Universidad Católica del Ecuador

Centro de educación virtual

Maestría en Biología Computacional

USO DE LAS HERRAMIENTAS SEURAT Y CLUSTERPROFILER PARA LA IDENTIFICACIÓN Y ANÁLISIS FUNCIONAL DE TIPOS CELULARES PRESENTES EN MUESTRAS DE TEJIDO MAMARIO A PARTIR DE DATOS SINGLE-CELL.

PROYECTO DE TITULACIÓN COMO REQUISITO PREVIO PARA LA OBTENCION DEL TITULO DE MAGISTER EN BIOLOGIA COMPUTACIONAL

Director: Andrés Romero Carvajal

Centro de educación virtual, Pontificia Universidad Católica del Ecuador

Av. 12 de Octubre 1076 y Roca. Quito, Ecuador

Rebeca Contreras Marcillo

Quito, Ecuador.

Agosto, 2023

Índice

Índice de figuras 4
Resumen 5
Summary 5
Introducción 6
Heterogeneidad celular y la tecnología single-cell 6
Análisis de datos single-cell con la herramienta Seurat y clusterProfiler 8
Control de calidad 8
Obtención de grupos celulares 10
Identificación y caracterización de los grupos celulares encontrados. 13
Análisis de células del tejido mamario. 14
Componentes celulares del tejido mamario 14
Enfoque del trabajo 16
Hipótesis 17
Objetivos 17
Objetivo general 17
Objetivos específicos. 17
Materiales y métodos 18
Obtención de los datos 18
Obtención de grupos celulares 18
Control de calidad y normalización 18
Pre-procesamiento de los datos. 18
Reducción de dimensionalidad y Clustering 18
Integración de los datos 19
Identificación de células 19
Análisis funcional 19
Resultados 20
Control de calidad 20
Obtención de grupos celulares datos sin integrar 21
Obtención de grupos celulares en datos integrados 24
Identificación de los grupos celulares 26
Etiquetado 41
Discusión 42

Conclusión 45
Referencias 46
Anexo 51
    Características de los tipos celulares 51

Índice de figuras

Figura 1. Diferencia entre las tecnologías Bulk y Single-cell, ... 7
Figura 2. Aislamiento de células por tecnologías Chromium Single Cell. 9
Figura 3.Componentes estructurales de la mama. 15
Figura 4.Tipos de células de la glándula mamaria. 16
Figura 5.Control de calidad (QC) de la matriz de cuentas de Pal et al (2021) 20
Figura 6.Grupos celulares obtenidos con las funciones de Seurat 22
Figura 7.Grupos celulares obtenidos de los datos integrados con las funciones SelectIntegrationFeatures(), FindIntegrationAnchors() y IntegrateData(). 25
Figura 8.Identificación de los grupos celulares (3,7,2,17,1 y 9). 27
Figura 9.Exploración de la expresión del marcador EPCAM de células epiteliales. 28
Figura 10. Identificación y caracterización de poblaciones de células luminales progenitoras. 29
Figura 11. Identificación y caracterización de poblaciones específicas de células luminales maduras. . 30
Figura 12.Identificación y caracterización de poblaciones de células basales. 31
Figura 13.Identificación de los grupos celulares 4,6,11,8,10,13,0,5 y 16. 33
Figura 14. Identificación y caracterización de poblaciones de fibroblastos 34
Figura 15.Identificación y caracterización de poblaciones de células endoteliales. 35
Figura 16. Identificación y caracterización de poblaciones de pericitos. 36
Figura 17.Identificación de los grupos celulares 12,14 y 15. 37
Figura 18. Identificación y caracterización de poblaciones de macrófagos. 38
Figura 19. Identificación y caracterización de poblaciones de linfocitos T. 39
Figura 20. Identificación y caracterización de poblaciones de células plasmáticas. 40
Figura 21.Etiquetado de los grupos celulares obtenidos. 41

Resumen

La heterogeneidad es una propiedad fundamental de los sistemas biológicos que les da ventajas adaptativas y funcionales. Hasta hace poco, las células de los tejidos habían sido estudiadas como poblaciones mediante las tecnologías de secuenciación obteniendo un promedio de los perfiles genómicos y transcriptómicos ignorando el aporte que tiene cada tipo de célula diferente en el comportamiento total de la población, tejido u organismo. Las tecnologías *single-cell* permiten obtener los perfiles moleculares de cada célula por separado capturando las diferencias entre las distintas células que componen el tejido. El primer paso para aprovechar estas tecnologías es la identificación y caracterización de tipos celulares para después poder compararlas entre diferentes condiciones biológicas y entender el efecto de estos tratamientos sobre cada tipo celular. Sin embargo, el proceso requiere mayor conocimiento de herramientas bioinformáticas lo que puede ser una limitante para los investigadores, por esto se han desarrollado una serie de marcos de trabajo como Seurat que son paquetes que contienen una serie de herramientas para realizar los pasos básicos de un análisis completo de *single-cell:*control de calidad, reducción de complejidad de los datos, agrupamiento de células, identificación de marcadores, por otro lado, clusterProfiler es un herramienta que permite extraer información de los marcadores obtenidos, pudiendo identificar los grupos formados y hacer un análisis funcional mediante diferentes pruebas de enriquecimiento y sobrerrepresentación. En este trabajo se utilizarán estas herramientas para identificar los tipos celulares presentes en el tejido mamario normal obtenidos de la base de datos GEO (GSE161529) pertenecientes al trabajo de Pal *et al* (2021).

Summary

Heterogeneity is a fundamental property of biological systems that gives them adaptive and functional advantages. Until recently, tissue cells have been studied as populations by sequencing technologies, in which an average of the genomic and transcriptomic profiles are obtained, ignoring the contribution that each different cell type has in the total behavior of the population, tissue or organism. Single-cell technologies make it possible to obtain the molecular profiles of each cell separately, capturing the differences between the different cells that make up the tissue. The first step to take advantage of these technologies is the identification and characterization of cell types, which will later allow the comparison between different biological conditions and understand the effect over every cell type. However, the identification process requires more knowledge of bioinformatics tools, which can be limiting for researchers. For this reason, a series of frameworks have been developed, such as Seurat, which are packages that contain a series of tools to carry out the basic steps for a single-cell analysis: quality control, data complexity reduction, cell clustering , and to fin markers. On the other hand, clusterProfiler is a tool that allows extracting information from the markers obtained, being able to identify the groups formed and perform a functional analysis through different enrichment and overrepresentation tests. In this work, these tools will be used to identify the cell types present in normal breast tissue obtained from the GEO database (GSE161529) from the work of Pal et al (2021).

Introducción

Heterogeneidad celular y la tecnología single-cell

La heterogeneidad celular es una propiedad fundamental de los sistemas biológicos que les da ventajas adaptativas y funcionales. En organismos multicelulares se requieren diferentes tipos de células especializadas para el correcto funcionamiento de sus tejidos y sistemas. El estudio de células a nivel individual ha permitido entender fenómenos como el desarrollo embrionario en el que se toman decisiones a nivel de una célula y cada una es esencialmente distinta (Griffiths, Scialdone & Marioni, 2018). En inmunología, permite develar la particularidad de los diferentes receptores expresados por las células. En el contexto del cáncer, la heterogeneidad representa una ventaja adaptativa que le permite adquirir resistencia a los fármacos (Venkatesan& Swanton.,2016; McErlean & Brauer-Krisch., 2016; Piraino, Thomas, Donovan & Furney.,2019)

Las tecnologías de secuenciación son herramientas importantes para el estudio de las células ya que permiten obtener el conjunto de moléculas de ADN y ARN presentes en un momento determinado. Hasta hace poco, estas tecnologías se aplicaban sobre poblaciones de células extraídas de una muestra (la que puede contener células de diferentes tipos) obteniendo un perfil promedio de la expresión de todas (Griffiths, Scialdone & Marioni, 2018), sin que este refleje la verdadera expresión de ninguno de los tipos presentes en la muestra (figura1). Esto puede enmascarar perfiles menos abundantes pero relevantes (como perfiles asociados con resistencia en células tumorales) impidiendo identificar la verdadera fuente de una patología, otras veces, puede enmascarar mecanismos de interacción entre las diferentes células que es importante en procesos como el desarrollo o el cáncer en el que cada tipo de célula sufre un cambio transcripcional diferente según su propia función. En contraste, las técnicas de secuenciación de una sola célula (single-cell por su nombre en inglés), permiten obtener los perfiles moleculares de cada célula por separado (figura 1) lo que a su vez permite encontrar diferencias moleculares vinculadas a tipos de células específicos (Griffiths, Scialdone & Marioni, 2018). Usando estas tecnologías, se han creado una serie de atlas celulares que buscan construir mapas de referencia de todos los tipos y estados celulares en condiciones fisiológicas y patológicas las cuales están siendo reunidas en iniciativas como el atlas de células humanas, The Human Cell Atlas (HBCA) (Rozenblatt-Rosen,Stubbington, Regev & Teichmann,. 2017) o el atlas de células mamarias (HBCA) (Kumar *et al*., 2023).

Por su parte, la tecnología de secuenciación de RNA de una célula (scRNA-seq) presenta muchas ventajas para el estudio de heterogeneidad celular ya que que el transcriptoma representa, en gran parte, el subconjunto del genoma que se encuentra activo que se está expresando y por lo tanto determina sus características y procesos particulares (a diferencia del genoma, que es el mismo en todas las células) y técnicamente es más eficiente y tiene menos dificultades que las tecnologías de secuenciación de proteínas (Adams., 2008, Jehan., 2019).

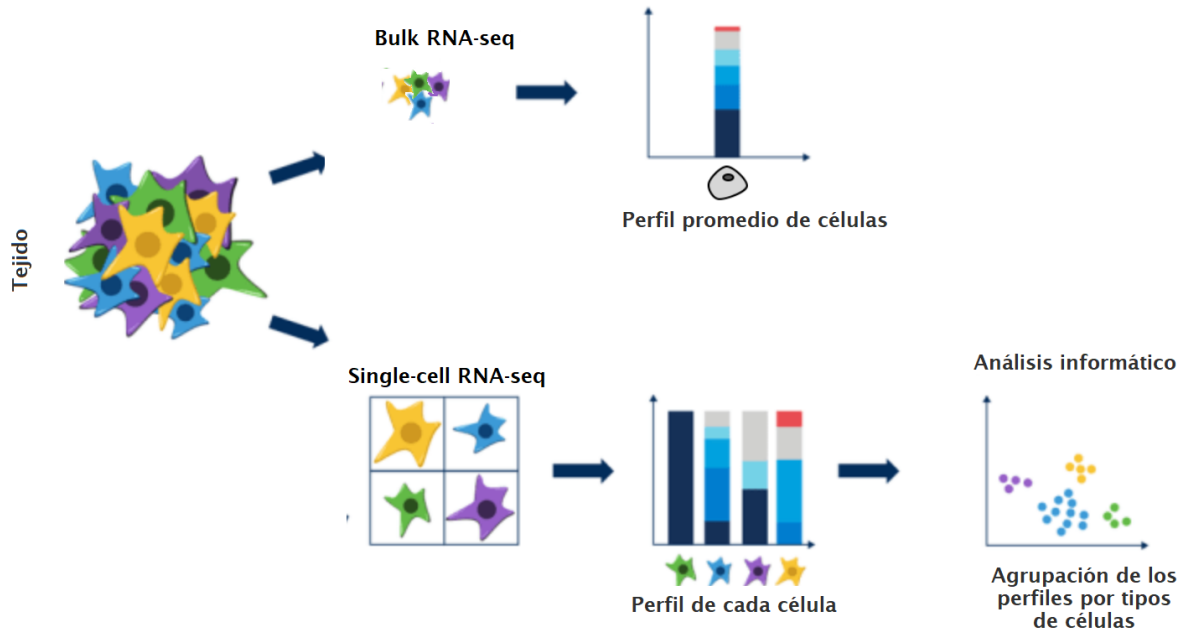


Figura 1. Diferencia entre las tecnologías Bulk y Single-cell,
En las técnicas bulk se obtiene un perfil promedio de las diferentes células que representan a una célula que no existe. Las técnicas single-cell permiten identificar perfiles verdaderos, sin embargo, implican un paso extra de análisis. (imagen adaptada de Nieto., 2022).

Sin embargo, para aprovechar esta tecnología se requiere un análisis informático más extenso y un mayor conocimiento de herramientas bioinformáticas. Dado que del secuenciador solo se obtiene una lista de los genes expresado en cada célula sin tener una clasificación por tipo o estado biológico (figura 1), el primer paso es la identificación y caracterización de tipos celulares en la muestra, este paso es esencial para después compararlas entre diferentes condiciones sanos/patológicos, tratamiento/sin tratamiento, distintas especies o distintas etapas de desarrollo y entender su papel en estos procesos. Para esto, después de obtener los perfiles transcriptómicos y alinearlos, a diferencia de las técnicas *bulk,* hay que identificar patrones similares entre las diferentes células que permitan su agrupación en tipos. Esto implica un análisis e interpretación de los datos en diferentes pasos lo que puede representar una limitante para los investigadores. Se han desarrollado una serie de entornos integrados o "frameworks" como Seurat, Scanpy y Scater que son paquetes que contienen una serie de herramientas para realizar los pasos básicos de un análisis completo de *single-cell*: exploración de datos, control de calidad, obtención de tipos celulares, identificación de genes diferencialmente expresados y visualización de los datos. Uno de los más populares, Seurat (Hao, Hao, et al.,2021), es un paquete creado para el lenguaje R que tiene un buen rendimiento junto con una interface amigable con para el usuario y cuenta con una amplia comunidad activa por lo que hay disponibles numerosos tutoriales y recursos que facilitan el aprendizaje (Mangiola, Doyle & Papenfuss., 2021). Por esto, representa un buen acercamiento al análisis de los datos *single-cell* Para este trabajo se eligió aplicar estas herramientas en la identificación de tipos celulares en muestras de tejido mamario normal provenientes del trabajo de de Pal *et al* (2021).

A grandes rasgos, el proceso de identificación tiene los siguientes pasos básicos:

- Control de calidad: Se intenta reducir el efecto de fallas técnicas y sesgos en los datos mediante el filtrado y ajuste de los datos.
- Reducción de complejidad de los datos: Se filtran los genes que no varíen entre diferentes grupos y reducción de dimensionalidad que faciliten la evaluación de similitudes entre perfiles transcriptómicos al agrupar los genes en variables que idealmente den cuenta de firmas relacionadas, permitiendo evaluar las características de las células en términos de procesos, vías o tipos en lugar de evaluarlas en términos de genes individuales al mismo tiempo que selecciona las variables que capturen la mayor variabilidad entre células.
- Agrupamiento de células: En el que se evalúa la similitud transcripcional (usando las variables generadas en el paso anterior) entre las células para determinar cuáles son más parecidas entre sí por lo que pertenecen a un mismo grupo.
- Identificación de los grupos: determinar la identidad de los grupos de células formados, relacionando los marcadores, firmas y características de cada grupo con una referencia.

Análisis de datos single-cell con la herramienta Seurat y clusterProfiler

Control de calidad

Contexto experimental

Para aislar las células se usan diferentes métodos siendo las tecnologías microfluídicas Chromium las más populares. Estas tecnologías se basan en un sistema formado por dos fluidos inmiscibles que forman dos fases: una continua y una dispersa. Cuando la muestra de células se incorpora al sistema, este "dispersa" uno de los fluidos en forma de pequeñas gotas dentro del otro. El equipo está diseñado para entregar en las gotas, las células individuales (en la mayoría de los casos), los reactivos de transcripción inversa (RT), oligonucleótidos con índices a través de perlas de gel. Estos van a formar vesículas de reacción llamadas perlas de gel en emulsión o GEM. Cada GEM se convierte en una unidad de reacción donde la construcción de la biblioteca (extracción de ARN, transcripción inversa a ADNc, agregado de índices y amplificación) ocurre por separado para cada célula. El ADNc generado en cada uno de estos GEM (pertenecientes idealmente a una sola célula) se podrá identificar durante el análisis gracias los índices añadidos que son específicos para cada muestra, célula y read (índice denominado UMI) antes de la amplificación. Los UMI permiten distinguir entre copias amplificadas de la misma molécula de ARNm. Después de la construcción de la biblioteca, los GEM se desintegran y las bibliotecas se agrupan (multiplexan) dentro de la fase continua en donde va a ocurrir la secuenciación (Luecken & Theis, 2019)

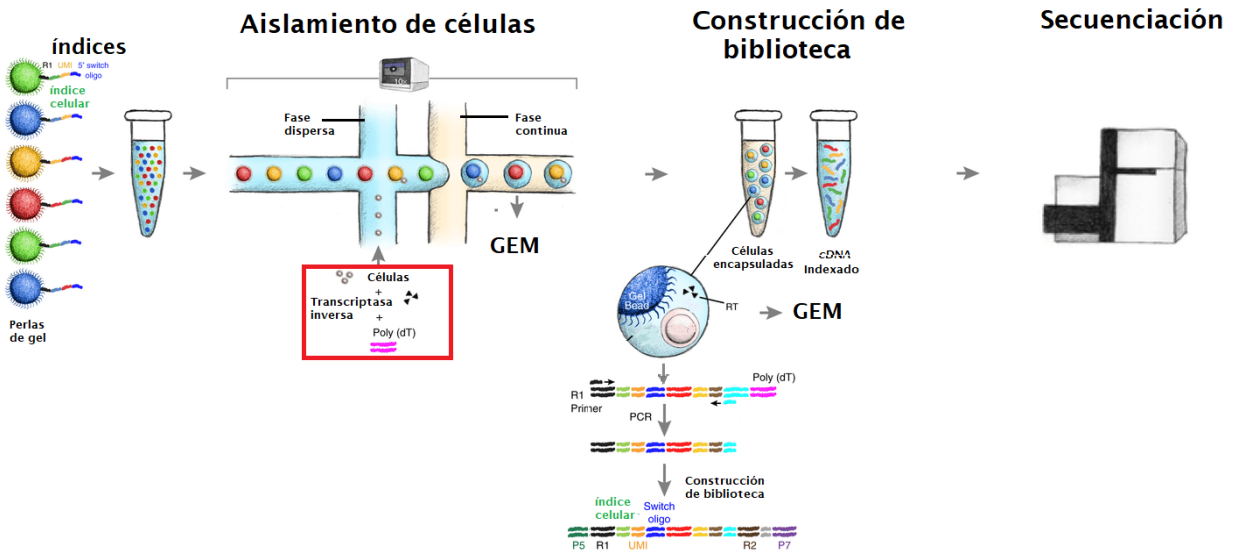


Figura 2. Aislamiento de células por tecnologías Chromium Single Cell.
Proceso de aislamiento de las células consiste en la formación de GEM que es la unidad de construcción de biblioteca para cada célula y la secuenciación que ocurre en un pool conjunto (imagen adaptada de Goldstein,Chen, Wu, Chaudhuri, Hsiao, Schneider & Seshagiri, 2019..

Obtención de matriz de cuentas

Los datos generados en la secuenciación se procesan para obtener matrices de cuentas moleculares (de células) y recuentos de reads. Por medio de distintos paquetes como Cell Ranger o indropsse se pueden realizar todos los pasos necesarios para obtener la matriz de cuentas el control de calidad de lectura (QC), "desmultiplexación" que es la identificación de índice de GEM y UMIS para cada read, alineamiento y cuantificación obteniendo finalmente una matriz de cuentas en la que se ha cuantificado la cantidad de reads alineados a cada gen (que idealmente indican la cantidad de transcritos generados por ese gen) y clasificado según la GEM y el transcrito de origen (Luecken & Theis, 2019, Slovin *et al*.,2021).

Hay que considerar que después de este paso se obtiene la matriz de cuentas etiquetadas solo según su unidad de secuenciación (la GEM de la cual provienen) y UMI, esto quiere decir que no se sabe si provienen de una o más células (dependiendo si se capturaron más en la unidad), ni a qué tipo de célula pertenece. Entonces para poder identificar tipos celulares hay que agrupar las células según sus perfiles transcriptomicos. Esto se pueden lograr usando distintas funcionas de la herramienta Seurat.

