GUIÓN:

Tradicionalmente, cuando pensamos en **análisis de datos**, imaginamos a grandes empresas estudiando cómo vender más o cómo reducir costes económicos. Sin embargo, los datos tienen un poder mucho más profundo.

Hoy queremos demostrar que las mismas herramientas que se usan para entender el mercado pueden usarse para **entender la vida y la enfermedad**. Analizar una base de datos biológica no es solo mirar números; es descifrar el mensaje oculto en nuestras células para intentar mejorar la salud de las personas.

¿Qué es el Cáncer y cómo progresa?

Para entender nuestro análisis, primero debemos saber qué estamos combatiendo. El cáncer empieza cuando una célula de nuestro cuerpo muta genéticamente. El ADN (la base molecular de nuestros genes) es lo que guía a la célula en su ciclo de vida y en la realización de sus tareas, de forma que su expresión es la que controla las pautas que debe seguir una célula para su correcto crecimiento, desarrollo y división.

La mutación de estos genes, provoca que la célula deje de comportarse de forma natural, y empiece a dividirse sin control (evadiendo la muerte celular), provocando crecimientos de masas anormales (lo que conocemos como tumores).

La cuestión es que, aunque tratemos de agrupar en una sola enfermedad a esta división descontrolada, lo cierto es que no hay dos cánceres iguales. Al igual que cada persona tiene una huella dactilar única, cada tumor desarrolla un **perfil molecular específico**.

Aún así, los médicos clasifican el cáncer en etapas (del I al IV). El **Estadio I** suele ser un tumor pequeño y localizado, mientras que el **Estadio IV** indica que la enfermedad es más agresiva o se ha extendido. Incluso el mismo cáncer en la misma persona no tiene el mismo nivel de expresión si este se le diagnostica antes o después (la rápida división de la masa tumoral hace que se produzcan más mutaciones todavía).

El Cáncer de Tiroides: Nuestro caso de estudio

Nuestro dataset agrupa datos clínicos y genéticos de muestras tomadas de pacientes con cáncer de tiroides. El cáncer de tiroides es la neoplasia endocrina más común, siendo el Carcinoma Papilar (THCA) su variante más frecuente. Se diferencian también en este dataset datos de pacientes que también presentan la variante folicular (poco agresiva) y la Tall Cell, con una tendencia a ser más agresiva.

¿Por qué buscamos "Biomarcadores"?

Aunque el cáncer sea diferente en cada paciente, existen genes driver que actúan activando/desactivando rutas metabólicas completas o regulando la expresión de otros genes. Estos se pueden considerar biomarcadores de diagnóstico, y son necesarios para construir una medicina de precisión:

- Ayudan a confirmar diagnósticos de certeza que son imprecisos con pruebas físicas o ecografías. La expresión o silenciamiento de ciertos genes en una muestra tumorigénica nos puede asegurar que un tumor sea benigno o maligno.
- Predicción del comportamiento y selección de terapias dirigidas hacia esos focos.

Acompañanos en este viaje por las redes genéticas humanas para descubrir que se esconde tras el cáncer de tiroides. ¿Seremos capaces de encontrar patrones que nos permitan hipotetizar sobre posibles biomarcadores de la progresión o, por el contrario, el perfil genético de este tipo de cáncer en concreto será completamente heterogéneo?

VISUALIZACIÓN DE LOS DATOS

Encontramos esta información sobre los pacientes:

- Su edad (pacientes de entre 15 y 89 años)
- Género
- Estado del cancer… entre otras

Además de información sobre la muestra recogida:

- La expresión normalizada de sus genes
- Su pureza tumoral, que de hecho hemos usado en la limpieza de nuestros datos (lo explicamos en el análisis)

PREPARACIÓN DE ENTORNO: CARGA, LIMPIEZA Y NORMALIZACIÓN

Los datos biológicos suelen ser ruidosos. En esta fase, tratamos los valores no deseados:

