.B., Skolnick A.H., et al. Association between advanced age and vascular disease in different arterial territories: a population database of over 3.6 million subjects. J Am Coll Cardiol. 2013;61(16):1736-1743.
https://doi.org/10.1016/j.jacc.2013.01.054

Savarese G., Becher P.M., Lund L.H., Seferovic P., Rosano G.M.C., Coats A.J.S. Global burden of heart failure: a comprehensive and updated review of epidemiology [published correction appears in Cardiovasc Res. 2023;119(6):1453]. Cardiovasc Res. 2023;118(17):3272-3287.
https://doi.org/10.1093/cvr/cvac013

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (New York, N.Y.), 1995;270(5235):467–470.
https://doi.org/10.1126/science.270.5235.467

Schmermund A., Möhlenkamp S., Stang A., et al. Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the Heinz Nixdorf RECALL Study. Risk Factors, Evaluation of Coronary Calcium and Lifestyle. Am Heart J. 2002;144(2):212-218.
https://doi.org/10.1067/mhj.2002.123579

Sedelnikova OA, Pilch DR, Redon C, Bonner WM. Histone H2AX in DNA damage and repair. Cancer Biol Ther. 2003;2(3):233–5.

Sharma P. ,Narinder S Sahni S. N., Tibshirani R., Skaane P., Urdal P, Berghagen H., Jensen M., Kristiansen L., Moen C., Sharma P., Zaka A., Arnes J., Sauer T., Akslen L. A., Schlichting E., Børresen-Dale A, Lönneborg A. Early detection of breast cancer based on gene-expression patterns in peripheral blood cells. Breast Cancer Res. 2005;7:634–644

Sheriff A., Kayser S., Brunner P., Vogt B. C-Reactive Protein Triggers Cell Death in Ischemic Cells. Front Immunol. 2021;12:630430. Published 2021 Feb 10.
https://doi.org/10.3389/fimmu.2021.630430

Shreya D., Zamora D.I., Patel G.S., et al. Coronary Artery Calcium Score - A Reliable Indicator of Coronary Artery Disease?. Cureus. 2021;13(12):e20149. Published 2021 Dec 3.
https://doi.org/10.7759/cureus.20149

Singh R.B., Mengi S.A., Xu Y.J., Arneja A.S., Dhalla N.S. Pathogenesis of atherosclerosis: A multifactorial process. Exp Clin Cardiol. 2002;7(1):40-53.

Singh M., Kumar A., Khanna N.N., et al. Artificial intelligence for cardiovascular disease risk assessment in personalized framework: a scoping review. EClinicalMedicine. 2024;73:102660. Published 2024 May 27.
https://doi.org/10.1016/j.eclinm.2024.102660

Sima P., Vannucci L., Vetvicka V. Atherosclerosis as autoimmune disease. Ann Transl Med. 2018;6(7):116.
https://doi.org/10.21037/atm.2018.02.02

Skelly D.A., Ronald J., Akey J.M. Inherited Variation in Gene Expression. Annual Review of Genomics and Human Genetics. 2009;10:313-332.
https://doi.org/10.1146/annurev-genom-082908-150121

Sopić M., Karaduzovic-Hadziabdic K., Kardassis D., Maegdefessel L., Martelli F., Meerson A., Munjas J., Niculescu L.S., Stoll M., Magni P, Devaux Y.Transcriptomic research in atherosclerosis: Unravelling plaque phenotype and overcoming methodological challenges. Journal of Molecular and Cellular Cardiology Plus. 2023;6:100048. ISSN 2772-9761.
https://doi.org/10.1016/j.jmccpl.2023.100048

Sordet O, Redon CE, Guirouilh-Barbat J, Smith S, Solier S, Douarre C, et al. Ataxia telangiectasia mutated activation by transcription- and topoisomerase I-induced DNA double-strand breaks. EMBO Rep. 2009;10(8):887–93.