Control de calidad de células

Las matrices de recuento contienen las cuentas capturadas en cada GEM la cual se espera contenga un sola célula, sin embargo, en realidad esta puede haber capturado células no viables, más de una célula (doblete) o ninguna. Para evaluar cada una de estas situaciones se miden las variables: cantidad de transcritos por GEM (profundidad de conteo), la cantidad de genes por GEM, fracción de cuentas por GEM, cuentas de genes mitocondriales. Por ejemplo, GEMS con una profundidad reads baja, pocos genes detectados y una fracción alta de conteos mitocondriales indican células no viables, que han perdido su ARNm citoplasmático por la pérdida de integridad de la membrana y solo conservan el ARNm ubicado en las mitocondrias. Por el contrario, las células con un número de cuentas muy alto y una gran cantidad de

genes detectados pueden representar dobletes. Por otro lado, la matriz puede contener información que generan ruído como cuentas de genes que se expresan en pocas células y, por lo tanto, no son informativas (Luecken & Theis, 2019, Linderman.,2021,McCarthy & Hemberg.,2021). Luego de la evaluación, se filtran las GEMS con la función subset(),en el que se especifican los números el mínimo y máximo de genes con el parámetro Feature_RNA y el porcentaje máximo de genes mitocondriales debe tener para la retención con el parámetro percent.mt < 5 que tengan estas características, obteniendo una matriz de cuentas que idealmente contenga solo las GEMS que representen una sola célula viable.

Normalización

La matriz de cuentas contiene una lista de genes asociado a un número de cuentas que debería reflejar el nivel transcrito por cada gen para luego poder ser comparados. Sin embargo, genes con el mismo nivel de expresión entre dos condiciones pueden aparecer como diferencialmente expresadas debido a factores técnicos como haber pasado por diferente número de rondas de amplificación y tener diferentes tamaños de transcritos. Esto impide que los niveles no sean comparables directamente. Para abordar este problema se realiza una normalización en la que se transforma el valor absoluto de las cuentas a un valor relativo a las variables mencionadas anteriormente. En Seurat la normalización se implementa con la función NormalizeData(), que divide el número de cuentas de cada célula en las cuentas totales generados en ella y lo multiplica por un factor de escala, que generalmente es 1000. Luego, se aplica una transformación logarítmica log(x+1) lo que permite que las distancias entre los valores de expresión representen veces de cambio logarítmico, que es la medida estándar para los cambios en la expresión y además, reduce la asimetría de los datos y logrando que los datos se acerquen a esta distribución normal, lo que es importante porque es un requisito para los siguientes pasos (Luecken & Theis, 2019, McCarthy & Hemberg.,2021, Slovin *et al*.,2021).

Obtención de grupos celulares

Encontrar tipos celulares significa encontrar perfiles transcriptómicos comunes entre algunas células pero diferente con las otras, para esto se calculan los perfiles que generen mayor variabilidad.

Identificación de genes altamente variables

Los datos que quedan del control de calidad pueden contener una cantidad muy grande de genes que pueden dificultar el análisis y solo algunos de estos genes contienen información relevante para el estudio, esto es, solo algunos varían en su expresión entre condiciones o células debido a factores biológicos, mientras que el resto se mantiene en niveles similares o presenta una variación mínima atribuible a ruido técnico. El identificar estos genes permite reducir el tamaño de los datos, filtrando el ruido y manteniendo solo los genes con la señal biológica más fuerte. Para esto se calcula la varianza media de la expresión y los genes que presenten una relación media-varianza más alta se mantienen. En Seurat, este paso lo cumple la función FindVariableFeatures() que utiliza el método VST transformación estabilizadora de varianza, que ajusta el cálculo considerando que no todos los genes se mueven en los mismos rangos (Luecken & Theis, 2019, McCarthy & Hemberg.,2021, Slovin *et al*.,2021).

Estandarización de los datos

Luego, en los siguientes pasos se va a proceder a la identificación de grupos, para lo cual se quiere identificar características diferenciantes evaluándolas por la variabilidad que presentan en los datos, sin embargo, para esto, primero deben tener una variabilidad comparable. El problema es que algunos genes pueden moverse en rangos de expresión muy grandes por lo que van a diferenciarse en varios órdenes de magnitud entre células, mientras otros pueden presentar variaciones menores pero significativas. Por esto

se debe ajustar la variabilidad de manera que los genes que se mueven en mayores rangos no aparezcan como más significativos que el resto de genes. Para esto se estandarizan los datos de modo que la expresión media para los genes entre células sea 0 y la varianza, 1 mediante la función ScaleData() de Seurat.

Reducción de dimensionalidad

Como se mencionó anteriormente se quiere encontrar la forma separar los datos según sus perfiles de expresión, sin embargo, este tipo de datos es multidimensional (tiene muchas características, miles de genes), lo que hace difícil la comparación entre células/muestras en base a cada gen y puede llevar a sesgos, por lo que se realiza un análisis de reducción de dimensiones que permite combinar la información de los miles de genes en menos variables. En Seurat se utiliza el método de análisis de componentes principales (PCA) con la función RunPCA(). Esta técnica permite por un lado resumir información y por otro encontrar la información que capture la mayor variabilidad entre grupos de células. Para resumir la información, se calcula la covarianza entre genes, y se van combinando en una sola variable en forma de una función formada por grupos de genes cuyos valores se ajustan para que represente la mayor variabilidad entre células (en un proceso equivalente al de encontrar la mejor recta que represente la relación entre variables en los análisis de regresión lineal). Las variables generadas se denominan componentes principales y al ser construidas en base a la covarianza permiten capturar una especie de firma, con grupos de genes que tiendan a co-expresarse (activarse o desactivarse juntos), y que podría indicar firmas asociadas un mismo proceso o un mismo tipo de célula. Luego estos se van ordenando según los que tengan menor correlación entre ellos de manera de obtener las firmas más distintivas que presenten la mayor variabilidad entre grupos de células (Luecken & Theis, 2019, Andrews, Kiselev, McCarthy & Hemberg.,2021). Finalmente se mantienen los componentes que permitan dividir los datos en grupos. Esto se puede visualizar en un sistema de coordenadas en la que cada componente es un eje en el que las células se van a posicionar, entonces al graficar estos componentes, puede que un componente representando el eje “y” esté reflejando firmas relacionadas con células epiteliales y otro componente en el eje “x” esté relacionado con el estado de proliferación, luego las células se van a posicionar a lo largo del eje “y” según su similitud con células epiteliales y a lo largo del eje “x” si están proliferando o no. Sin embargo, hay que considerar que las firmas obtenidas pueden captar ruido generado por ejemplo por efectos de lote en lugar de características biológicas relevantes, por lo que las células se van agrupar en función la técnica de extracción o tecnología de secuenciación aplicada.

Visualización reducción dimensional no lineal (UMAP)

Después del PCA los datos se reducen a unos 20-50 variables (componentes) o más. Para visualizarlos todos se requerirían 20-50 ejes. Esto se puede visualizar en 2 ejes aplicando otro algoritmo de reducción de dimensiones de manera similar al PCA pero que utiliza técnicas de reducción de dimensión no lineales como UMAP que permite trabajar con datos tipo clúster (agrupados) y es uno de los que utiliza Seurat mediante la función RunUMAP() (Luecken & Theis, 2019, McCarthy & Hemberg.,2021).

Clustering de células

En este paso llamado *clustering* se asignan las células a un grupo determinado. Para esto, hay que establecer el nivel de similitud que debe haber entre las células para constituir un grupo, basándose en el espacio generado por las firmas del PCA. Se puede llegar a diferentes grados de resolución identificando células que tengan el mismo linaje, pero sean diferente tipo o que estén en diferentes estados de activación, proliferación u otras características. En Seurat se utilizan 2 algoritmos, el primero evalúa la similitud entre células, registra el nivel de relación entre ellas y mapea sus conexiones, pero sin asignarlas

a un grupo en particular. Para la asignación se aplica un segundo algoritmo que detecta los grupos de células que presenten una densidad mayor de conexiones entre sí y que por lo tanto constituyan un mismo grupo (Luecken & Theis, 2019, Andrews, Kiselev, McCarthy & Hemberg.,2021, Slovin *et al*.,2021).

Para el primer paso en Seurat mediante se utiliza la función FindNeighbors() que usa el algoritmo k vecinos más cercanos o KNN (por k-nearest neighbors en inglés) para construir un grafo de SNN (shared nearest neighbors) que refleje el nivel de relación que hay entre las células. El algoritmo KNN ocupa las variables generadas por el PCA, toma las diferentes células como nodos e identifica a un determinado número K de células vecinas que tengan un perfil transcripcional parecido (K es un parámetro establecido por el investigador) y las conecta. Luego, dado que las células pueden formar muchas combinaciones de grupos, se les asigna un peso a estas conexiones basándose en el índice de Jaccard que mide cuántos vecinos comparten dos células en comparación con el número total de vecinos únicos que tiene cada una. El resultado de estos cálculos un grafo que refleja las conexiones entre las células según su similitud de expresión y la cohesión que tenga con su grupo (Luecken & Theis, 2019, Andrews, Kiselev, McCarthy & Hemberg.,2021).

Para el segundo paso se asigna cada célula a un grupo, Seurat implementa el algoritmo de Louvain, detectando grupos de células que tienen una mayor densidad de conexiones entre ellas que con células de los otros grupos. La densidad necesaria para que un conjunto de células constituya un solo grupo, se define por el parámetro de resolución r. En la función FindClusters() para cada parámetro de resolución, el algoritmo realizará la agrupación de las células, una mayor resolución, se traduce en más grupos de células que podría determinar mayores características entre los grupos aunque también podría llevar a capturar ruido entre los grupo. Finalmente, cada grupo obtenido se puede interpretar como tipo de células o estado biológicos (Luecken & Theis, 2019, McCarthy & Hemberg.,2021). La elección de la resolución es un proceso iterativo en que el investigador debe ir evaluando y ajustando a medida que va a analizando los datos.

Integración de datos

Este es un paso que va antes de la normalización, pero se utiliza solo en ciertos casos ya que su función es identificar poblaciones celulares comunes entre muestras o conjuntos de datos, se puede utilizar con el fin de corregir efectos por lote o sesgos biológicos como en el caso de tomar muestras en diferentes ubicaciones de un tejido (Luecken *et al*., 2019, Luecken *et al*., 2022).

Para esto, se identificar un conjunto de genes comunes entre los conjuntos de datos y se realiza un análisis de correlación canónica (CCA) que de manera similar a la técnica de PCA, hace combinaciones lineales de genes en cada conjunto de datos (llamados componentes canónicos) y se buscan las células entre los conjuntos de datos con la correlación más alta entre sus componentes canónicos que den cuenta patrones de expresión génica similares. Estos son identificados como células comunes o anclajes. Luego con el algoritmo de vecinos más cercanos mutuos (MNN),que sigue la misma lógica de KNN, se buscan agrupar los mismos tipos de células, identificando pares de células dentro de k vecinos más cercanos de las células representativas, pero a diferencia de KNN en lugar de buscarlos entre los datos de su misma muestra, los hace entre las diferentes muestras. Con esto, se estarían encontrando las células que sean similares entre las muestras y que corresponderían a un mismo tipo (Ryu, Han, Jung & Hwang.,2023). Tras la integración se corren los procesos normales del flujo de trabajo encontrar genes variables, estandarización, PCA, agrupamiento de células e identificación.

Cabe notar, que al realizar este paso se puede perder heterogeneidad biológica, ya que puede haber muestras que contengan células únicas con genes que se expresen de manera diferencial y que sean de interés biológico (Haghverdi, Lun,Morgan & Marioni.,2017).

Identificación y caracterización de los grupos celulares encontrados.

Luego de agrupar las células según sus perfiles, se deben anotar (identificar) para lo cual hay 3 opciones el método manual, método por correlación y por clasificación supervisada.

El método por correlación utiliza conjuntos de datos scRNA-seq ya etiquetados como entrada para la identificación del tipo de célula y evaluando la similitud entre los conjuntos de datos de referencia y de consulta. Para esto calcula la correlación los perfiles de expresión de cada célula en el conjunto de consulta con los perfiles de las células de referencia. Para cada célula en el conjunto de datos scRNA-seq, le asigna la etiqueta que muestre la mayor correlación con su patrón de expresión génica. (Pasquini, Arias, Schäfer & Busskamp,2021). Los métodos por aprendizaje supervisado, entrena un clasificador con diferentes referencias etiquetadas tras lo cual se puede utilizar para consultar sus conjuntos de datos (Pasquini, Arias, Schäfer & Busskamp,2021).

En el método manual, se calcula genes expresados diferencialmente comparando la expresión de cada grupo encontrado con los demás, y los toma como marcadores, cabe notar que la elección de cuán diferencialmente expresado debe estar un gen para usarse como marcador, puede ser en parte arbitraria, y se establece después de una exploración de los dato, aunque se usan generalmente sobre 0.5 y 1 de veces de cambio como puntos de corte. Estos se evalúan por su coincidencia con listas de genes marcadores de tipos de células conocidos de las referencias. La significancia estadística de la coincidencia se puede determinar por medio de herramientas bioinformáticas que incluyen pruebas estadísticas como el análisis de sobrerrepresentación (ORA) y el análisis de enriquecimiento de conjuntos de genes (GSEA) (Pasquini, Arias, Schäfer & Busskamp,2021). Este tipo de anotaciones suficientemente efectiva cuando se trabajan con tejidos ampliamente estudiados y con gran disponibilidad de marcadores (como es el caso del tejido mamario) y la mecánica es fácil de entender para un primer acercamiento a las técnicas single-cell. Este fue uno de los métodos usados para la identificación de células en el trabajo de Kumar *et al*., 2023 que es parte del atlas de células mamarias HBA.

Para encontrar marcadores de cada grupo, Seurat presenta la función FindAllMarkers() que utiliza la prueba de suma de rangos de Wilcoxon, modelo lineal generalizado de poisson, la prueba t de Student entre otras pruebas estadísticas y las evalúa para cada célula componente del grupo.

Con el análisis de expresión diferencial idealmente se obtienen listas de genes que llevaron a la separación de los grupos y que son propias de cada uno. Luego, como se mencionó anteriormente, se puede comparar si esta lista coincide significativamente con una lista de marcadores de un tipo celular en particular (o de una vía, o función molecular en particular) mediante ORA o GSEA. ClusterProfiler es una herramienta que permite hacer estos tipos de análisis. En particular mediante la función enricher() se puede implementar un análisis de sobrerrepresentación para comparar los marcadores obtenidos con listas de marcadores de tipos celulares. Enricher() aplica la prueba exacta de Fisher que evalúa si hay una diferencia estadísticamente significativa entre la proporción de genes diferencialmente expresados que se encuentra en la lista de marcadores referenciales vs la proporción de genes no diferencialmente expresados que se

encuentran en ella (Wu *et al*., 2021). Por otro lado, las listas de genes marcadores de tipos celular se pueden obtener en bases de datos como CellMarker y PanglaoDB que son las más usadas para single-cell.

En adición, los marcadores se pueden utilizar para caracterizar los grupos celulares mediante análisis funcional, clusterProfiler tiene funciones especializadas para hacer estos análisis aplicando el mismo análisis de sobrerrepresentación sobre las bases de datos más conocidas como wikipathways, DAVID, Gene Ontology (GO) y la Enciclopedia de genes y genomas de Kyoto (KEGG). Estas últimas pueden utilizar para ORA con las funciones enrichgo y enrichkegg de clusterProfiler.

Análisis de células del tejido mamario.

Una de las aplicaciones más populares de la tecnología single-cell es el estudio del tejido mamario el cual ha adquirido una importancia especial debido a que el cáncer de mama es la principal causa de mortalidad por cáncer en las mujeres de todo el mundo (Easton & Wilcox., 2023). Dado que las células tumorales forman verdaderos ecosistemas compuestos por diferentes células malignas epiteliales, estromales e inmunitarias en la que cada una cumple una función diferente en la supervivencia y desarrollo de la enfermedad, se deben estudiar cada una por separado.

Por el potencial del estudio de las células a nivel individual para entender los mecanismos involucrados en patologías, se han creado una serie de estudios enfocados al análisis de heterogeneidad celular en tejido normal, incluyendo el tejido mamario en los trabajos de Kumar *et al*., 2023 y Bhat-Nakshatri *et al*., 2021. que buscan identificar y caracterizar las células presentes en la glándula mamaria y en su entorno de modo que sirvan de referencia de la biología del tejido en estado fisiológico para luego estudiar la biología de los estados patológicos. Por esto y por su importancia, se eligió analizar muestras tejido mamario normal en este trabajo.

Componentes celulares del tejido mamario

La mama es un órgano apocrino cuya función principal es producir y secretar leche después del embarazo. Está formada por la glándula mamaria (que es la parte funcional responsable de la producción de leche), tejido conectivo (que proporciona soporte estructural a los tejidos), adiposo (que le dan la forma e integran la glándula mamaria), por vasos sanguíneos (que suministran nutrientes y oxígeno) y conductos linfático (que drenan el exceso de líquido y transportan células inmunitarias) (Biswas, Banerjee,Baker, Kuo & Chowdhury.,2022).

La glándula mamaria, a su vez, está formada por unidades secretoras llamadas alvéolos y un sistema ductal. Los alvéolos se organizan en estructuras más grandes llamadas lóbulos que se conectan por el sistema ductal compuesto por los ductos galactóforos que colectan y transportan la leche desde los alvéolos hasta el pezón figura 3 (Biswas, Banerjee, Baker, Kuo & Chowdhury.,2022).

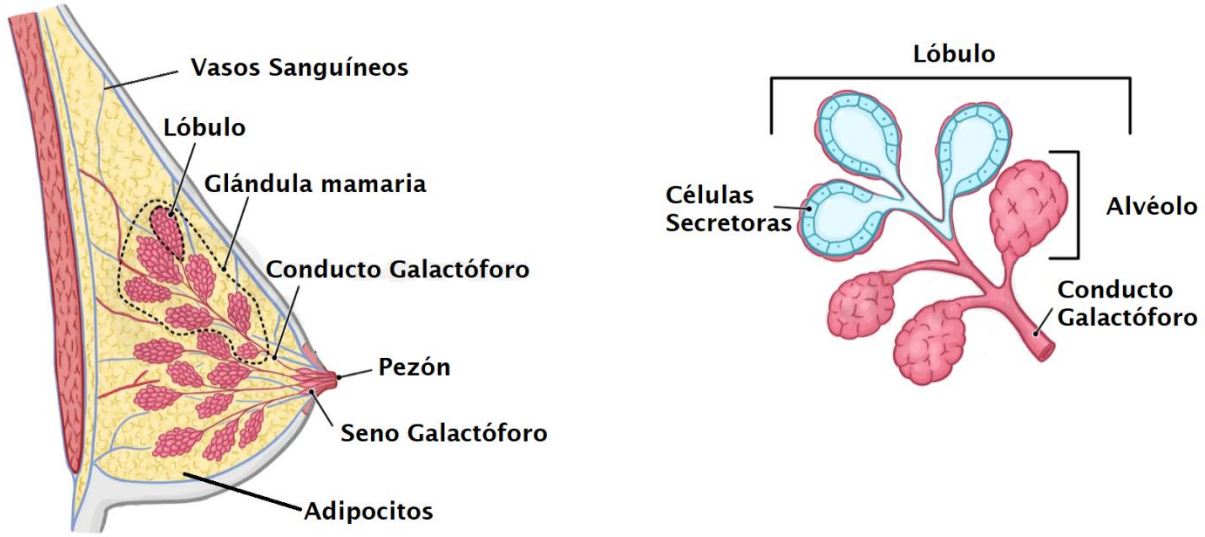


Figura 3.Componentes estructurales de la mama.
La parte funcional compuesta por una parte secretora y un sistema ductal es soportada y protegida por tejido conectivo, músculo y grasa, y es alimentado por vasos sanguíneos y linfáticos y el pezón es la conexión exterior (adaptado de https://www.osmosis.org/learn/Anatomy_of_the_breast).