1. Nos desharemos de los nulos para nuestras variables directoras.
2. Los datos de expresión de los genes ya se encuentran normalizados (por lo que nos ahorra trabajo).
   Se trabajó con datos transformados en log_2, lo que permite comparar genes con niveles de abundancia muy dispares de forma estadísticamente justa. Esos números representan el nivel de abundancia de los mensajeros (mRNA) de un gen en la muestra, pero no están en unidades naturales (como 1, 2, 3 copias), sino en una escala logarítmica.
3. Filtrado para eliminar “ruido”: De los 19,927 genes iniciales, se eliminaron aquellos con baja expresión media (por debajo de 0.5) y baja variabilidad (percentil 25), quedándonos únicamente con los genes que realmente aportan información diferencial entre pacientes. Explicación:
   - Filtro de Media (mean>0.5):Ciertos genes tienen casi todos sus datos de expresión a 0. Esos genes están apagados o son errores de lectura. Si la media es muy baja, el gen no tiene suficiente señal biológica para ser confiable.
   - Filtro de Varianza (Percentil 25): Si casi todos los pacientes tienen el mismo valor no obtendremos expresión diferencial. Si haces un análisis para ver qué genes causan una enfermedad, este gen no te dirá nada porque no cambia entre pacientes.

   *Inciso: Aun así, esta variabilidad es entre todas las muestras. Aunque hayamos sacado información de los genes que presenten mayor variabilidad, estos no tienen por qué ser posibles biomarcadores, para ello tiene que existir variabilidad durante el progreso de la enfermedad (lo veremos más adelante en el análisis bivariante).

PREGUNTAS PREVIAS

A partir de la visualización de todas las variables clínicas, y tras la limpieza, sacamos las hipótesis que intentaremos responder durante el desarrollo del proyecto:

- **¿Existen características demográficas que afecten al diagnóstico/agresión de la enfermedad?**
- **¿Se observan diferencias clave que permitan admitir que hay tipos histológicos con mayor agresividad?**
- **Diseminación: La presencia de ganglios (y el numero de ellos) y la metástasis, ¿están relacionadas?¿Y con el estadío?¿Y con la edad?**
- **¿Existe relación con la supervivencia para aquellos pacientes ya fallecidos?**
- **¿Hay alguna relación entre los pacientes que han recibido radioterapia con una mayor o menor presencia de tumor residual?**

Y la más importante:

- **¿Se observan diferencias en el perfil molecular entre tumores en estadíos tempranos frente a los tardíos? ¿o entre diferentes tipos histológicos?**
  Con esta nos enfocamos en la búsqueda de posibles biomarcadores

ANÁLISIS UNIVARIANTE

Luego, a raíz del análisis univariante obtenemos diferentes conclusiones y más preguntas. Para comenzar, hemos clasificado nuestras variables clínicas en función del tipo de dato que presentan. Luego, a las categóricas les hemos calculado su distribución y a las numéricas las hemos sometido a un boxplot y hemos sacado su histograma y correspondiente función de densidad.
Observaciónes variables categóricas:

- Los **estadios** muestran que hay mayor cantidad datos recogidos de pacientes del estadio I que para el resto. Antes de agruparlos aún mas (ya que normalmente se agrupan estadio I y II por una parte, reconociéndose como estadios tempranos, y III y IV por otra como tardíos) nos gustaría ver si hay mayor parecido molecular entre I y II o entre II y III/IV. En el segundo de los casos, nos vendría mejor la agrupación, pues tendríamos mas o menos la misma distribución de valores para las agrupaciones. (esto lo haremos en análisis bivariante)
- Hay buena distribución entre aquellos cuyos tumores se han expandido a los **nodos linfáticos** y aquellos que tienen el tumor todavía focalizado en la zona en la que se produjo. Lo mismo nos conviene centrarnos mejor en esta agrupación para el estudio de la expresión de los genes en la progresión de la enfermedad
- Pocos datos de pacientes con **metástasis**, nos interesaría saber si aquellos pocos para los que hay datos (ya que hay gran cantidad de pacientes para los que no se ha recogido esta información) siguen vivos.
- Se observa la predisposición del cáncer al tipo clásico, los otros tipos son más raros.
- De este gráfico nos surge la duda de si este tipo de cáncer es más frecuente en mujeres.
- Para el resto del análisis no hemos tenido en cuenta la raza/etnia,puesto que intuimos que hay un sesgo en la población estudiada.