Stammers A.N. Susser S.E., Hamm N.C. The regulation of sarco(endo)plasmic reticulum calcium-ATPases (SERCA). Can J Physiol Pharmacol. 2015;93(10):843-854.
https://doi.org/10.1139/cjpp-2014-0463

Stanaway J.D., Roth G. The burden of Chagas disease: estimates and challenges. Glob Heart. 2015;10(3):139-144.
https://doi.org/10.1016/j.gheart.2015.06.001

Steenman M., Espitia O., Maurel B., et al. Identification of genomic differences among peripheral arterial beds in atherosclerotic and healthy arteries. Sci Rep. 2018;8(1):3940. Published 2018 Mar 2.
https://doi.org/10.1038/s41598-018-22292-y

Steensma D.P., Bejar R., Jaiswal S., et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. Blood. 2015;126(1):9-16.
https://doi.org/10.1182/blood-2015-03-631747

Strumberg D., Pilon A.A., Smith M., Hickey R., Malkas L., Pommier Y. Conversion of topoisomerase I cleavage complexes on the leading strand of ribosomal DNA into 5'-phosphorylated DNA double-strand breaks by replication runoff. Mol Cell Biol. 2000;20(11):3977–87.

Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, Lahr DL, Hirschman JE, Liu Z, Donahue M, Julian B, Khan M, Wadden D, Smith IC, Lam D, Liberzon A, Toder C, Bagul M, Orzechowski M, Enache OM, Piccioni F, Johnson SA, Lyons NJ, Berger AH, Shamji AF, Brooks AN, Vrcic A, Flynn C, Rosains J, Takeda DY, Hu R, Davison D, Lamb J, Ardlie K, Hogstrom L, Greenside P, Gray NS, Clemons PA, Silver S, Wu X, Zhao WN, Read-Button W, Wu X, Haggarty SJ, Ronco LV, Boehm JS, Schreiber SL, Doench JG, Bittker JA, Root DE, Wong B, Golub TR. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. Nov 2017;171(6):1437-1452.e17.
doi: 10.1016/j.cell.2017.10.049.

Szkudelski T., Szkudelska K. The relevance of AMP-activated protein kinase in insulin-secreting β cells: a potential target for improving β cell function?. J Physiol Biochem. 2019;75(4):423-432.
https://doi.org/10.1007/s13105-019-00706-3

Szumilas M. Explaining odds ratios. J Can Acad Child Adolesc Psychiatry. 2010;19: 227–229.

Takahashi T., Hato F., Yamane T. Activation of human neutrophil by cytokine-activated endothelial cells. Circ Res. 2001;88:422–9

Taslimitehrani V., Dong G., Pereira N.L., Panahiazar M., Pathak J. Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function. J Biomed Inform. 2016;60:260-269.
https://doi.org/10.1016/j.jbi.2016.01.009

Thorat, S. B., Banarjee S. k., Gaikwad D. D., Jadhav S. L., Thorat R. M. Clinical Trial: A review. Vishal Institute of Pharmaceutical Education And Research. 2010;1(2):019. ISSN 0976–044X

Tikkanen E., Havulinna A.S., Palotie A., Salomaa V., Ripatti S. Genetic risk prediction and a 2-stage risk screening strategy for coronary heart disease.
Arterioscler Thromb Vasc Biol. 2013;33(9):2261-2266.
https://doi.org/10.1161/ATVBAHA.112.301120

Tran B., Tran D., Nguyen H., Ro S., Nguyen T. scCAN: single-cell clustering using autoencoder and network fusion. Scientific Reports. 2022;12(1):10267.
https://doi.org/10.1038/s41598-022-14218-6

Triposkiadis F., Xanthopoulos A., Starling R.C., Iliodromitis E. Obesity, inflammation, and heart failure: links and misconceptions. Heart Fail Rev. 2022;27(2):407-418.
https://doi.org/10.1007/s10741-021-10103-y

Thompson A.L. Developmental origins of obesity: early feeding environments, infant growth, and the intestinal microbiome. Am J Hum Biol. 2012;24(3):350-360.
https://doi.org/10.1002/ajhb.22254

Tcheandjieu C., Zhu X., Hilliard A.T., et al. Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. Nat Med. 2022;28(8):1679-1692.
https://doi.org/10.1038/s41591-022-01891-3

Voora D., Coles A., Lee K.L., et al. An age- and sex-specific gene expression score is associated with revascularization and coronary artery disease: Insights from the Prospective Multicenter Imaging Study for Evaluation of Chest Pain (PROMISE) trial. Am Heart J. 2017;184:133-140.
https://doi.org/10.1016/j.ahj.2016.11.004

Waddell H.M.M., Mereacre V., Alvarado F.J., Munro M.L. Clustering properties of the cardiac ryanodine receptor in health and heart failure. J Mol Cell Cardiol. 2023;185:38-49.
https://doi.org/10.1016/j.yjmcc.2023.10.012