Estas estructuras están formadas por diferentes tipos de células, células epiteliales que forman alvéolos y ductos y constituyen una barrera con funciones especializadas, células endoteliales y pericitos que forman la vasculatura sanguínea y linfática que forman una barrera especializada y regulan su hemostasis, fibroblastos que forman parte del tejido conectivo y sintetizan los componentes de la matriz extracelular y las células del sistema inmune que reconocen sustancias extrañas y coordinan la respuesta con ls demás células (Bernard *et al*., 2018). Cada tiene expresan genes particulares, presentando perfiles de expresiones que permiten identificarlos. Detalles en el anexo.

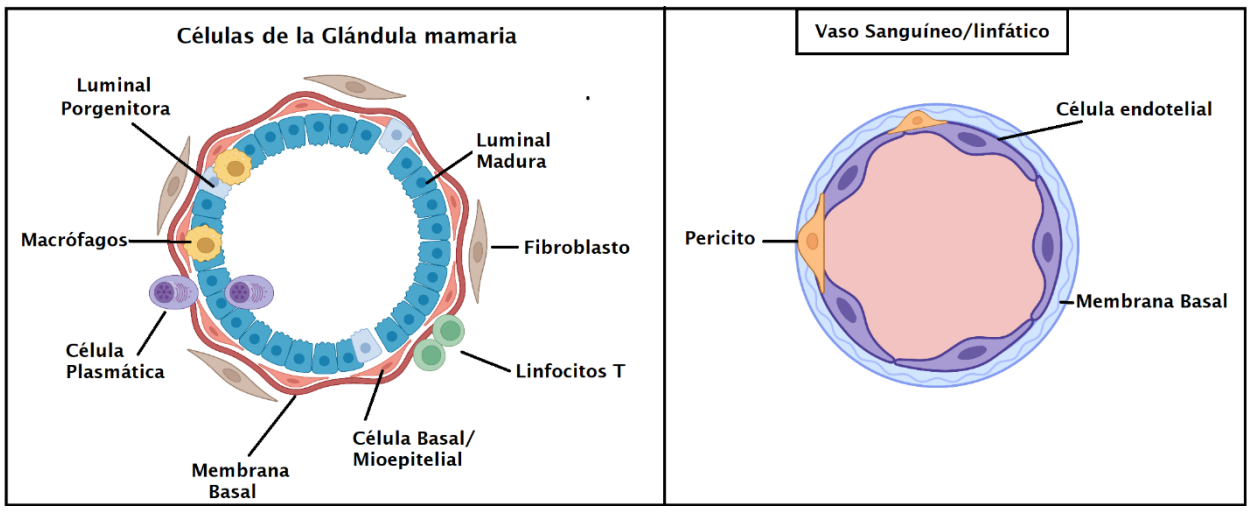


Figura 4.Tipos de células de la glándula mamaria.

Está compuesta por ductos y alveolos están compuestos por células epiteliales luminales que dan al lumen, las cuales están revestidas de células mioepiteliales, que a su vez están anclados a la membrana basal que se encuentra rodeada por una variedad de células estromales que forman la matriz extracelular. En combinación con una variedad de células inmunitarias (macrófagos, linfocitos B y células plasmáticas). Las células endoteliales vasculares cuyos nutrientes y productos viajan a través de la matriz, están cubiertas por pericitos y una membrana basal (imagen de creación propia https://app.biorender.com/).

Enfoque del trabajo

Los datos con los que se va a trabajar corresponden al trabajo de Pal *et al*., (2021) en el que se estudiaron las diferencia de los perfiles transcripcionales inter e intra paciente entre distintos tipos de células de tejido mamario: normal de mujeres (pre y post menopausicas), tumoral de (mujeres y hombres de los subtipos ER+, HER2+ y triple negativo (TNBC)) y prenoplásico. Para lo cual primero crearon atlas de cada tipo de muestra para identificar los tipos celulares encontrados en cada uno y para luego poder compararlos entre estados. Este enfoque les permitió identificar varios patrones con potencial terapéutico, desde cambios en la abundancia relativa de las distintas poblaciones de células, diferencias transcripcionales especificas a cada tipo inter pacientes e intra paciente y diferencias entre células del mismo tipo en el mismo paciente.

Para el análisis de los datos de Pal *et al*., (2021) se usó Seurat para la obtención de grupos celulares. Para la identificación de las células utilizaron dos métodos diferentes. Para las células epiteliales utilizaron un método propio en el que crearon firmas de cada tipos de célula (luminal progenitora, luminal madura y basal) mediante perfiles transcriptomicos obtenidos por bulk RNA-seq de células divididas según estos tipos por cell-sorting y contras la cual mapearon los perfiles transcriptomicos de las células obtenidas por single-cell RNA-seq en un diagrama ternario. Para la identificación de del resto de tipos celulares, también se utilizó el método manual en base a marcadores de cada agrupación, sin embargo, para la identificación de marcadores se hizo un análisis de pseudo-bulk en el que se toman las células por grupos al que perteneces y por paciente, se juntan sus expresiones y se analizan como un solo perfil (de manera análoga a los perfiles obtenidos por RNA-seq) lo que le da mayor poder estadístico ya que le permite en cuenta la variación biológica entre diferentes pacientes (Chen at al., 2022). Los análisis funcionales se realizaron también mediante sobrerrepresentación pero utilizando la función kegga y goanna del paquete limma

Mientras el estudio de Pal *et al*., (2021) estaba enfocado en el análisis de las diferencias entre todos los diferentes estados omitiendo resultados no relevantes para el estudio como los relacionados con el proceso de obtención de los grupos de células y la caracterización funcional de cada uno. El presente trabajo se limita a la identificación y caracterización funcional de las células presentes en el tejido normal, la descripción del proceso de obtención de los grupos con la herramienta Seurat y la identificación con clusterProfiler como parte del aprendizaje del análisis bioinformático de datos single-cell. Para esto se van a descargar las cuentas ya alineadas y demultiplexadas, y se van a analizar los resultados del proceso de agrupación celular por Seurat y de la identificación y caracterización por clusterProfiler.

Hipótesis

Las firmas y asignación de grupos por medio de las funciones de Seurat a partir de los perfiles transcriptómicos de las células de tejidos mamarios, permiten su agrupación según tipos celulares, identificación y caracterización funcional.

Objetivos

Objetivo general

- Identificación y caracterización funcional de los tipos celulares típicos presentes en una muestra de tejido de glándula mamaria mediante la herramienta Seurat.

Objetivos específicos.

Agrupamiento de las células en base a sus perfiles transcriptómicos.

- Obtención de los perfiles de células únicas viables
- Identificación de genes y grupos de genes distintivos que permitan la agrupación de células.
- Asignación de las células a un grupo.

Obtención de marcadores de cada grupo celular formado.

- Obtención de genes diferencialmente expresados entre los grupos celulares formados

Identificación con células típicas del tejido mamario.

- Identificación de los marcadores con marcadores conocidos para tejidos mamarios

Materiales y métodos

Obtención de los datos

Se descargaron las matrices de cuenta (ya alineadas y demultiplexadas) GEO con número de acceso GSE161529, proveniente del trabajo de Pal *et al* (2021) pertenecientes a tejidos de glándula mamaria normal de 12 mujeres con los siguientes códigos de acceso GEO

GSM4909253 Pre-menopausia, 4966 células

GSM4909254 Pre-menopausia, 7130 células

GSM4909257 Pre-menopausia, 7412 células

GSM4909261 Pre-menopausia, 3443 células

GSM4909263 Pre-menopausia, 1678 células

GSM4909265 Pre-menopausia, 5665 células

GSM4909266 Pre-menopausia, 4605 células

GSM4909268, pre-menopausia, 7371 células

GSM4909270 Post-menopausia, 10178 células

GSM4909271 Post-menopausia, 2320 células

GSM4909272 Post-menopausia, 2463 células

GSM4909274 Post-menopausia , 1914 células

Obtención de grupos celulares

Para obtener grupos celulares se utilizaron las funciones disponibles en el paquete Seurat versión 4.3.0.1

Control de calidad y normalización

Se filtraron las células que tuvieran menos de 200 genes y los genes que se expresaran en menos de 3 células. Para lo cual se ajustaron los parámetros min.cells y min.features de la función **CreateSeuratObject()**

Luego para eliminar datos de GEMS vacías o que hayan capturado más de una célula, se filtraron los índices según el número de transcritos por célula y porcentaje de recuentos mitocondriales. Para esto se utilizó la función subset() conservando las células con más de 200 y menos de 2500 transcritos por célula y con menos del 5 % de recuentos mitocondriales. Luego, los datos se normalizaron con la función NormalizeData() con el método LogNormalize.

Pre-procesamiento de los datos.

Se seleccionaron las 2000 genes más variables utilizando la función FindVariableFeatures() con el método "vst" y ajustado con el parámetro nfeatures = 2000, y un factor de escala de 1000.

Reducción de dimensionalidad y Clustering

A continuación, se escalaron los datos con la función ScaleData() y se realizó el PCA con la función RunPCA() calculando los primeros 50 componentes por defecto y para identificar los componentes que explicaran la

mayor variabilidad se ocupó la función ElbowPlot() que grafica la desviación/varianza estándar en el conjunto de datos atribuidos a cada componente principal. Para encontrar los grupos se utilizaron las funciones FindNeighbors() y FindClusters() en el que la resolución se determinó manualmente.

Integración de los datos

Se separaron los datos por paciente, y se aplicó la normalización y la selección de genes variables por separado. Luego, se seleccionaron los genes variables en común entre las muestras con la función SelectIntegrationFeatures(), y se encontraron las células comunes con la función FindIntegrationAnchors(). se corrigieron de los valores de expresión de todos las muestras con la función IntegrateData(). Tras la integración de datos se volvieron a correr los pasos desde la normalización para obtener los grupos celulares.

Identificación de células

Se encontraron los genes marcadores para cada tipo celular usando la función de Seurat FindAllMarkers() que determina los genes diferencialmente expresados de cada grupo, al realizar la prueba de suma de rangos Wilcoxon entre cada grupo generado, con el resto de grupos,sobre los genes que se expresaran en al menos el 25% de las células del grupo con el parámetro min.pct = 0.25. Luego se evaluaron los marcadores encontrados con el programa clusterProfiler versión 4.6.2 sobre los marcadores con veces de cambio promedio mayor a 0.5 (medida en logaritmo de base 2) y un *p value* menor a 0.05. Para lo cual se usó la función enricher() contra lista de marcadores de humanos (Homo sapiens) obtenidos PanglaoDB sitio https://panglaodb.se/. Además, se evaluó la expresión de genes marcadores clásicos de los tipos célulares y marcadores encontrados en los atlas de single cell de tejido de mama de Kumar *et al*.,2023 y Pal *et al*., 2021 mediante la función FeaturePlot() de Seurat.

Análisis funcional

Se usaron los marcadores obtenidos con FindAllmarkers() con un *p value* menor a 0.05. Las veces de cambio a usar se establecieron evaluando las cantidades de genes capturados en los procesos, funciones y vías. mediante el programa clusterProfiler usando las funciones enrichKEGG() y enrichGO()

Resultados

Control de calidad

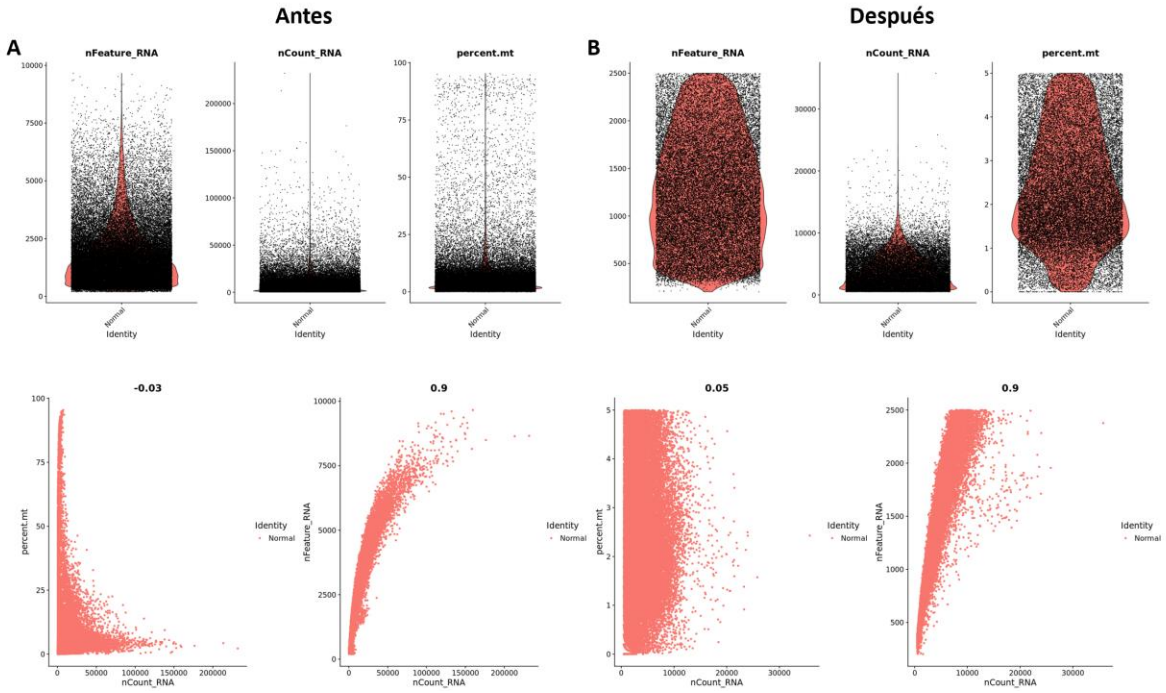


Figura 5.Control de calidad (QC) de la matriz de cuentas de Pal et al (2021)

En la parte superior, diagramas de violín de distribución número de genes detectados en cada GEM (nFeature), número total de cuentas detectadas en cada GEM (nCount), y el porcentaje de genes mitocondriales mapeados cada GEM (percent.mt). En la sección inferior, relación entre número de cuentas y porcentaje de genes mitocondriales presentado en cada GEM y número de cuentas vs número de genes presentados por cada GEM. Cada punto de datos representa el código de un GEM. (A) distribuciones antes del filtrado (B) distribuciones después del filtrado mediante la función subset().

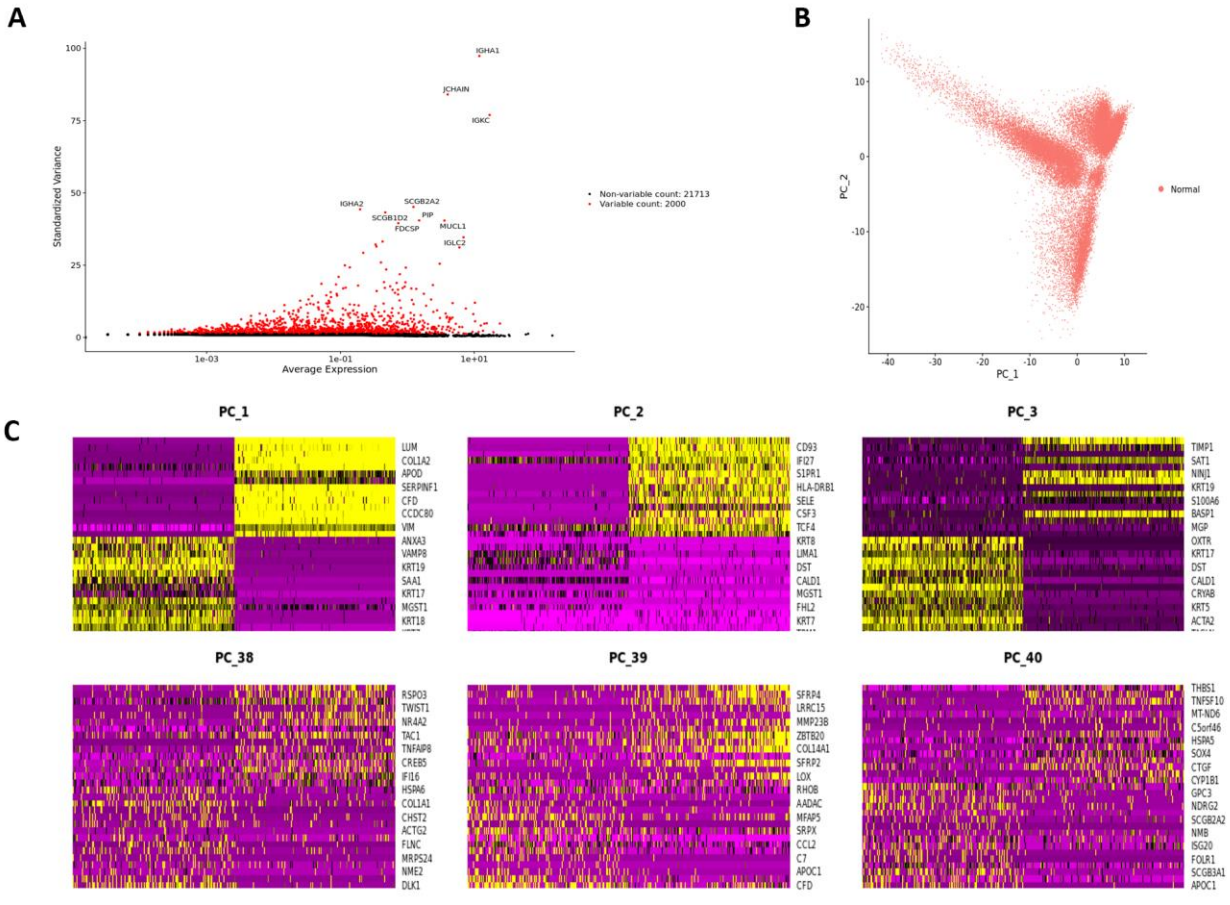
En la figura 5.A se muestra el análisis de las características (número de genes detectados en cada GEM, número total de cuentas detectadas en cada GEM y el porcentaje de genes mitocondriales mapeados cada GEM) que se usaron para evaluar la calidad de los datos y determinar si los GEM generados cumplían con los requisitos mínimos para ser considerados como una célula viable. Antes de la filtración, se obtuvieron 23713 genes y 57440 GEM, en las cuales el número de genes expresado variaba entre 0 y 10000. Aunque la mayoría de las células expresaba entre 0 y 2500 genes. Lo mismo pasaba con el número de cuentas cuya distribución iba de 0 a los 200000 por célula, aunque la mayoría se encontraba entre 0 y 25000.En cuanto al porcentaje de genes mitocondriales, este se movía entre 0 y 100, pero la mayoría se encontraba entre 0 y 20.

En la sección inferior de la figura 5 se muestra la relación entre el número de cuentas presentes en los GEM y su porcentaje de genes mitocondriales, se espera que no muchas cuentas estén asociadas a células

con un alto porcentaje de genes mitocondriales detectados. Antes del filtrado, las células con bajo porcentaje de genes mitocondriales se asocian en general con un mayor número de cuentas, mientras las que presentan un alto porcentaje, se acercan a un número de cuentas de 0. Luego se ve la relación entre el número de cuentas y el número de genes, el cual debe tener una relación lineal, es decir, mientras más genes se detecten, debería haber más cuentas. Si fuera una línea recta sobre el eje "x" significaría que solo se detectaron pocos genes y estos pocos han sido secuenciados muchas veces. Por el contrario, si fuera una línea sobre el eje "y" significaría que se detectaron varios genes pero con poca profundidad de secuenciación. La tendencia lineal se mantiene tanto en los datos sin filtrar como en los filtrados, hasta las células que tienen 5000 genes.