- Con la distribución del estatus observamos la proporción baja de muertos- es cancer poco agresivo

Observaciones variables numéricas:

Respecto a las variables numéricas, el cuadro clínico es el siguiente:

- **Diversidad en Edad y Afectación Ganglionar:** La distribución de la edad y el número de ganglios afectados cubre todo el espectro clínico, desde pacientes jóvenes sin afectación hasta casos complejos con múltiples metástasis ganglionares. Esta amplia distribución es lo que nos permite afirmar que los hallazgo no son sólo estadísticos, sino un reflejo de la biología real del cáncer de tiroides en diferentes tipos de personas. Aun así, hay mayor proporción de ciertos parámetros, lo que podría opacar la búsqueda de resultados para otros casos menos representados.
- **La pureza tumoral** la hemos tomado como parámetro para aumentar la limpieza de nuestro dataframe como hemos dicho antes. De modo que sólo aquellas muestras con más de un 60% de pureza serán examinadas en el análisis. Esto no es más que una forma de aportar mayor robustez a nuestros resultados (de hecho, este umbral debe superarse siempre en conjuntos de datos usados para investigación.

  Un tumor no es una masa aislada de cáncer; es un ecosistema. La pureza disminuye cuando hay una alta presencia de:

  - **Células del Estroma:** Fibroblastos y vasos sanguíneos que el tumor "recluta" para crecer.
  - **Células Inmunes:** Linfocitos o macrófagos que el cuerpo envía para atacar al tumor o que el tumor manipula para protegerse.

  La pureza afecta directamente la interpretación de los resultados:

  - **Dilución de la señal:** Si una muestra tiene solo un 30% de pureza, la señal de los genes que son exclusivos de las células cancerosas será más débil porque está "diluida" por el ARN de las células normales.
  - **Falsos positivos/negativos:** Algunos genes que marcaste como diferenciales podrían no ser del cáncer en sí, sino de la respuesta del cuerpo (sistema inmune) que aumenta en estadios avanzados.
- La supervivencia global representa el **tiempo** (normalmente en días) desde el diagnóstico hasta el último contacto.
  - Para los que tienen **status 0**, ese número es simplemente cuánto tiempo llevaban vivos la última vez que se tomaron muestras de ellos para el estudio.
  - Para los que tengan **status 1**, ese número indica exactamente cuántos días sobrevivieron

  Nos interesaría quizás hacer una separación para observar los valores de expresión de genes para aquellos pacientes que ya hayan fallecido.

Se identificaron outliers para estos 3 valores, y, tal y como hemos dicho, eliminamos aquellos que no superaran el 60% de pureza tumoral. Con el número de supervivencia no

hicimos nada, porque, lejos de aportarnos valor real sobre el pronóstico de la enfermedad, solo nos aporta datos sobre cuando fue diagnosticado el paciente.

Además, un punto crítico que hemos identificado al analizar la distribución de nuestros datos es que la **mediana de supervivencia parece inusualmente baja** para una enfermedad que, por definición, progresa de forma muy lenta. Debemos interpretar este dato con cautela y rigor**:** En el cáncer de tiroides, los pacientes pueden vivir décadas tras el diagnóstico. Si nuestra mediana de supervivencia es corta, no significa necesariamente que la enfermedad sea letal a corto plazo, sino que probablemente **no disponemos de datos clínicos suficientes a largo plazo** para los pacientes diagnosticados recientemente o en etapas muy tempranas. Nos hubiese gustado encontrar datos sobre la remisión del cancer, quizás le hubiesemos encontrado algun sentido a estos datos.

Podemos observar los histogramas y funciones de densidad de edad (bien distribuida) y pureza (con los outliers ya eliminados), y por otro lado, hemos categorizado la variable de número de gánglios linfáticos (ahora tenemos pacientes sin diseminación, de 1 a 5 ganglios afectados y con más de 5. También hemos guardado a los pacientes con valores nulos.