Walley A.J., Asher J.E., Froguel P. The genetic contribution to non-syndromic human obesity. Nat Rev Genet. 2009;10(7):431-442.
https://doi.org/10.1038/nrg2594

Wang N, Hua J, Fu Y, et al. Updated perspective of EPAS1 and the role in pulmonary hypertension. Front Cell Dev Biol. 2023;11:1125723. Published 2023 Feb 27.
https://doi.org/10.3389/fcell.2023.1125723

Wang Y., Lv Q., Wu H. Comparison of MESA of and Framingham risk scores in the prediction of coronary artery disease severity. Herz. 2020;45(Suppl 1):139-144.
https://doi.org/10.1007/s00059-019-4838-z

Wang Z., Gerstein M., Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63.
https://doi.org/10.1038/nrg2484

Weale M.E., Riveros-Mckay F., Selzam S., et al. Validation of an Integrated Risk Tool, Including Polygenic Risk Score, for Atherosclerotic Cardiovascular Disease in Multiple Ethnicities and Ancestries. Am J Cardiol. 2021;148:157-164.
https://doi.org/10.1016/j.amjcard.2021.02.032

Weihrauch-Blüher S., Schwarz P., Klusmann J.H. Childhood obesity: increased risk for cardio-

metabolic disease and cancer in adulthood. Metabolism. 2019;92:147-152.
https://doi.org/10.1016/j.metabol.2018.12.001

Woodward M. Cardiovascular Disease and the Female Disadvantage. Int J Environ Res Public Health. 2019;16(7):1165.
https://doi.org/10.3390/ijerph16071165

Wu T., Hu E., Xu S. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2(3):100141. Published 2021 Jul 1.
https://doi.org/10.1016/j.xinn.2021.100141

Wu C, Zhang Z, Zhang W, Liu X. Mitochondrial dysfunction and mitochondrial therapies in heart failure. Pharmacol Res. 2022;175:106038.
https://doi.org/10.1016/j.phrs.2021.106038

Xing Y., Yang X., Chen H. The effect of cell isolation methods on the human transcriptome profiling and microbial transcripts of peripheral blood. Mol Biol Rep. 2021;48(4):3059-3068.
https://doi.org/10.1007/s11033-021-06382-1

Xu J., Liao K., Yang X., Wu C., Wu W. Using single-cell sequencing technology to detect circulating tumor cells in solid tumors [published correction appears in Mol Cancer. 2022 Apr 18;21(1):100.
doi: 10.1186/s12943-022-01564-2]. Mol Cancer. 2021;20(1):104.
https://doi.org/10.1186/s12943-021-01392-w

Yépez VA, Gusic M, Kopajtich R, et al. Clinical implementation of RNA sequencing for Mendelian disease diagnostics. Genome Med. 2022;14(1):38. Published 2022 Apr 5.

Zeleznik R., Foldyna B., Eslami P. et al. Deep convolutional neural networks to predict cardiovascular risk from computed tomography. Nat Commun. 2021;12:715.
https://doi.org/10.1038/s41467-021-20966-2

Zhao S., Ye Z., Stanton R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. RNA. 2020;26(8):903-909.
https://doi.org/10.1261/rna.074922.120

Zhao, Y., Li, MC., Konaté, M.M. et al. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. J Transl Med 2021;19:269.
https://doi.org/10.1186/s12967-021-02936-w

Zhou B., Tian R. Mitochondrial dysfunction in pathophysiology of heart failure. J Clin Invest.

2018;128(9):3716-3726.
https://doi.org/10.1172/JCI120849

Zhou J., Troyanskaya O. G. Predicting effects of noncoding variants with deep learning-based sequence mode. Nature Methods. 2015;12:931-934.
https://doi.org/10.1038/nmeth.3547

Zhou K., Cai C., He Y., Chen Z. Potential prognostic biomarkers of sudden cardiac death discovered by machine learning. Computers in Biology and Medicine. 2022;150:0010-4825.
https://doi.org/10.1016/j.compbiomed.2022.106154

Zdravkovic S., Wienke A., Pedersen N.L., Marenberg M.E., Yashin A.I., De Faire U. Heritability of death from coronary heart disease: a 36-year follow-up of 20966 Swedish twins. Journal of internal medicine. 2002;252(3):247-254