Estos datos se filtraron para conservar las células con más de 200 y menos de 2500 genes expresados por célula y con menos del 5 % de recuentos mitocondriales. Esto se ve reflejado en las distribuciones post filtración en la figura 5.B de la cual quedaron 23713 genes y 30073 células. Por otro lado, se ve que la mayoría de las cuentas corresponden a células con un porcentaje de genes mitocondriales menor al 5% y que la relación entre genes y número de cuentas es lineal, cumpliendo con las condiciones para realizar el análisis

Obtención de grupos celulares datos sin integrar



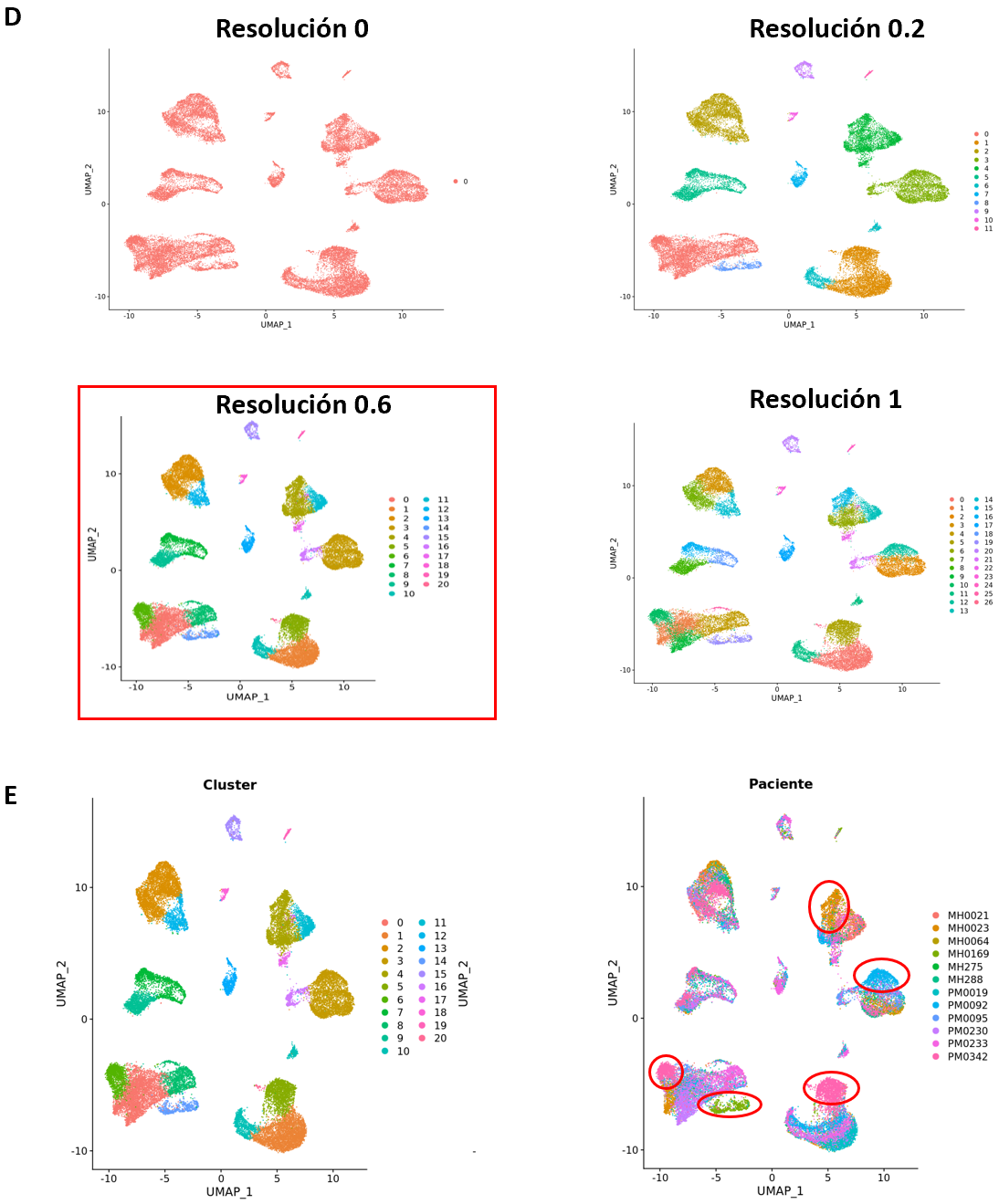


Figura 6.Grupos celulares obtenidos con las funciones de Seurat

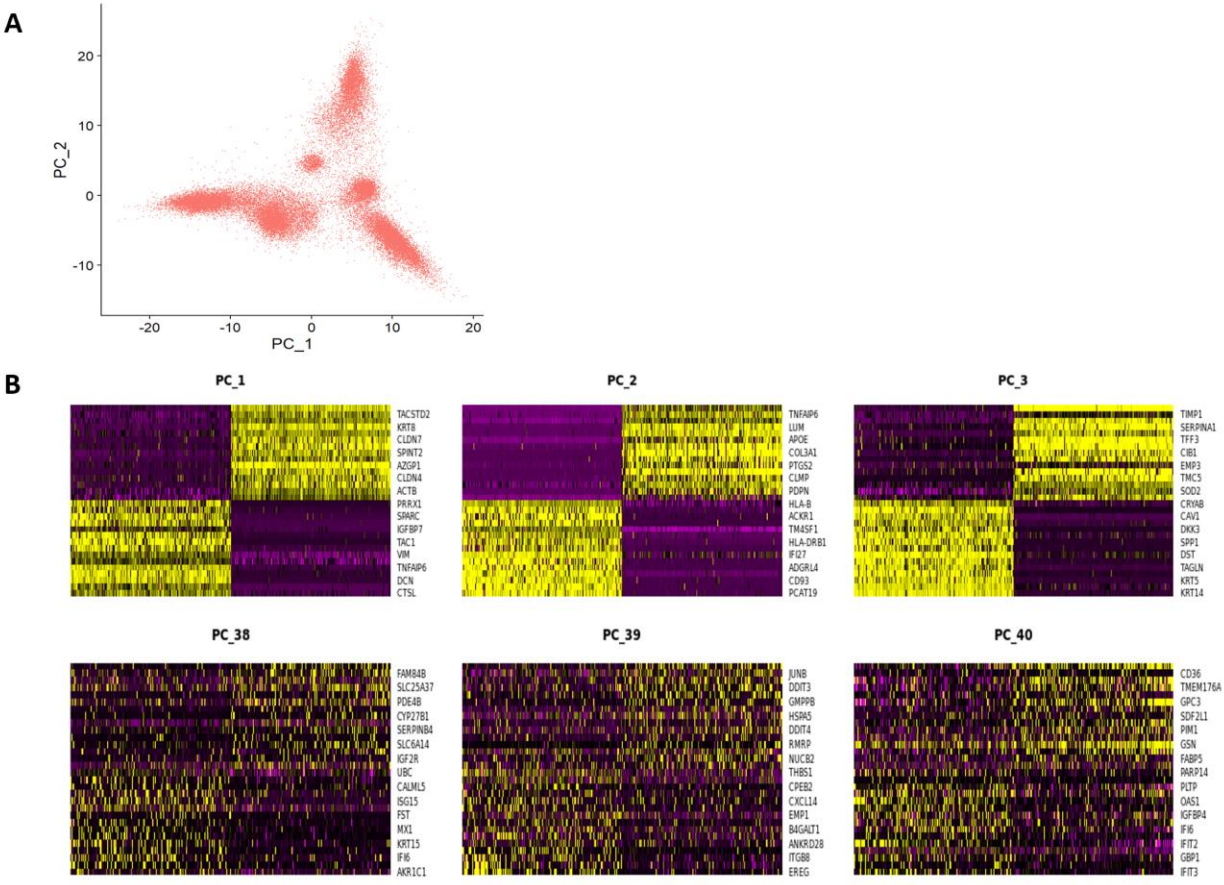
(A) identificación de genes que presentaron mayor variabilidad entre muestra con la función FindVariableFeatures() (B). Grupos formados por los dos primeros componentes resultantes de la reducción dimensional con la función RunPCA(). (C). Heatmap de las firmas pertenecientes los 3 primeros vs los 3 últimos componentes. (D) grupos obtenidos de células con las funciones FindNeighbors() y FindClusters() bajo distintas resoluciones (en rojo la resolución elegida).(E) células coloreadas según al que el grupo celular que pertenecen vs al paciente al que pertenecen.

En 6.A se ve el resultado de los genes más variables, entre los cuales se encuentran IGHA1 JCHAIN, IGKC, IGHA2, SCGB2A2, SCGB1D2, PIP, MUCL1, FDSCP e IGLC2.

En 6.B se ve la representación del resultado de PCA en el espacio de los 2 primeros componentes, en los cuales las células se separan en 4 grandes grupos diseminados, por lo que estos dos no son suficientes para separar correctamente los datos en grupos celulares discretos. Para determinar cuántos componentes usar, para los siguientes pasos, se utilizó el método del codo con el cual se ve la desviación/varianza estándar en el conjunto de datos atribuidos a cada componente principal y se determinó que hasta los 40 primero componentes principales, se capturaba gran parte de la variabilidad y se usaron estos para el siguiente paso. Esto se puede visualizar en la figura 6.C, en las que se construyeron heatmaps de las firmas de genes asociada a los 3 primeros componentes vs los 3 últimos y se ve que mientras en los 3 primeros componentes se ve una clara división de los datos, se ve que en los últimos 3 la definición de las divisiones es menor.

En 6.D por medio de UMAP con resolución 0 se pueden visualizar cómo se agrupan las células utilizando los primeros 40 componentes del PCA, en un espacio de dos dimensiones. En este punto se asignan células a grupos celulares. Se evaluaron diferentes resoluciones que permite establecer la distancia entre las células para considerarse como perteneciente a un mismo grupo y se ve cómo se van formando los grupos. En primera instancia con resolución 0 se tienen 10 grupos los cuales fueron aumentando hasta 26 grupos al aumentar la resolución. Por otro lado, se ve que al aumentar los grupos hay una menor definición de los grupos, es decir, las células de un grupos se van mezclan con los de otros grupos adyacentes, por lo que al final se decidió utilizar la resolución 0.6 que se consideró con una buena definición en la que los grupos celulares no se mezclan, pero captan espacios distintivos dentro de cada agrupación, manteniendo la posibilidad de detectar distintos estados o subtipos dentro de los grupos celulares. Por último, se evaluó cómo estos tipos celulares se integran en el contexto de los pacientes, si es que hay células que se agruparon en función de características asociadas a los pacientes en lugar de características asociadas a los tipos celulares. Se vio que en general hay una buena integración de las células, los grupos celulares formados presentan mezclas de células provenientes de los distintos pacientes, por lo que lo más probable es que las células se hayan agrupado por sus características biológicas y no por efectos de lote. Sin embargo, hay ciertos grupos que solo pertenecen a un paciente, que podría indicar contaminación con otros tipos de células externas a la glándula, aunque también podría ser células de interés, que al contrario, no se lograron obtener en los otros pacientes. Para el fin de este trabajo que tiene como objetivo un análisis simplificado, y se realizará la integración para analizar solo los tipos celulares comunes entre las diferentes personas.

Obtención de grupos celulares en datos integrados



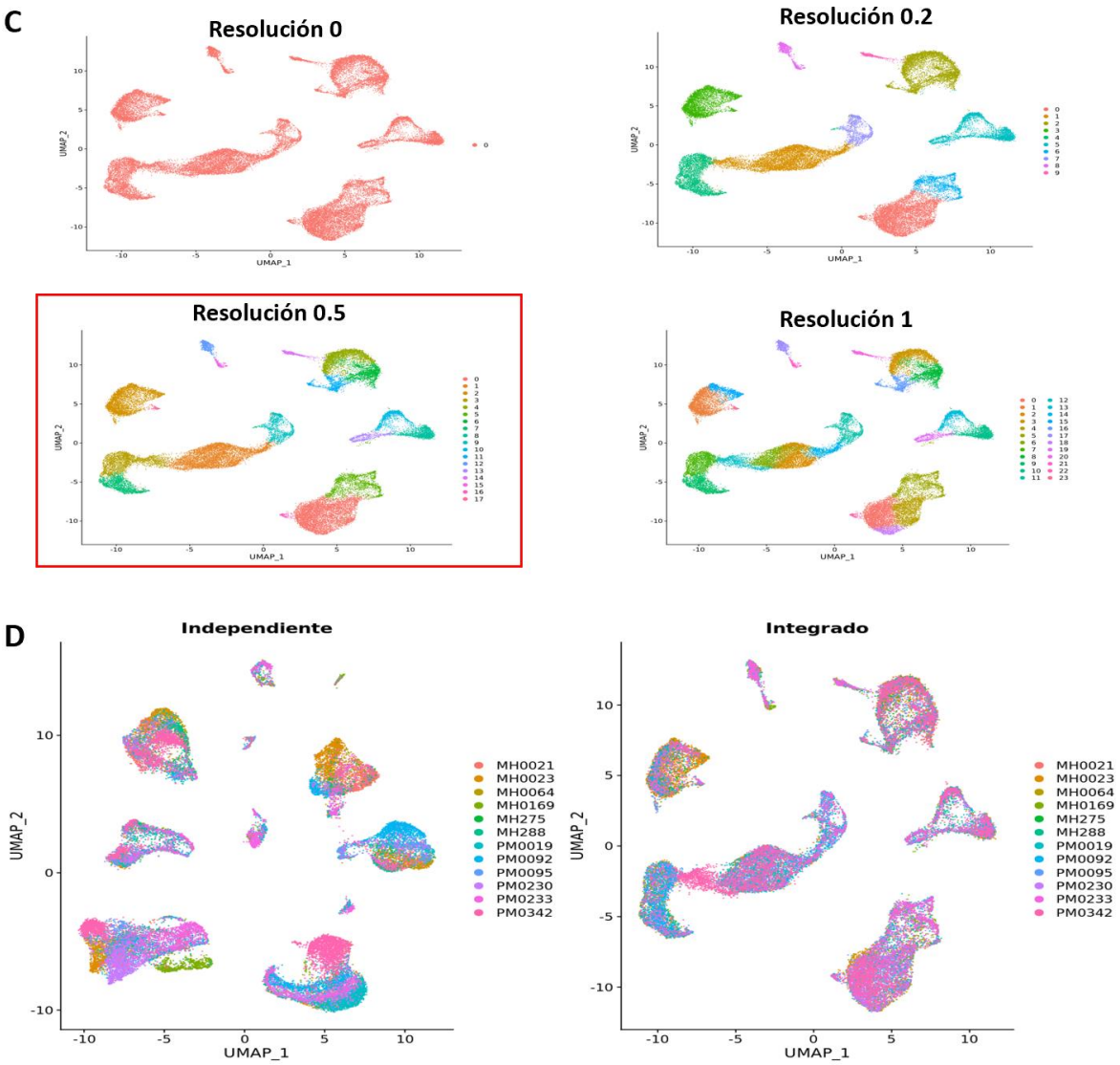


Figura 7.Grupos celulares obtenidos de los datos integrados con las funciones SelectIntegrationFeatures(), FindIntegrationAnchors() y IntegrateData().

(A). Grupos formados por los dos primeros componentes resultantes de la reducción dimensional con la función RunPCA(). (B). Heatmap de las firmas pertenecientes los 3 primeros vs los 3 últimos componentes. (C) grupos obtenidos de células con las funciones FindNeighbors() y FindClusters() bajo distintas resoluciones (en rojo la resolución elegida).(D) células coloreadas según al que el grupo celular que pertenecen vs al paciente al que pertenecen.

Tras la integración de datos se volvieron a correr los pasos desde la normalización y se volvieron a evaluar las características del proceso.

En 7.A se ve la representación del resultado de PCA en con los 2 primeros componentes, en los cuales se forman 6 grandes grupos más discretos en comparación con los grupos formados en los datos sin integrar. De todas maneras, se evaluó con el método del codo cuántos componentes captaban la mayor variabilidad para los siguientes pasos y se decidió tomar hasta los primeros 40 componentes. Se graficó un heatmap de la firma de genes asociada a los componentes generados en los 3 primeros componentes vs los 3

últimos (figura 7.B) y se ve que mientras los 3 primeros componentes se asocian con una división clara de los datos, se ve que en los últimos 3 las divisiones pierden resolución. Las agrupaciones generadas por los 40 componentes se pueden visualizar usando UMAP (7.C resolución 0). Luego al evaluar la asignación de grupos con diferentes resoluciones, se obtuvieron desde 9 hasta 23. Finalmente se eligió una resolución de 0.5 quedando con 17 grupos para analizar. Estos grupos se pueden modificar tras el análisis de identificación, evaluando la distribución de marcadores de tipos celulares.

Identificación de los grupos celulares

Para identificar los grupos obtenidos se hizo etiquetado manual. Como primer enfoque, se realizó un análisis de expresión diferencial para cada grupo vs el resto de grupos celulares encontrados con la función FindAllMarkers() con el fin de identificar marcadores que correspondan a las firmas características distintivas de cada tipo celular. Estos marcadores se evaluaron por un análisis de sobrerrepresentación con enricher de clusterProfiler contra las listas de marcadores celulares de la base de datos PanglaoDB, las etiquetadas encontradas se corroboraron o corrigieron mediante la función FeaturePlot() la cual marca las células según el nivel de expresión de un gen consultado, esto permitió encontrar qué células expresaban marcadores conocidos clásicos de las células esperadas. También se utilizó la función VlnPlot() que gráfica en un diagrama de violín la distribución de la expresión de los marcadores en cada grupo de células para corroborar la especificad de los marcadores.

Células epiteliales

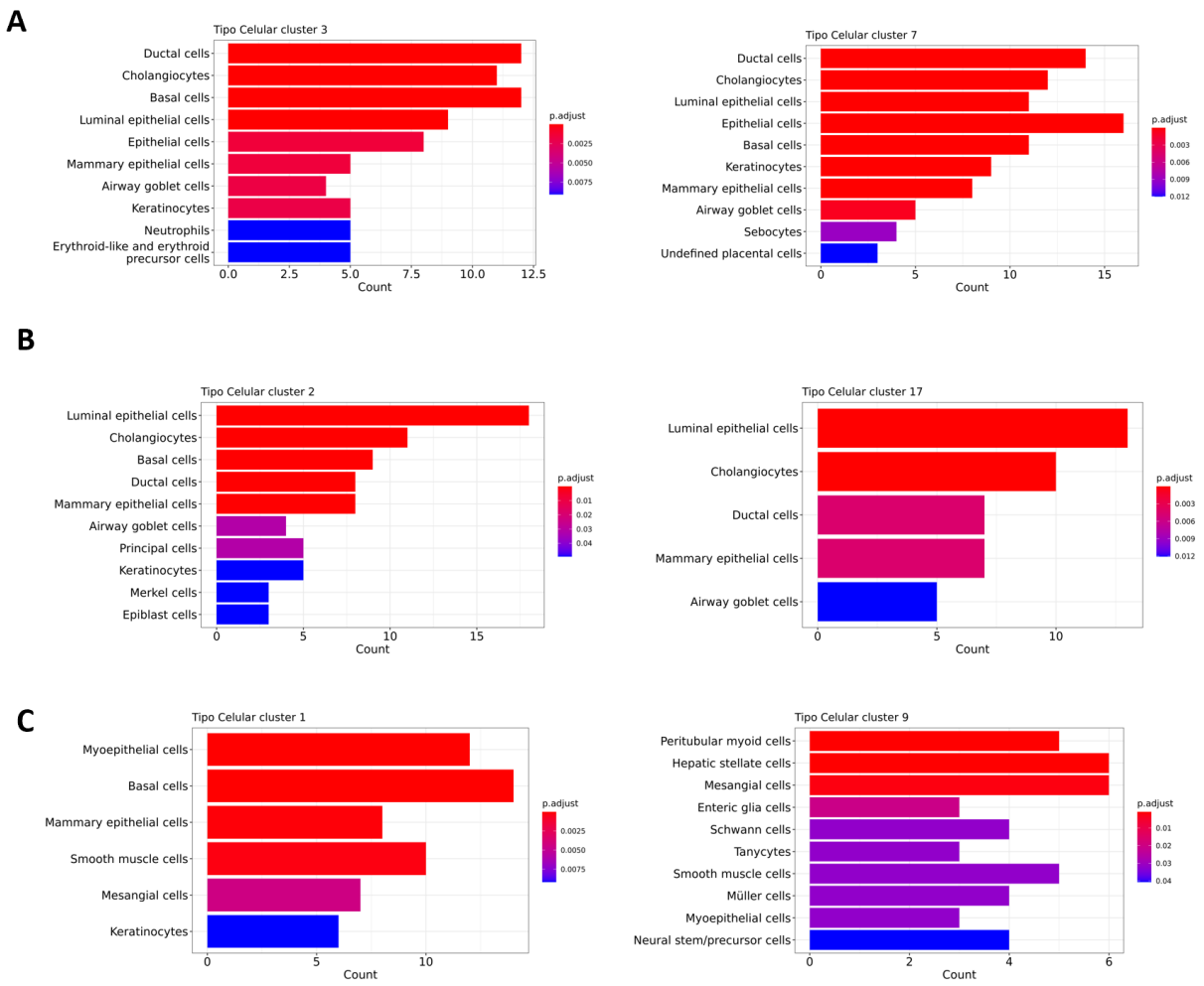


Figura 8.Identificación de los grupos celulares (3,7,2,17,1 y 9).
Etiquetas encontradas por el análisis de sobrerrepresentación con la función enricher de clusterprofiler con los marcadores con un avg_logFC mayor a 0.5.de los grupos de (A) células 3 y 7.(B) análisis de sobrerrepresentación de los grupos de células 2 y 17. (C) análisis de sobrerrepresentación de los grupos de células 1 y 9.