Con respecto al análisis de la expresión genética, no podíamos permitirnos hacer un análisis univariante gen a gen, por lo que estudiamos su distribución en función de su expresión media. A nivel genético, la expresión media de nuestras muestras sigue una **distribución normal (Gaussiana)** casi perfecta, (los datos han sido previamente normalizados).

ANÁLISIS BIVARIANTE

Hemos empezado por el análisis donde las dos variables en estudio son categóricas.

1. Comparación del Estadío con la afectación ganglionar

   Podemos observar como si que hay una relación significativa entre el estadio y la afectación ganglionar, con un mayor porcentaje de expansión a los nodos linfáticos en estadíos más avanzados. Esto, en realidad, no había que comprobarlo ya que en realidad los estadíos son categorizaciones que utilizan los médicos para agrupar a los pacientes en grupos en función de cómo de avanzada esté la enfermedad.

   Vemos que para el estadío II hay una frecuencia mucho menor (tabla derecha) de diseminación a los gánglios linfáticos.

2. Con la metástasis ocurre lo mismo, la presentan grupos más avanzados, pero nunca una persona clasificada como en el estadío I. Normalmente, en la mayoria de cánceres, la metástasis está directamente relacionada con personas en el estadio IV. En este caso, hay personas en el estadío II (3 en particular) que ya lo presentan.

   Buscando en la literatura de este cáncer descubrimos que en realidad estos análisis tan solo le aportan robustez a nuestros datos. Como hemos explicado con anterioridad, los estadios no son más que una categorización artificial por parte de los médicos para clasificar a los pacientes: en este cáncer en particular el factor

edad (y lo veremos también en el análisis) afecta a esta clasificación, de tal forma que los jóvenes no son clasificados en estadio II hasta que no presentan metástasis, y aquellos pacientes mayores (>55), siguen la clasificación tradicional, reservando la metástasis a estadíos más tardíos.- Lo observaremos más tarde en el análisis multivariante

3. tipo histologico vs estadio

   El gráfico de barras confirma visualmente lo que sospechábamos:Mientras que el tipo **Clásico** tiene una gran base en el **Estadio I** (barra azul dominante), el tipo **Tall Cell** muestra un comportamiento radicalmente distinto: su barra más alta corresponde al **Estadio III** (barra verde). Esto demuestra que la variante *Tall Cell* tiende a diagnosticarse en estadios mucho más avanzados y es intrínsecamente más agresiva que la variante clásica

   Variante Folicular: Muestra un perfil similar al clásico pero con una presencia algo mayor de Estadio II. Sin embargo, esta variable está clasificada como menos agresiva. Buscaremos luego en función del perfil molecular qué conclusiones podemos sacar.

   El p-valor obtenido es extremadamente bajo. Es decir, la relación que se ve entre el tipo de tumor y el estadio no es casualidad. Hay una **dependencia biológica real** entre la histología del cáncer de tiroides y su capacidad de propagación. Sin embargo, hay que tener en cuenta que **Categoría "Otros":** Solo tienes datos para el Estadio I y IV. Al ser una muestra tan pequeña (tabla de frecuencias esperadas con valores menores a 5), los resultados para este grupo específico deben tomarse con cautela El enorme volumen de datos del Estadio I (en la gráfica de frecuencia, izquierda, no se aprecia, pero en la de la derecha se ve que hay muchisimas más muestras del tipo clásico) puede enmascarar lo que ocurre en los estadios tardíos.

   Los resultados aún así "encajan" con la realidad epidemiológica: el cáncer de tiroides suele detectarse en etapas tempranas (Estadio I) y el tipo Papilar Clásico es el más frecuente y de mejor pronóstico. El dataset es una representación fiel de la enfermedad, lo que da mucha validez a nuestros futuros hallazgos genómicos.

4. tipo histologico vs estadio_N: En el tipo **Clásico**, hay más pacientes con afectación ganglionar (**n1**: 168) que sin ella (**n0**: 141).Esto refuerza que es muy `probable que este cáncer salte a los ganglios del cuello incluso en sus variantes menos agresivas. El tipo **Tall Cell** mantiene esta tendencia, teniendo casi el doble de casos con ganglios afectados que sin ellos. Sin embargo, se observa que la variante folicular desentona con el resto, presentando un mucho menor numero de pacientes que presenten esa diseminación (esto concuerda más con nuestros datos).
5. tipo histologico vs Estadio M: La metástasis distantees extremadamente inusual en toda la cohorte, validando que la progresión es principalmente regional. Que el gráfico muestre 0 casos M1 para tall cell responde principalmente a una limitación de la muestra: el tamaño del subgrupo Tall Cell es muy pequeño comparado con el Clásico, y la metástasis a distancia es un evento extremadamente raro en tiroides que suele detectarse de forma tardía. Además, ya se advertía sobre la falta de datos registrados para la variable metástasis en muchos pacientes, lo que sugiere

que la agresividad del Tall Cell en el dataset queda mejor reflejada en su clara tendencia al **Estadio III** y a la afectación ganglionar (**N1**).

6. Con respecto a la progresión linfática y la metástasis: En principio esperaríamos que todos los valores de m1 se presentaran en n1, ya que se intuye que la progresión del cancer comienza siempre por la diseminación linfática y luego aparece la metástasis. Sin embargo, aunque la mayoría de los carcinomas de tiroides prefieren la vía linfática (hacia los ganglios) , el cáncer también puede viajar a través de la **sangre** (vía hematógena).

   Si las células cancerosas entran directamente en un vaso sanguíneo, pueden colonizar el pulmón o el hueso (**m1**) saltándose por completo la estación de los ganglios del cuello (**n0**).

   El caso es que en este caso, es casi igual de frecuente que la metástasis se presente con o sin diseminación linfatica. Esto puede deberse a dos casos:

   - Podría ocurrir que el paciente tenga micrometástasis en ganglios que no fueron extirpados en la cirugía, por lo que se registra oficialmente como **n0** (no se encontraron ganglios positivos en la muestra analizada) a pesar de tener ya enfermedad a distancia (**m1**).
   - O al error debido a la baja representación de pacientes con metástasis. No se puede afirmar nada con una muestra tan pequeña.

7. El análisis de estadío patológico vs género aporta un dato clínico muy relevante:
   - **Prevalencia Femenina**: Se confirma visualmente que el cáncer de tiroides es significativamente más frecuente en mujeres en todos los estadios.
   - **También observamos una Diferencia en Progresión**: El p-valor de **0.0349** indica una relación significativa. Aunque hay más mujeres en total, la proporción de hombres parece aumentar ligeramente en el **Estadio IV** en comparación con el Estadio I, lo que podría sugerir que, aunque los hombres se diagnostican menos, podrían presentar estadios algo más avanzados.
8. Sin embargo, no hay relación con el tipo histológico.
9. Relación entre radioterapia y tumor residual: A simple vista, parece que la radioterapia "empeora" el resultado porque hay más casos de tumor residual (r1). Sin embargo, en oncología clínica, esto suele interpretarse al revés:

   > **-Interpretación probable:** Los médicos suelen recetar radioterapia precisamente a los pacientes que tienen **tumores más agresivos, más grandes o más difíciles de extirpar**. Por lo tanto, es normal que en el grupo de "Sí radioterapia" haya más casos de tumor residual, no porque la terapia falle, sino porque esos pacientes ya tenían un pronóstico más complejo desde el inicio.
   >
   > Además, sin la recurrencia tampoco somos capaces de hacer un buen análisis de estos datos. No podemos sacar conclusiones, solo suponer.

10. Relación entre estadio patológico y perfil ganglionar:La gráfica confirma que el **Estadio IV** está fuertemente ligado a una **alta carga ganglionar (>5 ganglios)- se observa mejor en la de frecuencia-**, lo cual es un indicador de que el cáncer se ha extendido significativamente por el sistema linfático. Por el contrario, en el **Estadio I**, es más probable encontrar pacientes cuyos ganglios

están limpios. Faltan muchos datos para el perfil ganglionar de ciertos estadíos, como el II, por lo que no se puede hacer una apreciación a ciencia cierta.