Entre los tipos celulares con mayor número de genes coincidentes en el grupo 3 fueron células ductales del páncreas, luminales, basales y colangiocitos, en el 7 células epiteliales, células del ducto de páncreas, colangeocitos, luminales y basales. Se encontró en los grupos 2 y 17 un mayor consenso presentando mayor número de genes de células epiteliales luminales. El grupo 1 muestra de células mioepiteliales, basales de musculatura lisa todas correspondiente con un perfil de células mioepiteliales, mientras que en el grupo 9 se encontró correspondencia con células miodes peritubulares, células mesangiales y células estrelladas hepáticas, aunque en menor proporción también se encontraron células musculares y mioepiteliales coincidentes con el grupo 9. Dado que en algunos de los grupos no se encontraron etiquetas claras coincidentes con lo esperado para las células mamarias, se realizó un análisis visual de marcadores clásicos usados en histología (y que presentan una buena correlación con niveles de ARNm) mediante la función featureplot() que permite visualizar los niveles de expresión de cada gen en cada célula.

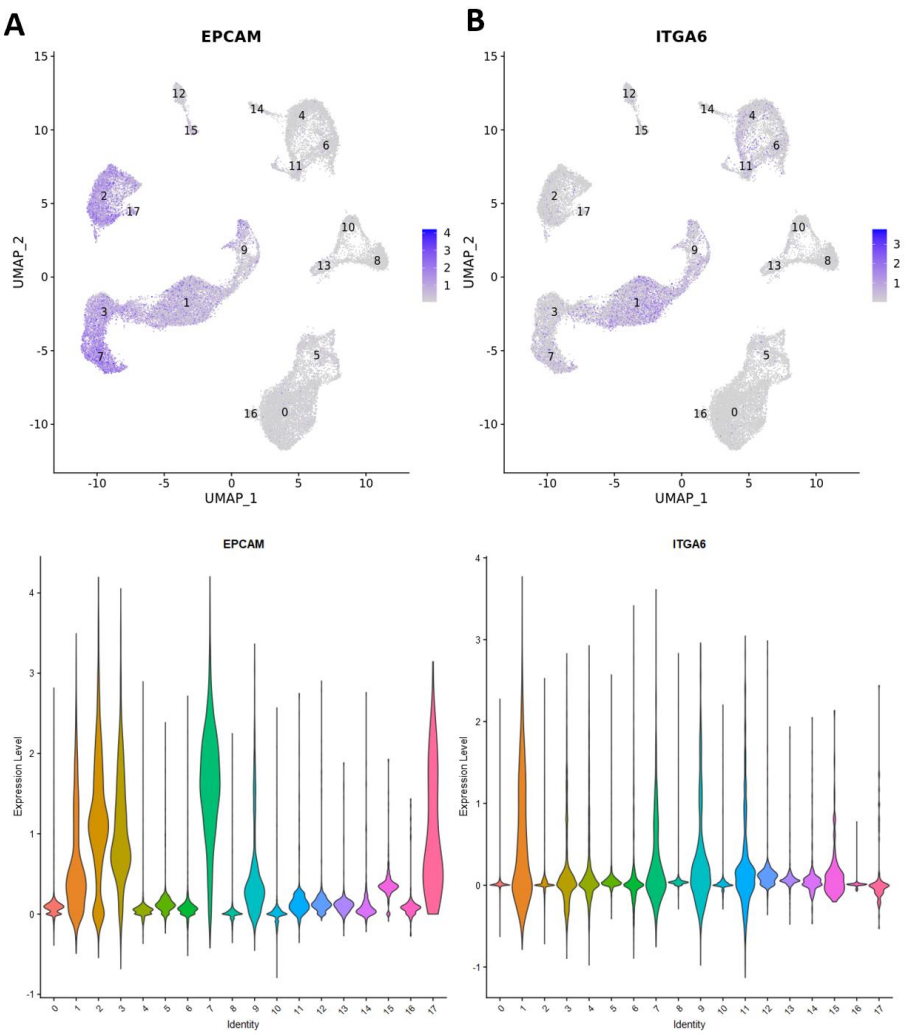


Figura 9.Exploración de la expresión del marcador EPCAM de células epiteliales.
UMAP que muestran los niveles de expresión del marcador epitelial EPCAM y la integrina ITGA6 en los grupos celulares obtenidos.

Se muestra qué grupos expresan el gen EPCAM que es un marcador clásico de células epiteliales. Se ve una mayor expresión en las células que forman los grupos 2,17 ,3,7,1 y 9. Por su parte, el marcador de integrina ITGA6 es una proteína involucrada en la unión a la matriz extracelular, que sirve de marcador negativo de las células epiteliales luminales, ya que se expresa en tejidos que se encuentren en contacto como las mioepiteliales, endoteliales y a cierto grado, las progenitoras. Para la expresión de esta proteína se ve que de las células de los grupos encontrados anteriormente que presentaban EPCAM, 2 y 17 muestran los niveles más bajos de ITGA6 por lo que estas podrían corresponder a células epiteliales luminales.

- Células luminales progenitoras

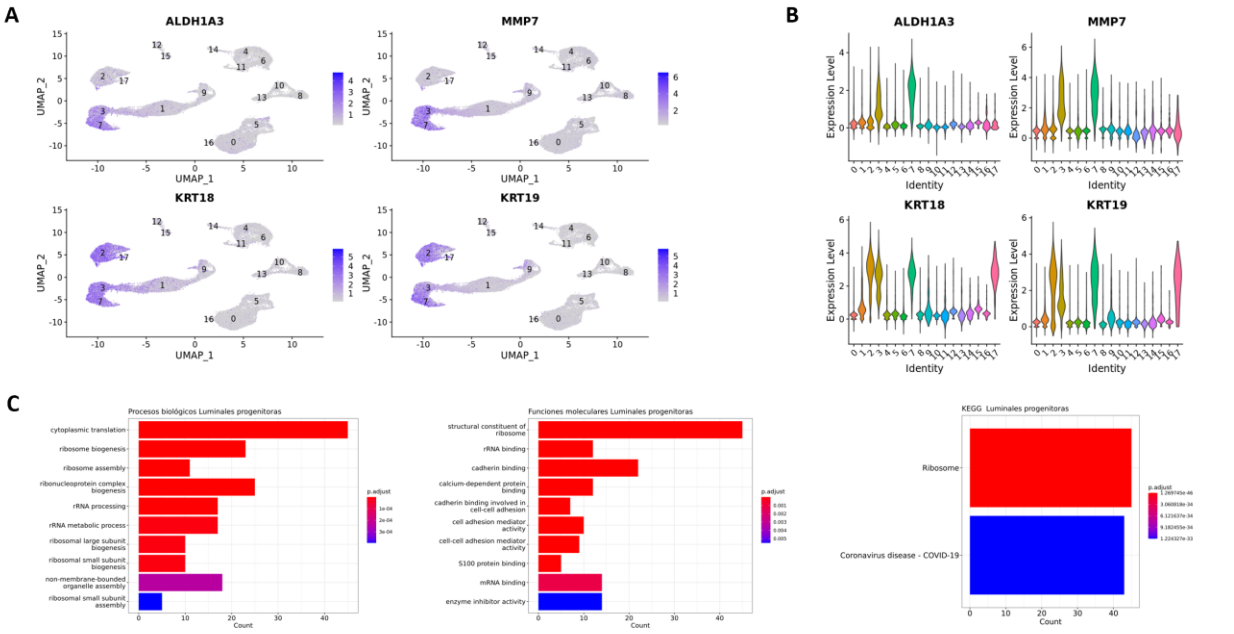


Figura 10. Identificación y caracterización de poblaciones de células luminales progenitoras.
(A)UMAP con los niveles de expresión de los marcadores epiteliales luminales progenitoras (ALDH1A3, MMP7, KRT18 Y KRT19). (B) Distribución de la expresión de los marcadores en los diferentes grupos de células(C) Análisis funcional del grupo positivo para los marcadores luminales progenitoras con avg_logFC mayor a 0.5.

Para la identificación de células progenitoras se utilizó una combinación de marcadores utilizados tanto en histología como en los análisis single-cell, los KRT18 Y 19 que se expresan tanto en células epiteliales maduras como progenitoras, ALDH13A que se expresan en tipos de células madre y progenitoras y el gen MMP7 encontrado en otro análisis single cell como marcador. KRT18 Y 19 se presentaron en las células de los grupos 3,7, 2 y 17. Mientras que los marcadores ALDH13A y MMP7 se encontraron en las células del grupo 3 y 7, por lo que se identificó a los grupos 3 y 7 como células epiteliales luminales progenitoras.

Entre los genes marcadores se encontraron varios genes de citoqueratinas KRT7,8,15,16,17,18,19 y 23, claudinas CLDN3, CLDN4 y CLDN7, genes relacionados con la señalización de ácido retinoico, el receptor de ácido retinoico RARRES2 y un gen de la enzima aldehído deshidrogenasa ALDH1A3 que produce ácido retinoico. Y varios genes de proteínas ribosomales entre ellas proteínas L ribosomal (RLPL 11, 12,13,14 entre otros) y proteínas ribosomales RPS2,21,23,24,27 entre otras.

Luego el análisis funcional se realizó en función sobre todos los grupos identificados como progenitores luminales. Los genes marcadores mostraron enriquecimiento en gran parte a funciones y las vía relacionadas con el ribosoma y el coronavirus (relacionada con altos niveles de traducción), biogénesis de ribonucleoproteinas, metabolismo de rRNA y la vía de biogénesis de ribosomas. Mientras en los procesos biológicos se encontraron, además de los relacionados con el ribosoma, procesos relacionados con la unión a cadherina y de adhesión celular.

- Células epiteliales luminales maduras

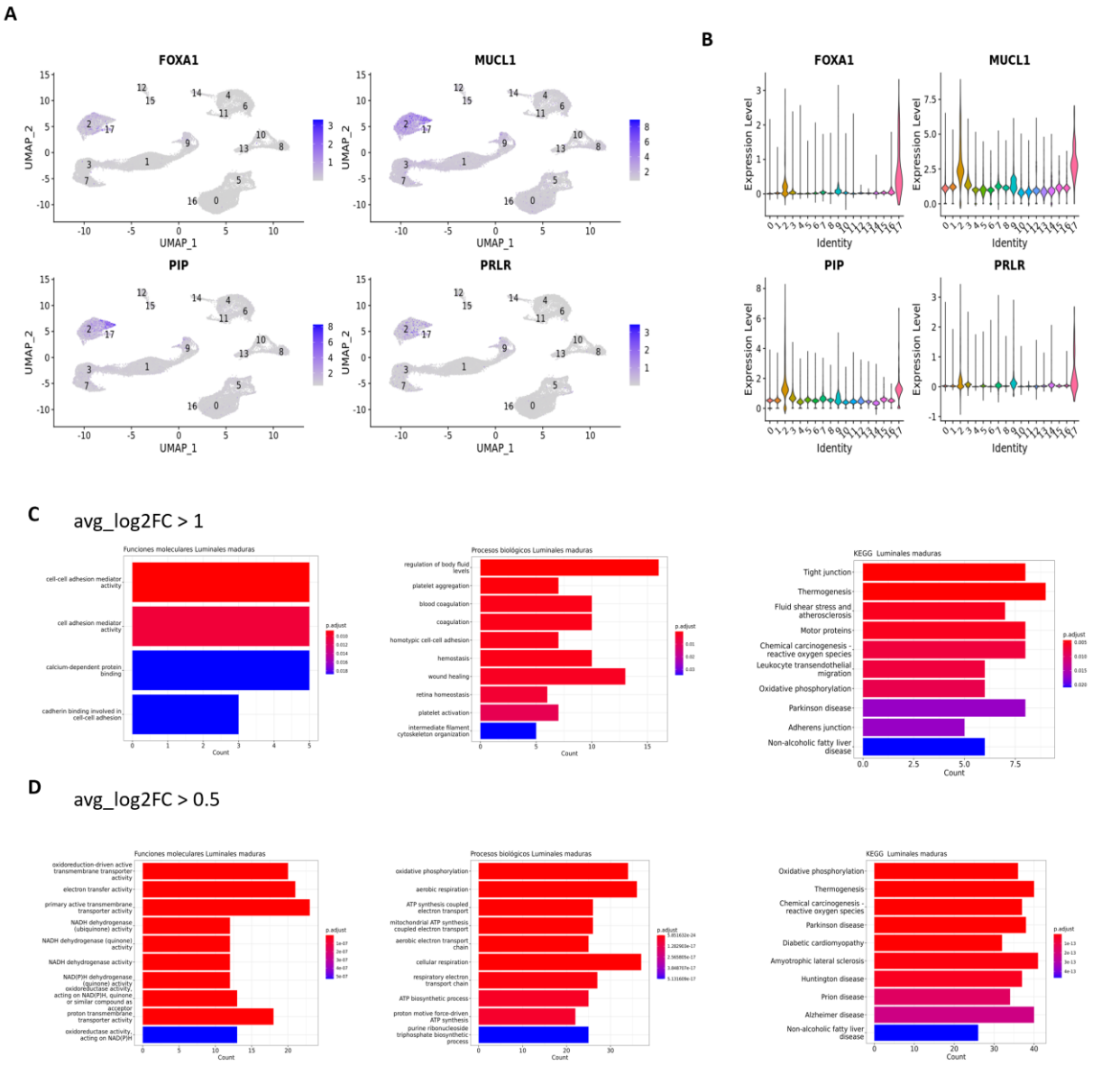


Figura 11. Identificación y caracterización de poblaciones específicas de células luminales maduras.

(A) UMAP con los niveles de expresión de los marcadores epiteliales luminales maduras (FOXA1, MUCL1, PIP y PRLR).(B) Distribución de la expresión de los marcadores en los diferentes grupos de células(C) Análisis funcional del grupo positivo para los marcadores luminales maduras con avg_logFC mayor a 1. (D) Análisis funcional del grupo positivo para los marcadores luminales maduras con avg_logFC mayor a 0.5.

En conjunto con los marcadores anteriores, EPCAM, KRT18 Y KRT19 que mostraban expresión en las células de los grupos 3 y 7. Se evaluaron genes marcadores usados ampliamente PIP y PRLR y los genes encontrados en otros trabajos single cell FOXA1, MUCL1. Se identificaron los genes 3 y 7 como células luminales maduras.

En los genes diferencialmente expresados se encontraron genes de citoqueratina (KRT7, KRT8, KRT18, KRT19 y KRT23) genes de uniones estrechas las claudinas CLDN7 y CLDN4, la caderina E (codificada CDH1), y genes de relacionados con su actividad secretora, (MUCL1 y PIP), genes que median respuestas hormonales FASN que media metabolismo lipídico controlado por las hormonas estrógeno y factores de crecimiento y el gen AREG de la anfirregulina que es un factor de crecimiento tipo EGF. Así como genes de actinas y miosinas de células no musculares ACTB, ACTG1, MYL12A, MYL12B, MYL6, MYH9 y MYO6 que se relacionan con tráfico vesicular.

En el análisis de procesos celulares, funciones moleculares y un análisis de vías en los genes con un valor de veces de cambio mayor a 1 resultaron enriquecidas, las funciones de unión célula-célula y unión dependiente de calcio y la regulación de niveles de fluidos corporales, el cual incluye procesos como regulación de la lactancia. En las vías enriquecidas se encontraron las uniones estrechas y adherentes y vías relacionadas al tráfico vesicular de fosforilación oxidativa, relacionado la esteatohepatitis no alcohólica. Por otro lado, se encontraron otros procesos menos relacionados al epitelio como coagulación sanguínea, agregación plaquetaria y hemostasis. Se extendió el corte de veces de cambio a 0.5 y aparecieron enriquecidos procesos relacionados con energía y transporte como transporte activo de membrana por oxidorreducción y la biosíntesis de trifosfatos de ribonucleósidos de purina y vías relacionadas como las vías implicadas en la esclerosis lateral amiotrófica, enfermedad de huntigton y termogénesis.

- Células basales

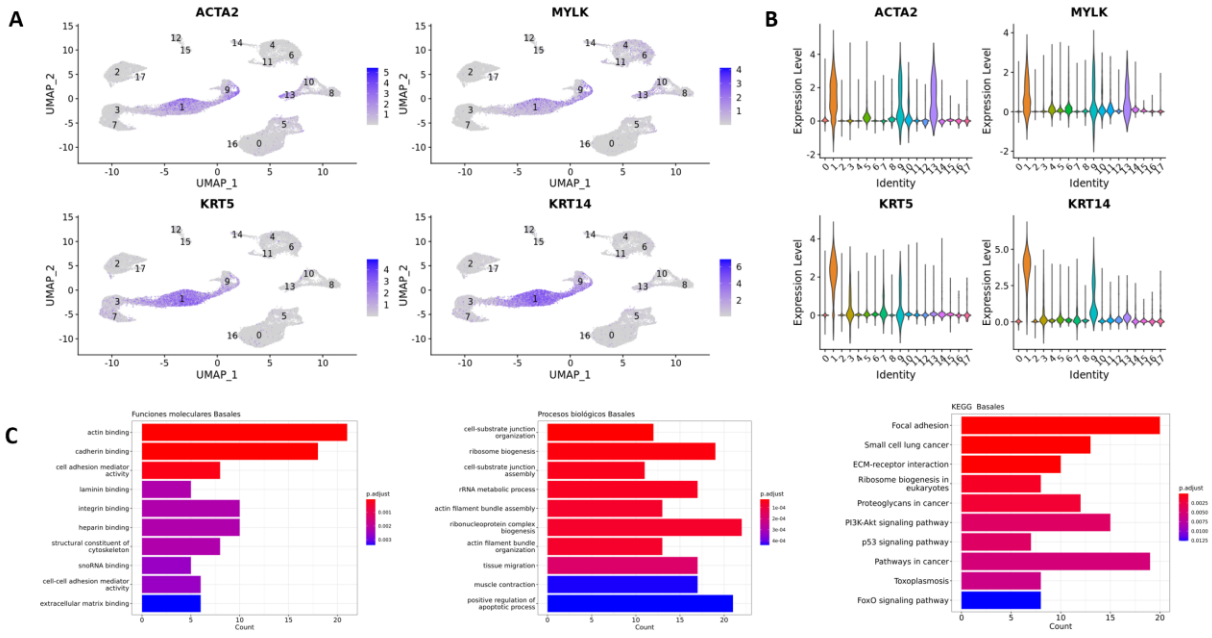


Figura 12.Identificación y caracterización de poblaciones de células basales.