11. Relación entre tipo histologico y perfil ganglionar: El tipo **Folicular** es notablemente más localizado (menos propenso a afectar ganglios, como ya habíamos observado antes), mientras que la categoría **Otros** muestra una clara tendencia a la diseminación linfática. **Importancia del Diagnóstico:** Esta gráfica sugiere que conocer el tipo histológico es fundamental para predecir si el paciente necesitará un tratamiento más agresivo en los ganglios.

Con respecto al análisis combinado entre categórica y numérica, podemos sacar:

1. Muestra la distribución de la **edad** de los pacientes según su **estadio patológico** (del I al IV). La gráfica sugiere una tendencia clara: el diagnóstico en **estadios tempranos (I)** ocurre con mayor frecuencia en personas **jóvenes**, mientras que los **estadios más avanzados (II, III y IV)** se presentan mayoritariamente en adultos de **mediana y tercera edad**. Aquí hemos agrupado los estadíos para obtener un histograma más claro, pero en ambos se puede ver la tendencia.

   -Solapamiento: Existe un área común entre los 45 y 65 años donde coinciden todos los estadios, lo que significa que en esa franja de edad es posible encontrar cualquier nivel de gravedad de la enfermedad.

2. Estudiando la edad frente al tipo histologico:La gráfica sugiere que el tipo histologico **Tall Cell** está más asociado a pacientes de edad madura, mientras que la categoría **Otros** identifica tumores que ocurren con mayor frecuencia en etapas más tempranas de la vida. Los tipos **Clásico** y **Folicular** se mantienen en un punto intermedio, afectando a un rango de edad muy diverso. Como con respecto a las muestras que parecen tener cierta tendencia tenemos tan pocos datos, no podemos dar ninguna respuesta clave.

3. Al observar el boxplot de **Supervivencia Global (solo para fallecidos) por Estadio (**ya que estas son las únicas muestras donde este dato indican los días de vida desde el diagnóstico hasta la muerte), surge una conclusión vital para tu guion:

   - No se observa la caída drástica de supervivencia que esperaríamos en otros cánceres al pasar al Estadio IV. De hecho, la mediana de supervivencia en el Estadio IV es similar a la del Estadio I.
   - **Justificación de la Firma Genética**: Esto confirma que la clínica "llega tarde". Como los pacientes viven mucho tiempo incluso en estadios avanzados, no podemos usar solo la supervivencia para medir el riesgo; necesitamos la firma genética para detectar la progresión de la enfermedad y la agresividad tras el diagnóstico

   El p valor nos indica de todas formas que esta relación no se puede extrapolar debido a su falta de significancia.

4. Boxplots de edad vs perfil ganglionar: Los datos de estos gráficos tampoco presentan datos significativos, por lo que no parece que haya una relación entre la cantidad de gánglios y la edad. Por tanto, a pesar de que la clasificación enfoque claramente que hay personas mayores con tendencia a estadíos más tardíos, no significa que estos presenten más ganglios afectados que los jóvenes. Nos gustaría saber a qué se debe esta clasificacion tan dependiente al rango de edad.

ANÁLISIS MULTIVARIANTE:

1. Primero, para comprobar la relación entre la edad con la categorización del estadio en función de la diseminación hemos realizado un bubble plot, donde el tamaño de los puntos representa la cantidad de ganglios afectados. Podemos observar como a edades tempranas, pacientes con gran cantidad de ganglios afectados siguen siendo categorizados como de estadío I. Solo hay un caso para el que esto no se cumple. Sin embargo, a edades más avanzadas, se observa como la clasificación sigue la estructura natural de todos los cánceres (los que presentan menor diseminación en los estadíos primarios y conforme progresa la enfermedad en tardíos).

2. Luego hemos realizado un Volcano plot comparando la expresión diferencial de genes en estadios tardíos vs tempranos. El gráfico tiene líneas punteadas que definen qué genes son "importantes":La línea horizontal: Es el umbral significativo (p-valor de 0.001). Todo lo que esté por encima es estadísticamente significativo. Las líneas verticales: Marcan el umbral de magnitud del cambio (1.2).