(A)UMAP con los niveles de expresión de los marcadores epiteliales basales (ACTA2, MYLK, KRT5 y KRT14).(B) Distribución de la expresión de los marcadores en los diferentes grupos de células(c) Análisis funcional del grupo positivo para los marcadores basales con avg_logFC mayor a 0.5.

Los genes de citoqueratinas, marcadores de células mioepiteliales se expresan en el grupo 1 y 9, mientras los genes de función contráctil (ACTA2 y MYLK) se expresan en el grupo 13, que se verá adelante que pertenece a grupos de pericitos que también tienen función contráctil.

Entre los genes diferencialmente expresados (material suplementario) se encontraron genes de citoqueratina para la formación de filamentos intermedios común células epiteliales (KRT5, KRT6B, KRT7, KRT14 Y KRT17). Se encontraron genes de integrina que permiten el anclaje del mioepitelio a la membrana basal, (ITGA2, ITGA6 e ITGB). También, se encontraron genes de actina (ACTG2 ACTA2 ACTN4 ACTN1), miosina (MYLK) y tropomiosinas (TPM1 y TPM2) relacionados con su función contráctil, y proteínas regulación de la contracción del músculo liso, las calponinas TAGLN y CNN1 y el receptor de ocitocina OXTR.

El análisis ontológico mostró que los genes diferencialmente expresados presentan funciones moleculares común con las células epiteliales (unión a cadherina y funciones de organización de filamentos intermedios, desarrollo de epidermis y piel debido a su alta expresión de citoqueratinas) y funciones que le permiten su anclaje a la membrana basal, como la unión a integrina, laminina, colágeno y a glucosaminoglucanos. Estos se relacionan con los procesos de ensamblaje y organización de unión célula-sustrato, y las vías de adhesión focal, interacción matriz extracelular receptor, y proteoglicanos en cáncer (que resulta de una expresión de genes de relacionados con la adhesión celular como colágenos e integrinas y caveolinas). Se encontraron enriquecidos la función de unión a actina y el proceso de contracción muscular y la vía regulación de actina del citoesqueleto, que concuerdan con su función contráctil. Por otro lado se encontraron enriquecidos los procesos y vías relacionados con el ribosoma, biogénesis de ribonucleoproteinas, metabolismo de rRNA y la vía de biogénesis de ribosomas

Células estromales

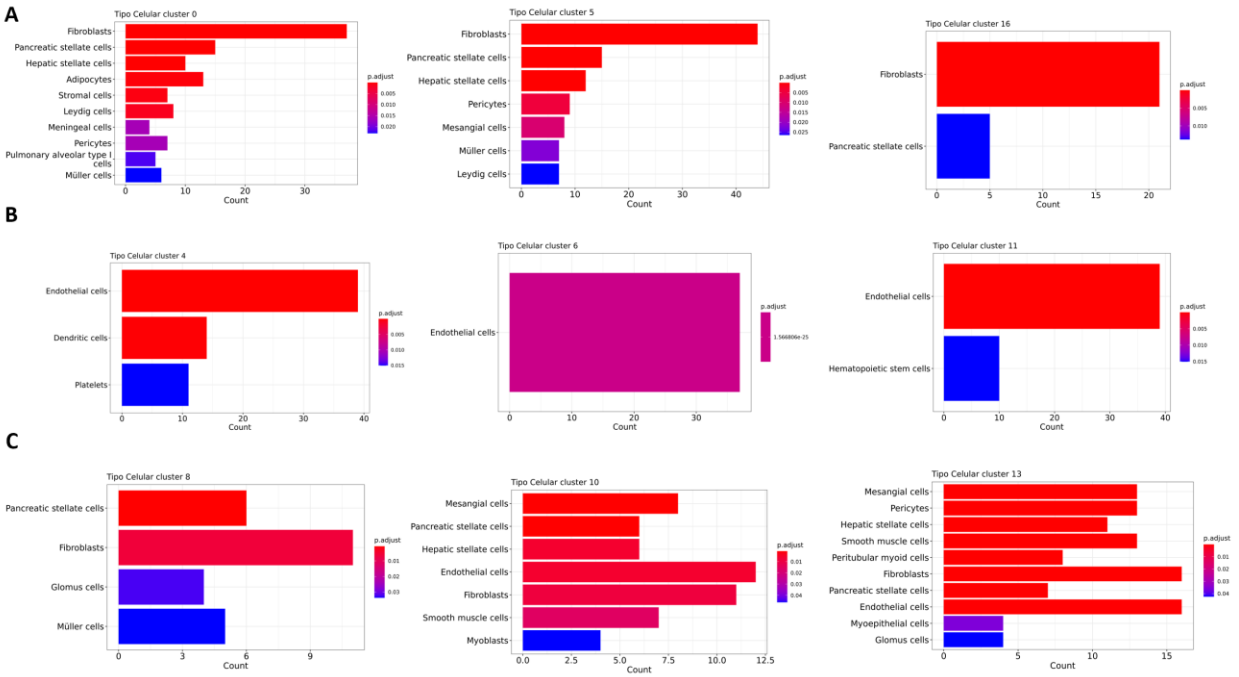


Figura 13.Identificación de los grupos celulares 4,6,11,8,10,13,0,5 y 16.

Análisis de sobrerrepresentación de los marcadores con un avg_logFC mayor a 0.5 de los grupos de(A) células 0,5 y16.(B) análisis de sobrerrepresentación de los grupos de células 4,6 y 11. (C) análisis de sobrerrepresentación de los grupos de células 8,10 y 13.

Para los grupos 0,5 y 16 se obtuvieron de manera consistente etiquetas de fibroblastos. Por su parte, los grupos 4,6 y 11 mostraron de manera consistente también etiquetas de células endoteliales. Del grupo 8 se obtuvo perfiles de células de fibroblastos y células estrelladas hepáticas. Los grupos 10 y 13 mostraron una mayor coincidencia con células mesangiales, endoteliales y fibroblásticas.

- Fibroblastos

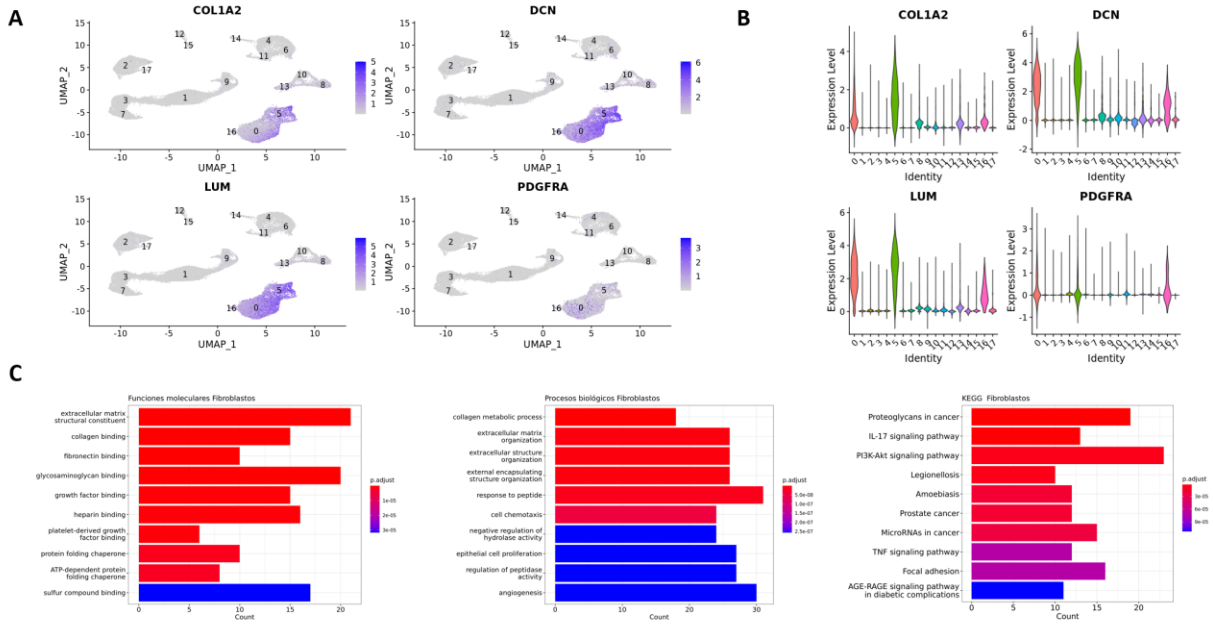


Figura 14. Identificación y caracterización de poblaciones de fibroblastos

(A)UMAP con los niveles de expresión de los marcadores de pericitos (COL1A2, DCN, LUM y PDGFRA).(B) Distribución de la expresión de los marcadores en los diferentes grupos de células(C) Análisis funcional del grupo positivo para los marcadores para fibroblastos con avg_logFC mayor a 0.5**.**

Los marcadores COL1A2 se expresa mayormente en 5 y de manera variable en 0 y en bajos niveles en 16. DCN y LUM se expresa en los grupos 16,0 y 5 de manera más homogénea y PDGFRA se expresa en los grupos 0,5, y 16 presentando mayores niveles en 16.

Genes componentes de la matriz extracelular, colágeno como COL1A2 y COL1A1, COL3A1, COL6A1, COL6A2, COL6A3, fibulina FBLN1 y FN1 una glicoproteína implicada en la adhesión celular y la organización de la matriz extracelular. Genes relacionados con el remodelamiento de la matriz, el gen PCOLCE que participa en el procesamiento del colágeno y la regulación de la matriz extracelular, las metaloproteinasas de matriz MMP2, MMP3, MMP12, y MMP10 y inhibidores de la actividad de las metaloproteinasas TIMP1, TIMP2 y TIMP3 y el gen DCN de la decorina involucrada en la producción de componentes de la matriz extracelular.

Genes de receptores de factores de crecimiento FGF7, FGFR1 y PDGFRA

Los genes diferencialmente expresados mostraron funciones relacionadas con componentes de la membrana plasmática como unión a colágeno, compuestos sulfatados, heparina. Al evaluar los procesos biológicos relacionados con estos genes se encontró organización y estructura extracelular, metabolismo del colágeno y regulación de actividad peptidasa. También se encontraron las vías relacionadas con la matriz extracelular, como enriquecimiento proteoglican en cáncer, adhesión focal y vías de regulación de la actividad fibroblástica para inflamación y reparación de tejido como la vía de interleucina 17 y el factor de necrosis tumoral (TNF)

- Células endoteliales

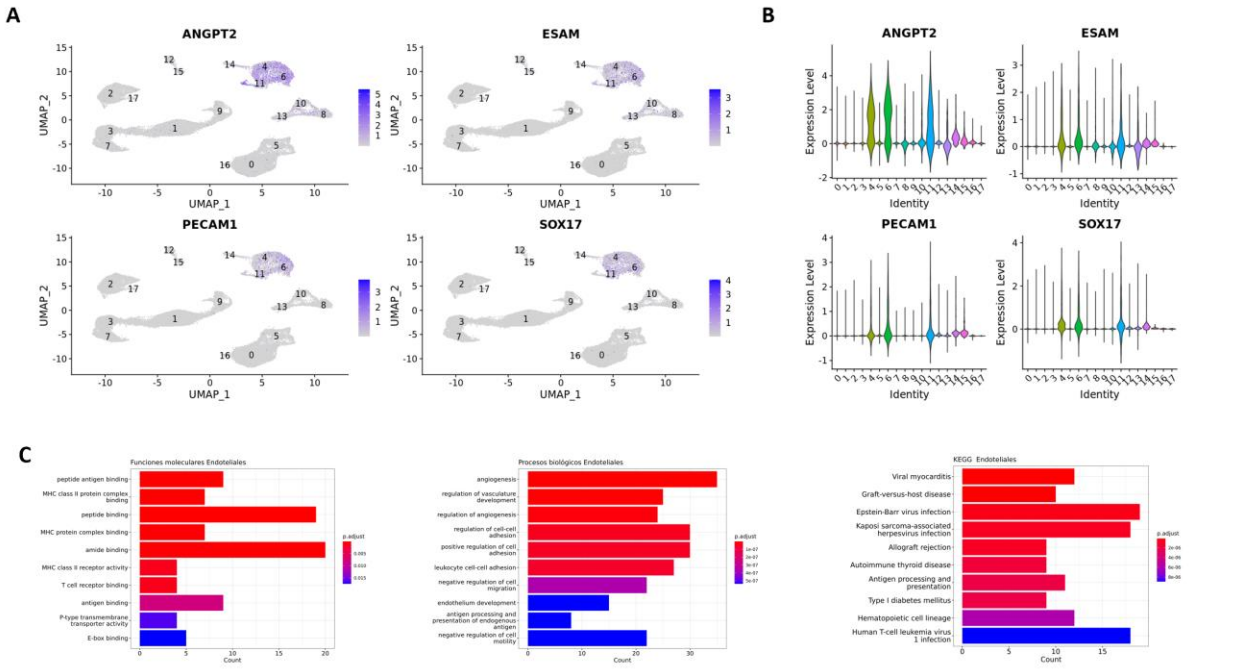


Figura 15.Identificación y caracterización de poblaciones de células endoteliales.
(A)UMAP con los niveles de expresión de los marcadores endoteliales (ANGPT2, ESAM, PECAM1 y SOX17).(B) Distribución de la expresión de los marcadores en los diferentes grupos de células(C) Análisis funcional de los marcadores células endoteliales con avg_logFC mayor a 0.5.

ANGPT2 se expresa en altos niveles de manera homogénea en los grupos 11,4 y 6 mientras ESAM y SOX17, se expresa de manera homogénea, pero en bajo niveles, mientras PECAM1 se expresa en mayoes niveles en 6 y 11.

Se encontraron genes de unión intercelular, el gen de claudina 5 CLDN5, proteína de unión estrecha exclusiva del endotelio, ESAM y el gen GJA1 de la proteína de unión comunicante conexina 43. También el gen de integrina ITGA5 que participa en la unión de las células con la matriz extracelular. Además, genes de adhesión celular relacionadas con la respuesta inmune (ICAM1, PECAM1 y SELE). Por otro lado, se encontraron genes relacionados con la angiogénesis y la coagulación, la angiopoetina 2 ANGPT2, gen de la endoglina, ENG que es un correceptor para ligandos del factor de crecimiento TGFB, INHBB (miembro de la familia de TGFB), FLT1 (miembro del receptor del factor de crecimiento endotelial vascular (VEGFR)) y el factor de transcripción SNAI1 (que regula la transición epitelial-mesenquimal), el receptor de trombina F2RL3 y la Trombomodulina , THBD. También se encontraron genes que participan en la mediación de la respuesta inmune como la interleucina 6 (IL-6) que actúa como citocina proinflamatoria y el receptor de quimiocinas ACKR1

El análisis ontológico encontró que los genes diferencialmente expresados presentan funciones moleculares de unión a péptidos, iones y amidas mientras en los procesos biológicos se encontraron diversos procesos relacionados con la hemostasis, desarrollo y diferenciación endotelial, regulación de la

vasculatura, regulación de la respuesta inmune como presentación de antígenos. Las vías enriquecidas se encuentran relacionadas con procesos inmunes como la presentación de antígenos, enfermedades autoinmunes y relacionadas con virus y vías procesos virales.

- Pericitos

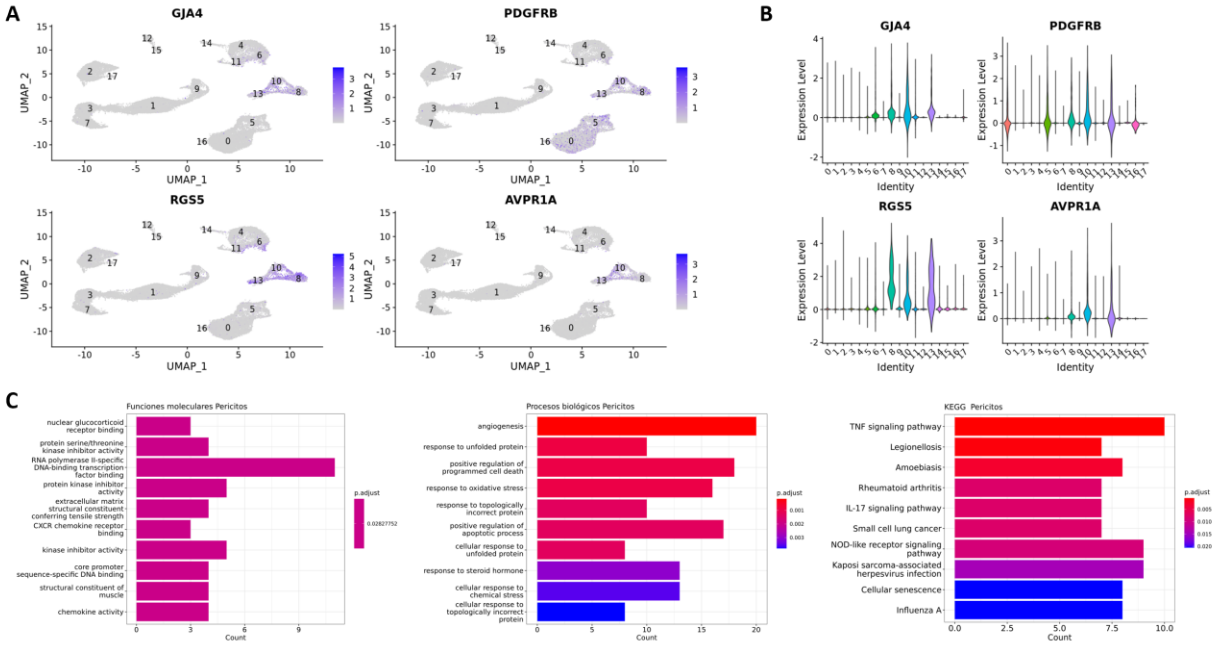


Figura 16. Identificación y caracterización de poblaciones de pericitos.
(A)UMAP con los niveles de expresión de los marcadores de pericitos (GJA4 ,PDGFRB , RGS5 y AVPR1).(B) Distribución de la expresión de los marcadores en los diferentes grupos de células(C) Análisis funcional del grupo positivo para los marcadores para pericitos con avg_logFC mayor a 0.5.

El marcador RGS5 que es el más usado para marcar pericitos se encontró de manera homogénea en 10, 13 y 8, PDGRB que es un marcador común con los fibroblastos se expresa en estos grupos también y en el grupo de los fibroblastos y AVPR1A se expresa mayoritariamente en 13 y 10 y a bajos niveles en 8. Por su parte GJA4 que se encontró como marcador en el atlas de células mamaria, se expresa mayoritariamente en 10 y en menores niveles en 8 y 13.

Se encontraron diferencialmente expresados genes de adhesión celular VASN y GJA4 de la Connexina 37 de unión comunicante, genes de componentes de la matriz celular, colágeno (COL4A1, COL6A2, COL18A1) y laminina (LAMA4). Así como genes de enzimas remodeladoras de la matriz extra celular, las metalotioneínas MT1A y MT1M, las metaloproteinasas ADAMTS1 y ADAMTS4 y un regulador de metaloproteinasa TIMP3. Por otro lado, se encontraron genes relacionados con su actividad contráctil como el componente de miosina MYL9 y la calmodulina CALD1 y genes relacionados con la hemostasis, PROCR (receptor de proteína C) que participa la anticoagulación, EPAS1, un factor de transcripción implicado en la respuesta a la hipoxia y el gen EDNRB receptor implicado en la señalización de endotelina que participa en la regulación del tono vascular, el receptor PDGFB, el gen PDGFRB que participa en el reclutamiento de pericitos en el endotelio.

Los genes enriquecidos en pericitios cumplen la función de unión del factor de transcripción de unión a ADN específica de ARN polimerasa II que están involucradas en los procesos de angiogénesis, y también procesos de respuesta a estresores particulares de tejido, regulación de la apoptosis, estrés oxidativo, acumulación de proteínas mal plegadas.