   Mientras los puntos Grises son genes que no cambiaron lo suficiente su perfil de expresión o cuyos cambios no son estadísticamente confiables, los Puntos Rojos son los genes de interés. Han pasado ambos filtros: cambian mucho y son significativos. Lado derecho (ej. COL11A1, MMP13, CCL17): Estos genes están sobreexpresión en los estadios avanzados. Podrían ser biomarcadores de progresión o responsables de que la enfermedad empeore. Estos genes se pierden o se "apagan" a medida que la enfermedad avanza- Hemos obtenido 39 genes.

3. Ahora, vamos a intentar ver si podemos obtener un claro mapa diferenciativo donde observemos claramente la diferencia de expresión comparando perfiles.

   1. Primero lo comparamos en función del estadío (tanto agrupado, en la primera diapositiva, como sin agrupar). Podemos observar que agrupados encontramos una clara diferencia de patrones. De los 39 genes, algunos en mayor medida que otros, están: 37 sobreexpresados en estadíos tardíos y 2 subexpresados (los dos últimos, se puede notar el cambio de color). Aún así, podemos ver como ciertas muestras no presentan del todo el perfil apagado esperado en los estadíos tempranos, lo que puede deberse justamente a la clasificación en estadíos dependiente de la edad. De hecho, si observamos la firma genética sin agrupar, nos damos cuenta de que el estadío II tiene claramente un perfil más apagado, y esto ya lo podíamos intuir antes con el bubble plot, ya que en este estadío solo encontrábamos pacientes con menor número de ganglios (menor progresión).

   2. Por otro lado, si estudiamos la expresión génica en función de el tipo histológico, también encontramos una clara diferencia de expresión en el tipo folicular, lo que tampoco nos sorprende, porque recapitulando, hemos visto que esta variante presentaba muchísima menor afectación ganglionar.

   3. Todo esto nos lleva a la misma consideración: El análisis de expresión génica más diferenciativo y donde se observará una mejor clasificación será aquel que se estudie en función de la diseminación ganglionar. Se observa en el heatmap una clara diferenciación. Esto pone más en duda esa dependencia de la clasificación por rango de edad, pues hemos visto claramente que el perfil genético asocia sus cambios de expresión al perfil ganglionar.

   4. Vamos, por último, a ver si claramente podemos observar una diferencia de expresión en función del perfil ganglionar, pero esta vez separando entre

pacientes mayores y menores de 55 años (el corte de edad al que nos referíamos cuando explicamos todo aquello de la diferente clasificación en estadíos). Se puede observar:

En el grupo de menores de 55 años:

- **Señal "Diluida"**: Aunque existe un bloque de expresión roja (alta agresividad) en los pacientes con metástasis ganglionar (**n1**, barra azul), este patrón es más **heterogéneo**. Hay muchos pacientes n1 que todavía muestran perfiles "fríos" (azules).
- **Infiltrados moleculares**: Se aprecian varios pacientes con estadio n0 (barra roja) que tienen una firma genética casi idéntica a los n1. Esto sugiere que en jóvenes, la maquinaria genética de invasión puede encenderse mucho antes de que se detecte el ganglio físicamente.

En el grupo de mayores de 55 años:

- **Bloques más compactos**: El bloque de pacientes n1 presenta una señal roja mucho más **densa y uniforme**.
- **Separación más clara**: La división entre el perfil "tranquilo" (n0) y el "invasivo" (n1) parece estar más definida visualmente, lo que correlaciona con el hecho de que en mayores la clínica es un predictor más fiable del riesgo que en jóvenes

Por último, vamos a realizar de nuevo el volcano plot, esta vez en función de la presencia o no de ganglios, a ver si esta vez cambia el asunto (ya que los genes que estamos estudiando los hemos sacado comparando el estadiaje e incluso así hemos podido observar en el heatmap que es mejor la separación en función del perfil ganglionar).