Células inmunes

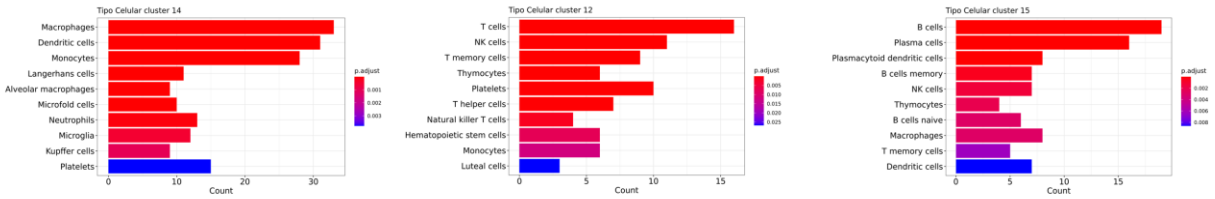


Figura 17.Identificación de los grupos celulares 12,14 y 15.

Análisis de sobrerrepresentación de los marcadores con un avg_logFC mayor a 0.5 de los grupos de células 12,14 y 15.

El grupo 12 mostró enriquecimiento en células T y NK, el 14 en macrófagos y dendríticas y el 15 las células B y plasmáticas.

- Macrófagos

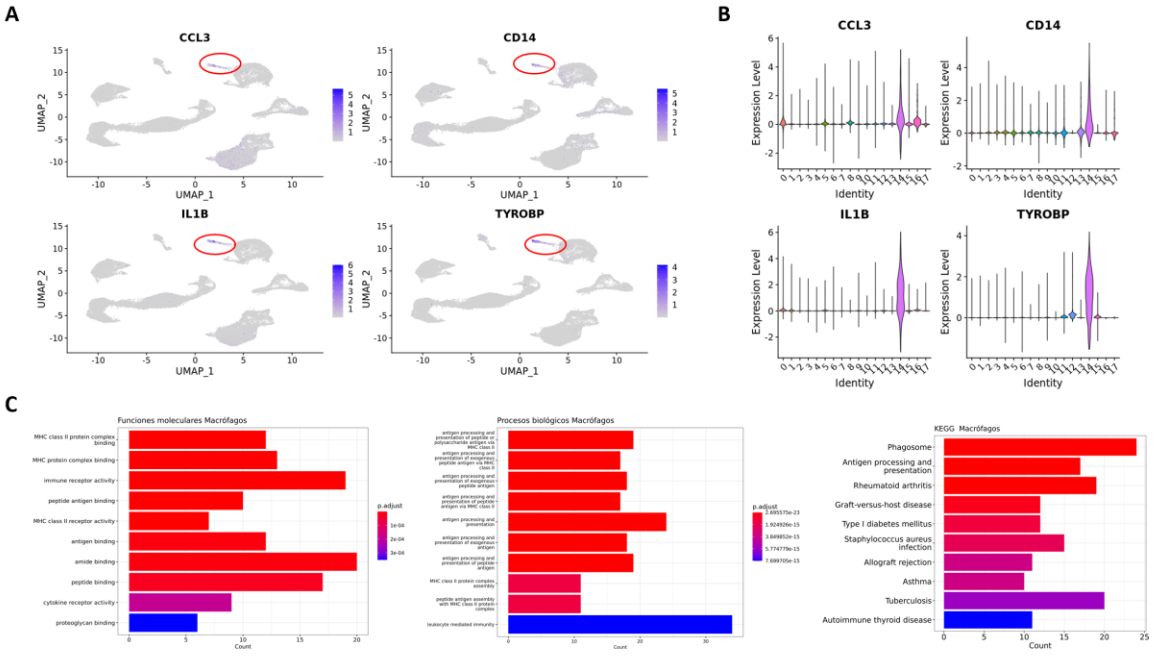


Figura 18. Identificación y caracterización de poblaciones de macrófagos.
(A)UMAP con los niveles de expresión de los marcadores de macrófagos (CCL3, CD14, IL1B y TYROBP).(B) Distribución de la expresión de los marcadores en los diferentes grupos de células(C) Análisis funcional del grupo positivo para los marcadores para macrófagos con avg_logFC mayor a 1**.**

Los marcadores CD14 y TYROBP se distribuyen homogéneamente en el grupo 14, sin embargo, CCL3 IL1B se expresa puntualmente en algunas de las células del grupo, pero de manera específica.

Entre los marcadores se encontraron genes de proteínas involucradas en el reconocimiento de patrones microbianos como CD14 y TLR2.

Genes parte del sistema de complemento, que participan en la activación del macrófago, C3AR1 y C5AR1. genes de parte de la comunicación con las otras células inmunas, interleuquinas, IL1B IL1R2 y citoquinas CXCL8 CXCL16 y receptores de interleuquinas y citoquinas IL1R2 IL10RA CXCR4. Moléculas del complejo mayor de histocompatibilidad de tipo II HLA-DMA, HLA-DPA1, HLA-DPB1, HLA-DQA1 y HLA-DQB1

Genes relacionados con la actividad fagocítica, el gen LYZ que codifica lisozimas, genes de la vesícula fagocítica y endosomas Rab20 y RAB31, RILPL2 y CD68, el citocromo CYBA, genes de cistatina reguladores de catepsinas CSTB, CST3, CTSB, CTSD y CTSL.

Los genes expresados diferencialmente en macrófagos mostraron enriquecimiento para unión a péptidos (los cuales incluyen antígenos) función de unión a antígeno que a su vez están relacionados con los procesos biológicos y vías de respuesta inmune mediados los complejos de histocompatibilidad

(ensamblaje, presentación, procesamiento de moléculas) y mediadas por leucocitos, así como las vías relacionadas con el fagosoma, presentación de antígeno, y de las enfermedad autoinmune artritis reumatoide y la enfermedad bacteriana, tuberculosis

- Linfocitos T

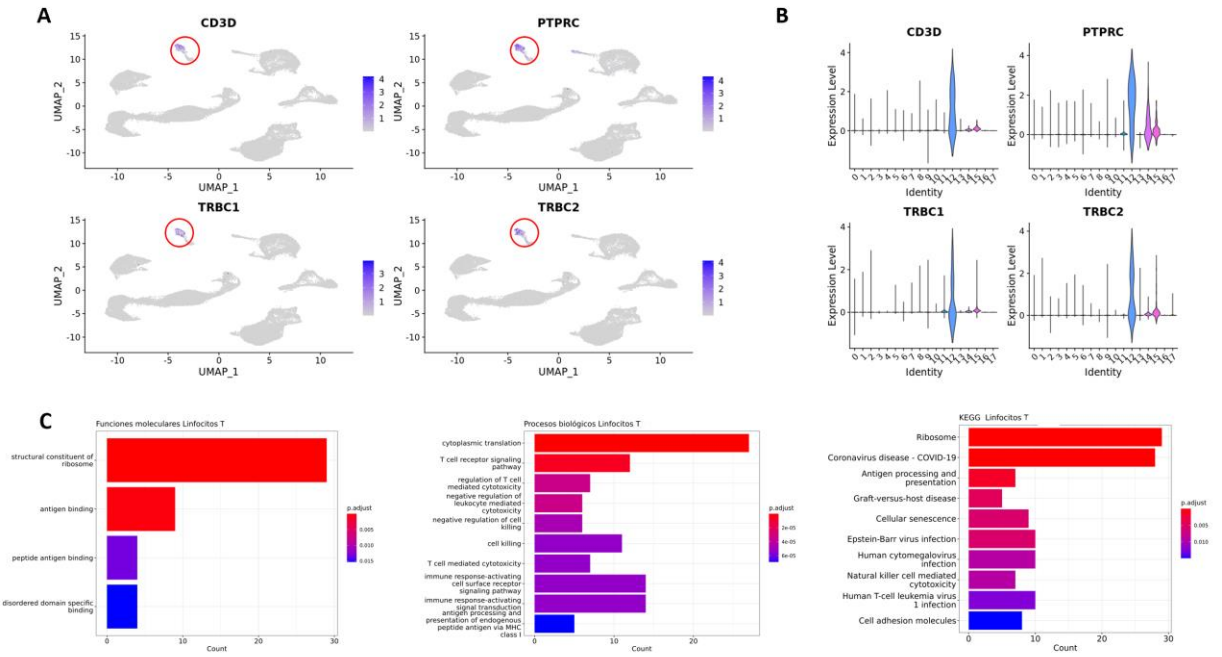


Figura 19. Identificación y caracterización de poblaciones de linfocitos T.
(A)UMAP con los niveles de expresión de los marcadores de linfocitos T (CD3D,PTPRC, TRBC1 Y TRBC2).(B) Distribución de la expresión de los marcadores en los diferentes grupos de células(C) Análisis funcional del grupo positivo para los marcadores para linfocitos T con avg_logFC mayor a 0.5.

PTPRC se expresa homogéneamente en el grupo 12, mientras TRBC1 Y TRBC2 (el cuerpo expresa una mezcla equitativa de linfocitos B que expresa una de estas dos unidades) en conjunto también se expresan homogéneamente en el grupo. CD3D se expresa en parte del grupo 12 aunque de manera específica.

Se encontraron los genes del TCR (TRAC TRBC1 y TRBC2 y CD3A), genes involucrados en la activación de las células T, entre ellos el gen FYN que codifica una proteína tirosina quinasa involucrada en la señalización por el complejo TCR/CD3 y una subunidad de la PI3-quinasa PIK3R1 y el regulador PIK3IP1 así como el gen de PTPRC (CD45), que regula los eventos de fosforilación y el marcador de la activación de linfocitos T, CD69 y el gen CD8A de las células CD8+. Además, al igual que el resto de células inmunitarias, se encontraron genes relacionados con la comunicación como la quimocina CCL5 y los receptores de interleucinas y quimicinas IL7R y CXCR4

Los genes diferencialmente expresados están enriquecidos en la función de unión a antígeno y constituyentes estructurales del ribosoma y en los procesos biológicos y vías de respuesta inmune mediados por células T, Citotoxicidad mediada por linfocitos T, regulación de citotoxidad y respuestas mediadas por unión a antígenos, además de los relacionados con la traducción y ribosoma.

- Células plasmáticas y linfocitos B

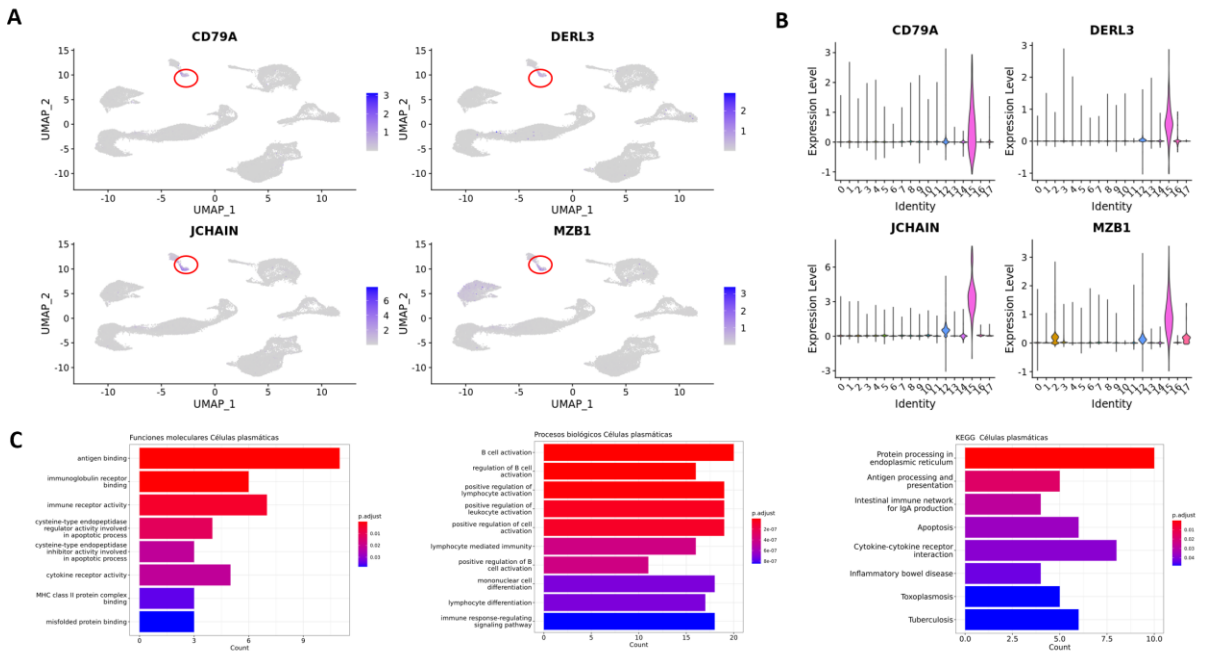


Figura 20. Identificación y caracterización de poblaciones de células plasmáticas.
(A)UMAP con los niveles de expresión de los marcadores de células plasmáticas (CD79A, DERL3, JCHAIN, MZB1).(B) Distribución de la expresión de los marcadores en los diferentes grupos de células(C) Análisis funcional del grupo positivo para los marcadores para células plasmáticas con avg_logFC mayor a 0.5.

Los marcadores de macrófagos y células dendríticas JCHAIN y MZB1 se expresan en el grupo 15 de manera homogénea, mientras CD79A y DERL3 se expresa en parte del grupo pero de manera especifica.

Entre los genes diferenciales se encontraron las inmunoglobulinas IGHA1, IGLC2, IGLC3, IGHA2 y IGKC. Genes relacionados con la síntesis de proteínas, JCHAIN, la proteína chaperona HSPA5. Consistente con su alta tasa de síntesis se encontraron genes relacionados con la homeostasis del retículo endoplasmático, ERN1 que codifica una proteína implicada en la respuesta de proteínas no plegadas (UPR). Además, genes involucrados en la migración y localización en los tejidos el receptor de quimicinas CXCR4 y CCR7 y el gen PRDM1, regulador maestro de la transición de células B a células plasmáticas.

Las células plasmáticas mostraron enriquecimiento en las funciones de unión a antígeno, receptores de inmunoglobinas, complejo mayor de histocompatibilidad y actividad de la endopeptidasa en apoptósis, las cuales están relacionadas con los procesos de activación y regulación de linfocitos B e inmunidad mediada por linfocitos. Así también las vías enriquecidas se relacionaron con su función, como procesamiento en el retículo endoplasmático, procesamiento y presentación de antígenos.

Etiquetado

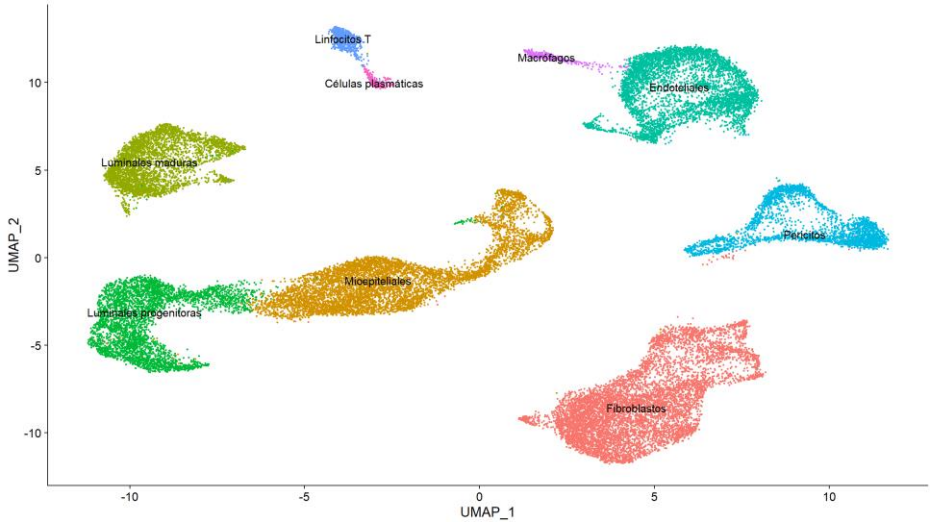


Figura 21.Etiquetado de los grupos celulares obtenidos.

En conjunto tras los análisis anteriores, se llegó a la conclusión que los grupos 2 y 17 corresponden a células luminales maduras, los grupos 2 y 7 a células luminales progenitoras, los grupos 1 y 9 a células basales, los grupos 0,5 y 16 a fibroblastos, los grupos 8,10 y 13 a pericitos, los grupos 4,6 y 11 a células endoteliales, el grupo 14 a macrófagos, el 12 a linfocitos T y el 14 a linfocitos B y células plasmáticas. Quedando 9 de los 17 grupos originales. Este es un etiquetado tentativo basado en el conjunto de los análisis de sobrerrepresentación y de la evaluación de los niveles de expresión de marcadores elegidos para los tipos celulares.

Discusión

Mediante la herramienta Seurat se busca agrupar las células en tipos celulares relevantes biológicamente. Para esto, uno de los pasos principales es la reducción de dimensionalidad que facilita la agrupación de manera que, en lugar de tener que evaluar las células en base a cada gen se evalúen en base a unas pocas características. En Seurat se usa PCA que permite capturar la información de todos los genes en unas cuantas variables, los componentes principales. Cada componente consiste en grupos de genes con una alta covarianza mientras el conjunto de componentes elegidos como características consiste en la combinación de los componentes que tengan la menor correlación entre sí (McCarthy & Hemberg.,2021, Slovin *et al*.,2021). Al visualizar en un heatmap la capacidad individual de cada componente para dividir los datos, se pudo concluir que las firmas capturadas por los componentes del PCA permitían dividir las células en grupos distintos, aunque, como se esperaba, esta capacidad se iba perdiendo a medida que se avanzaba hacia los últimos componentes.

Luego, al visualizar mediante UMAP el conjunto de componentes se concluyó que permitieron la obtención de grupos discretos aptos para ser usados para la identificación. Dado que del PCA solo se obtienen divisiones sin que las células se hayan evaluado por su similitud, el otro paso principal para la obtención de grupos, es el clustering que consiste en la combinación de los algoritmos KNN, SNN (Louvan , McCarthy & Hemberg.,2021, Slovin *et al*.,2021), estos establecen una medida de distancia para establecer que un conjunto de células pertenecen al mismo grupo, se espera que detecte diferencias relevantes entre células (ya sea que pertenezcan a diferentes tipos o estén en un estado biológico diferente) asignándolas a grupos distintos y no divida células en base a ruido de los datos. El desempeño de este paso no se pudo evaluar, ya que tras la integración se obtuvieron 17 grupos los cuales se redujeron a 9 ya que, al usar las firmas de los grupos generados para etiquetar las células, se encontró que muchos grupos que estaban cerca tenían la misma etiqueta, o tenían etiquetas parecidas y al evaluar la distribución de los marcadores celulares muchos abarcaban más de un grupo. En otros casos, se vio que, si bien no se distribuían siempre de manera homogénea en los grupos, tampoco la distribución coincidía con los grupos formados. sin embargo, no se puede concluir que las divisiones formadas se deban a ruido y no a características biológicas. Hay que considerar que la asignación de grupos es parte de un proceso más largo que implica la iteración de pruebas con distintos marcadores celulares que permitan detectar subtipos, puede que la resolución de los marcadores que se usaron para el análisis de sobrerrepresentación y para la evaluación de los niveles de expresión no fuera suficiente. Por otro lado, para determinar si las diferencias si bien no deben a subtipos, se pueden deber a diferentes estados queda por realizar análisis funcional de todos los grupos formados.