Hemos podido observar una mayor cantidad de genes diferencialmente expresados, por lo que hemos aumentado el umbral de magnitud de cambio a 2. De esta forma, el heatmap que obtendremos será incluso más robusto. En este caso obtenemos 27 genes.

Esta vez solo vamos a plantear el heatmap en función del perfil (separando en edades de nuevo. Observamos que, a pesar de que los cambios sean más observables en personas mayores de 55, encontramos que para esta clasificación si hay una diferencia de expresión con más lógica en jóvenes. Por ello, concluimos que estos genes tienen mayor confianza. Sigue habiendo sobre todo ciertos perfiles de jóvenes que tienen un mayor parecido al cáncer diseminado, sería interesante estudiarlos por separado. Además, a pesar de que en la distribución de edad con respecto al perfil ganglionar hayamos visto que hay un mayor porcentaje de jóvenes que presenten un mayor número de ganglios afectados, parece verse que no hay muestras que tengan una diferencia de expresión más alta que otras. El simple hecho de que el cáncer se disemine ya activa/desactiva dichos genes.

CONCLUSIÓN

Por último, cabe destacar que los genes no actúan de forma aislada, sino que forman parte de un complejo **gene networking**: una red de señales interconectadas donde el cambio en un solo nodo puede desestabilizar todo el sistema celular. La activación coordinada de nuestra firma de 27 genes podría ser la que realmente ejecuta el programa de invasión tumoral, actuando como los ejecutores de un plan maestro liderado por "genes directores" que originan este caos. Identificar a estos directores de orquesta nos permitiría hallar dianas terapéuticas estratégicas para desmantelar la progresión tumoral desde su origen.

Sin embargo, es crucial reconocer que este análisis se basa exclusivamente en datos de **RNA-seq (expresión génica)**, lo cual nos ofrece una visión valiosa pero parcial. Para comprender esta orquesta molecular en su totalidad, el siguiente paso lógico es la integración de **diferentes capas ómicas** (genómica, epigenómica, proteómica y metabolómica). Solo mediante un enfoque multiómico podemos observar qué es lo diferencial en cada paciente: mientras que algunos pueden presentar alteraciones a nivel de transcripción, otros podrían mostrar el punto crítico en la estabilidad de sus proteínas o en modificaciones epigenéticas (si, el mundo celular es muy complejo y lleno de capas difíciles de explicar en un sólo proyecto).

Para procesar esta magnitud de información, necesitaríamos de una herramienta con mayor potencia computacional y bioinformática que nos permitiera obtener un verdadero **mapa de correlación e integración**. Este mapa no solo nos diría qué genes están "encendidos", sino cómo interactúan las distintas capas biológicas en tiempo real, permitiéndonos identificar los "puntos críticos" de la red que, al ser bloqueados, detendrían la progresión del cáncer de forma fulminante.

La validación de los principales biomarcadores en estos perfiles moleculares integrados abre una puerta fascinante hacia el **drug repurposing** (o reposicionamiento de fármacos). Al identificar proteínas específicas que se sobreexpresan en variantes agresivas como la *Tall Cell*, podemos buscar medicamentos ya aprobados para otras patologías que ataquen esas mismas dianas. Esto no solo aceleraría la llegada de tratamientos al paciente (puesto que buscar tratamientos nuevos conlleva AÑOS), sino que permitiría ofrecer una solución terapéutica personalizada incluso cuando la clínica tradicional aún no tiene una respuesta clara.

En definitiva, nuestro análisis demuestra que el futuro de la oncología no reside en tratar estadios clínicos genéricos, sino en **hackear redes biológicas personalizadas** mediante la integración de toda la información ómica disponible, y todo esto cada vez es más posible y personalizable gracias a los modelos de aprendizaje profundo e IA, que estamos deseando abarcar en las siguientes semanas.

Por último, os queríamos dejar una imagen generada con IA de lo que se podría verdaderamente hacer con un dataset de esta calaña. Nosotras hemos basicamente llegado hasta el punto dos, pero, tal y como hemos explicado, hay muchos más pasos que seguir para llegar a una conclusión