Por otro lado, como ya se adelantó, tras estos pasos una vez que se obtienen los grupos de células queda la pregunta si las agrupaciones se hicieron en base a características biológicas relevantes o características relacionadas con la manipulación de los datos. Por un lado, como se en el paso de integración, se puede ver cómo las células se distribuyen en el contexto de la muestra de origen, si la agrupación de las células coincide con su muestra podría significar que la agrupación se hizo en función de características asociadas a la manipulación. Esto se puede complementar, evaluando cómo se distribuyen marcadores específicos de células en los grupos formados. Al evaluar los marcadores específicos de tipos celulares se vio que estos se concentran en ciertos grupos y que los marcadores obtenidos por el análisis de expresión diferencial y el análisis funcional de los grupos obtenidos pudo capturar algunos procesos y características

relacionadas con cada tipo celular. Esto grupos coincidieron con los mayores tipos de células en el tejido mamario encontrados en los atlas de células mamarias Pal *et al*., 2021 y Kumar *et al*., 2023, Bhat-Nakshatri *et al*., 2021 . Entre las células epiteliales se encontraron: células luminales, progenitoras y mioepiteliales. En las células estromales: fibroblastos, células endoteliales y pericitos, y en células inmunitarias, macrófagos, linfocitos T y células plasmáticas. Al igual que en las referencias no se identificaron adipocitos, esta dificultad en las tecnologías scRNA-seq se atribuye al gran tamaño los adipocitos que impide su encapsulación en la plataforma microfluidicas.

Por medio del análisis de genes diferencialmente expresados se pudieron capturar algunas características típicas de cada célula.:

En las células progenitoras se encontraron genes enriquecidos en gran parte, en procesos y vías relacionados con el ribosoma, biogénesis de ribonucleoproteinas, metabolismo de rRNA y la vía de biogénesis de ribosomas, esto, aunque este proceso no está ampliamente documentado en células progenitoras epiteliales en la glándula mamaria, coincide con sus propiedades relativas a células madre en las que la biogénesis y la traducción de los ribosomas han surgido como vías reguladoras que controlan de manera eficiente su homeostasis, (Durand *et al*., 2023). Ya que, aunque residen principalmente en reposo, tras la activación, estas células deben pasar rápidamente de un estado de baja actividad metabólica a una alta actividad que permitan las reacciones anabólicas y el consumo de energía, por lo que aún cuando inactivas, deben mantener un conjunto elevado de ribosomas disponibles para preparar la expresión génica hacia los programas de diferenciación (Gabut, Bourdelais & Durand., 2020, Sharifi da Costa & Bierhoff 2020). Por otro lado, se ha reportado que la vía mTOR participa en la mantención de capacidad proliferativa de las células progenitora, como unidad asociada a ribosomas, el complejo mTOR 2 se ha la biogénesis de los ribosomas con la capacidad proliferativa (tanto en células madre como en células cancerosas) (Ding, Du,Yu,Zhou, Cui & Nie., 2021, Ebrahimi *et al*.,2022)

En las células luminales se encontraron funciones relacionadas con su función de barrera, como las funciones, los procesos y vías de uniones homotípicas de células como las uniones estrechas y adherentes. Pero también se encontraron procesos relacionados con el tejido endotelial esta discrepancia se puede deber a la alta expresión de genes de adhesión celular que es común con las células endoteliales. Se encontraron procesos relacionados con energía y transporte como transporte activo de membrana por oxidorreducción que se da en células que tienen altas demandas de energía y requieren mecanismos de transporte eficientes, la glándula mamaria al ser responsable de producir y secretar leche durante la lactancia, requiere de varios mecanismos de transporte de vesículas para el paso de componentes sintetizados (Ollivier-Bousquet, 2002) y nutrientes desde el torrente sanguíneo a las células epiteliales y sean secretadas en la leche. Al mismo tiempo, estos procesos son comunes con procesos como la termogénesis y se han visto implicados en enfermedades como la esclerosis lateral amiotrófica y alzheimer. Por otro lado, la biosíntesis de trifosfatos de ribonucleósidos de purina, podrían estar implicados en mantener los cambios metabólicos que ocurren en la glándula mamaria para apoyar la síntesis y secreción de leche. Estos genes podrían expresarse en bajos niveles aun en mujeres durante la no lactancia.

En las células basales/mioepiteliales se obtuvieron funciones coincidentes con sus características conocidas como fuente de anclaje a la matriz extracelular como los procesos biológicos relacionados con organización y ensamblaje de célula sustrato así como las vía de adhesión focal y unión matriz extracelular.

Al mismo tiempo se pudo captar características de su actividad contráctil, contracción muscular y sistema muscular genes involucrados en enfermedades relacionado con la musculatura lisa como las cardiomiopatías (Haaksma,Schwartz & Tomasek, 2011, Gieniec & Davis., 2022).

En las células endoteliales se encontraron funciones moleculares relacionadas con las moléculas de adhesión, unión a péptidos, iones y amidas, coincide con que las células endoteliales requieren de una serie de proteínas para la regulación del tono vascular, el intercambio de nutrientes y la respuesta inmune (Alberts *et al*., 2002, Félétou.,2011). En los procesos biológicos y vías hay un sesgo hacia procesos relacionados con su función en la regulación de su respuesta inmune, perdiéndose algunos procesos relacionados con sus funciones como barrera, como las vías y procesos de uniones estrechas y adherentes de unión a matriz extracelular.

Los genes enriquecidos en pericitos cumplen la función de unión del factor de transcripción de unión a ADN específica de ARN polimerasa II que están involucradas en los procesos de angiogénesis, y también procesos de respuesta a estresores particulares de tejido, regulación de la apoptosis, estrés oxidativo, y acumulación de proteínas mal plegadas, están relacionadas a su función de síntesis de componentes de la matriz extracelular al igual que los fibroblastos (Zhao & Chappell., 2019)

En los fibroblastos se capturaron sus funciones y procesos principales de unión, metabolismo y organización de componentes de la matríz extracelular, y como participe de procesos de angiogénesis, proliferación epitelial ya que ambos procesos requieren de un remodelamiento de la matriz extracelular para dar paso al crecimiento de tejido y permitir la migración de células por la matriz extracelular (Yun *et al*.,2010, Inman,Robertson, Mott & Bissell.,2015). Sin embargo, no se capturó su función sintética de proteínas que es característica de fibroblastos quiescentes y activos, esperando procesos procesamiento en el retículo endoplasmático y estrés del retículo y bioseintesis de ribosomas.

En las células inmunitarias se lograron capturar funciones y procesos más relacionados con su función conocida, como unión a antígeno, para los 3 tipos. En el macrófago actividad del fagosoma, unión al complejo de histocompatibilidad tipo II, presentación de unión a antígenos. En las células B/plasmáticas se vieron sus procesos relacionados con diferenciación y actividad del retículo endoplasmático y en las células T procesos relacionados con su actividad citotóxica, unión a antígeno y señalización mediada por la unión a antígeno y traducción y actividades relacionadas al ribosoma que también se relación con sus actividades de síntesis de proteínas señalizadoras y enzimas (Marshall, Warrington,Watson & Kim., 2018).

Una de las limitantes del uso de genes expresados diferencialmente como medio de identificación de los procesos y tipos celulares de los grupos obtenidos fue suficiente para reconstituir los grupos celulares presentes en el tejido mamario y capturar varias de sus características, este método no fue suficiente para captar todos los procesos que estén ocurriendo ya que hay genes que se expresan de igual manera entre los diferentes grupos y que también son importantes para sus funciones, por lo que para complementar este tipo de análisis se pueden usar herramientas como AUCell (Aibar *et al.,* 2017)que utiliza los valores de los niveles de expresión por sí solo, permitiendo identificar la mayoría de los procesos y vías activas.

Conclusión

Con las funciones de la herramienta Seurat y clusterProfiler se pudo agrupar los transcriptomas de las células obtenidas por tecnología single-cell según tipos celulares presentes en la glándula mamaria, obteniendo grupos de células que mostraron marcadores correspondientes a células epiteliales luminales, progenitoras, mioepiteliales, fibroblastos, células endoteliales, pericitos, macrófagos, linfocitos T y células plasmáticas. Las agrupaciones formadas permitieron obtener marcadores que coincidieran con marcadores clásicos de estas células y que capturaran sus características conocidas. sin embargo, no se pudo determinar a qué resolución se encontraron los grupos, por otro lado, hubo características que no se pudieron capturar con la estrategia utilizada. Por lo que el análisis se debe complementar con otras herramientas.

Referencias

Abdulla, M., Traiki, T. B., Vaali-Mohammed, M. A., El-Wetidy, M. S., Alhassan, N., Al-Khayal, K., ... & Ahmad, R. (2022). Targeting MUCL1 protein inhibits cell proliferation and EMT by deregulating β-catenin and increases irinotecan sensitivity in colorectal cancer. International Journal of Oncology, 60(3), 1-12.

Adams, J. (2008) Transcriptome: connecting the genome to gene function. Nature Education 1(1):195

Aibar S, Bravo Gonzalez-Blas C, Moerman T, Huynh-Thu V, Imrichova H, Hulselmans G, Rambow F, Marine J, Geurts P, Aerts J, van den Oord J, Kalender Atak Z, Wouters J, Aerts S (2017). "SCENIC: Single-Cell Regulatory Network Inference And Clustering." Nature Methods, 14, 1083-1086. doi:10.1038/nmeth.4463.

Alberts B, Johnson A, Lewis J, *et al*. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Blood Vessels and Endothelial Cells. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26848/

Alberts B, Johnson A, Lewis J, *et al*. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Cell Junctions. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26857/

Althwaiqeb SA, Bordoni B. Histology, B Cell Lymphocyte. [Updated 2023 May 29]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK560905/

Andrews, T. S., Kiselev, V. Y., McCarthy, D., & Hemberg, M. (2021). Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nature protocols, 16(1), 1-9.

Aranda-Gutierrez, A., & Diaz-Perez, H. M. (2019). Histology, Mammary Glands.

Bhat-Nakshatri, P., Gao, H., Sheng, L., McGuire, P. C., Xuei, X., Wan, J., ... & Nakshatri, H. (2021). A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. Cell Reports Medicine, 2(3).

Bernard, S., Myers, M., Fang, W. B., Zinda, B., Smart, C., Lambert, D., ... & Cheng, N. (2018). CXCL1 derived from mammary fibroblasts promotes progression of mammary lesions to invasive carcinoma through CXCR2 dependent mechanisms. Journal of mammary gland biology and neoplasia, 23, 249-267.

Beyersdorf, N., Kerkau, T., & Hünig, T. (2015). CD28 co-stimulation in T-cell homeostasis: a recent perspective. ImmunoTargets and therapy, 111-122

Biswas, S. K., Banerjee, S., Baker, G. W., Kuo, C. Y., & Chowdhury, I. (2022). The mammary gland: basic structure and molecular signaling during development. International Journal of Molecular Sciences, 23(7), 3883.

Caminero, F., Iqbal, Z., & Tadi, P. (2020). Histology, cytotoxic T cells.

Cavaillon, J. M. (1994). Cytokines and macrophages. Biomedicine & pharmacotherapy, 48(10), 445-453

Chen, B. S. (2021). Systems Immunology and Infection Microbiology. Academic Press.

Chen, Y., Pal, B., Lindeman, G. J., Visvader, J. E., & Smyth, G. K. (2022). R code and downstream analysis objects for the scRNA-seq atlas of normal and tumorigenic human breast tissue. Scientific Data, 9(1), 96.

Colacino, J. A., Azizi, E., Brooks, M. D., Harouaka, R., Fouladdel, S., McDermott, S. P., ... & Wicha, M. S. (2018). Heterogeneity of human breast stem and progenitor cells as revealed by transcriptional profiling. Stem Cell Reports, 10(5), 1596-1609.

Conley, S. J., Bosco, E. E., Tice, D. A., Hollingsworth, R. E., Herbst, R., & Xiao, Z. (2016). HER2 drives Mucin-like 1 to control proliferation in breast cancer cells. Oncogene, 35(32), 4225-4234.

Dabravolski, S. A., Andreeva, E. R., Eremin, I. I., Markin, A. M., Nadelyaeva, I. I., Orekhov, A. N., & Melnichenko, A. A. (2023). The role of pericytes in regulation of innate and adaptive immunity. Biomedicines, 11(2), 600.

Deugnier, M. A., Teulière, J., Faraldo, M. M., Thiery, J. P., & Glukhova, M. A. (2002). The importance of being a myoepithelial cell. Breast Cancer Research, 4(6), 1-7.

Durand, S., Bruelle, M., Bourdelais, F., Bennychen, B., Blin-Gonthier, J., Isaac, C., ... & Gabut, M. (2023). RSL24D1 sustains steady-state ribosome biogenesis and pluripotency translational programs in embryonic stem cells. Nature Communications, 14(1), 356.

Easton, D., & Wilcox, N. (2023). Exome sequencing identifies breast cancer susceptibility genes and defines the contribution of coding variants to breast cancer risk.

Ebrahimi, M., Nourbakhsh, E., Hazara, A. Z., Mirzaei, A., Shafieyari, S., Salehi, A., ... & Barati, G. (2022). PI3K/Akt/mTOR signaling pathway in cancer stem cells. Pathology-Research and Practice, 237, 154010.

Eilken, H. M., Diéguez-Hurtado, R., Schmidt, I., Nakayama, M., Jeong, H. W., Arf, H., ... & Adams, R. H. (2017). Pericytes regulate VEGF-induced endothelial sprouting through VEGFR1. Nature communications, 8(1), 1574.

Fan, D., & Kassiri, Z. (2020). Biology of tissue inhibitor of metalloproteinase 3 (TIMP3), and its therapeutic implications in cardiovascular pathology. Frontiers in Physiology, 11, 661.

Félétou, M. (2011). The endothelium, Part I: Multiple functions of the endothelial cells--focus on endothelium-derived vasoactive mediators.

Gabut, M., Bourdelais, F., & Durand, S. (2020). Ribosome and translational control in stem cells. Cells, 9(2), 497.

Galley, H. F., & Webster, N. R. (2004). Physiology of the endothelium. British journal of anaesthesia, 93(1), 105-113.

Gandhi, D., Molotkov, A., Batourina, E., Schneider, K., Dan, H., Reiley, M., ... & Mendelsohn, C. (2013). Retinoid signaling in progenitors controls specification and regeneration of the urothelium. Developmental cell, 26(5), 469-482.

Garlanda, C., & Jaillon, S. (2016). The interleukin-1 family. In Encyclopedia of Immunobiology. Elsevier. . Así como citoquinas liberadas como tejidos como il7 Wang, C., Kong, L., Kim, S., Lee, S., Oh, S., Jo, S., ... & Kim, T. D. (2022). The role of IL-7 and IL-7R in cancer pathophysiology and immunotherapy. International Journal of Molecular Sciences, 23(18), 10412.

Gieniec, K. A., & Davis, F. M. (2022). Mammary basal cells: Stars of the show. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research, 1869(1), 119159.

Goff, S. L., & Danforth, D. N. (2021). The role of immune cells in breast tissue and immunotherapy for the treatment of breast cancer. Clinical breast cancer, 21(1), e63-e73

Goldstein, L. D., Chen, Y. J. J., Wu, J., Chaudhuri, S., Hsiao, Y. C., Schneider, K., ... & Seshagiri, S. (2019). Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. Communications biology, 2(1), 304.

Griffiths, J. A., Scialdone, A., & Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. Molecular systems biology, 14(4), e8046.

Haghverdi, L., Lun, A. T., Morgan, M. D., & Marioni, J. C. (2017). Correcting batch effects in single-cell RNA sequencing data by matching mutual nearest neighbours. biorxiv, 165118.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W. M., Zheng, S., Butler, A., ... & Satija, R. (2021). Integrated analysis of multimodal single-cell data. Cell, 184(13), 3573-3587.

Haaksma, C. J., Schwartz, R. J., & Tomasek, J. J. (2011). Myoepithelial cell contraction and milk ejection are impaired in mammary glands of mice lacking smooth muscle alpha-actin. Biology of reproduction, 85(1), 13-21.

Hipolito, V. E., Ospina-Escobar, E., & Botelho, R. J. (2018). Lysosome remodelling and adaptation during phagocyte activation. Cellular Microbiology, 20(4), e12824.

Hou, Y., Ding, Y., Du, D., Yu, T., Zhou, W., Cui, Y., & Nie, H. (2021). Airway basal cells mediate hypoxia-induced EMT by increasing ribosome biogenesis. Frontiers in Pharmacology, 12, 783946.

Inman, J. L., Robertson, C., Mott, J. D., & Bissell, M. J. (2015). Mammary gland development: cell fate specification, stem cells and the microenvironment. Development, 142(6), 1028-1042.

Iwaszko, M., Biały, S., & Bogunia-Kubik, K. (2021). Significance of interleukin (IL)-4 and IL-13 in inflammatory arthritis. Cells, 10(11), 3000.

Kumar, T., Nee, K., Wei, R., He, S., Nguyen, Q. H., Bai, S., ... & Navin, N. (2023). A spatially resolved single cell genomic atlas of the adult human breast. bioRxiv, 2023-04.

Kurn, H., & Daly, D. T. (2023). Histology, epithelial cell. In StatPearls [Internet]. StatPearls Publishing.

Jehan, Z. (2019). Chapter 1-Single-Cell Omics: An Overview.

Kato, H. (2002). Regulation of functions of vascular wall cells by tissue factor pathway inhibitor: basic and clinical aspects. Arteriosclerosis, thrombosis, and vascular biology, 22(4), 539-548.

Kumar, T., Nee, K., Wei, R., He, S., Nguyen, Q. H., Bai, S., ... & Navin, N. (2023). A spatially resolved single cell genomic atlas of the adult human breast. bioRxiv, 2023-04

Linderman, G. C. (2021). Dimensionality reduction of single-cell RNA-seq data. RNA Bioinformatics, 331-342.

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular systems biology, 15(6), e8746.

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Müller, M. F., ... & Theis, F. J. (2022). Benchmarking atlas-level data integration in single-cell genomics. Nature methods, 19(1), 41-50.

Mangiola, S., Doyle, M. A., & Papenfuss, A. T. (2021). Interfacing Seurat with the R tidy universe. Bioinformatics, 37(22), 4100-4107.

Marshall, J. S., Warrington, R., Watson, W., & Kim, H. L. (2018). An introduction to immunology and immunopathology. Allergy, Asthma & Clinical Immunology, 14(2), 1-10.

McErlean, C. M., & Brauer-Krisch, E. (2016). Institute of Cancer Research Repository https://publications. icr. ac. uk. Phys. Med. Biol, 61, 320.

Maezawa, Y., Cina, D., & Quaggin, S. E. (2013). Glomerular cell biology. Seldin and Giebisch's The Kidney, 721-755.

Malemud, C. J. (2019). Inhibition of MMPs and ADAM/ADAMTS. Biochemical pharmacology, 165, 33-40.

Meli, V. S., Veerasubramanian, P. K., Atcha, H., Reitz, Z., Downing, T. L., & Liu, W. F. (2019). Biophysical regulation of macrophages in health and disease. Journal of leukocyte biology, 106(2), 283-299.

Mofarrahi, M., & Hussain, S. N. (2011). Expression and functional roles of angiopoietin-2 in skeletal muscles. PloS one, 6(7), e22882.

Nieto Jaramillo, K. A. (2022). Análisis bioinformático del genoma y proteoma humano en diferentes subtipos celulares que conforman el corazón (Bachelor's thesis).

Ollivier-Bousquet, M. (2002). Milk lipid and protein traffic in mammary epithelial cells: joint and independent pathways. Reproduction Nutrition Development, 42(2), 149-162.

Pal, B., Chen, Y., Vaillant, F., Capaldo, B. D., Joyce, R., Song, X., ... & Visvader, J. E. (2021). A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. The EMBO journal, 40(11), e107333.

Pasquini, G., Arias, J. E. R., Schäfer, P., & Busskamp, V. (2021). Automated methods for cell type annotation on scRNA-seq data. Computational and Structural Biotechnology Journal, 19, 961-969.

Piraino, S. W., Thomas, V., O'Donovan, P., & Furney, S. J. (2019). Mutations: Driver versus passenger.

Radtke, D., & Bannard, O. (2019). Expression of the plasma cell transcriptional regulator Blimp-1 by dark zone germinal center B cells during periods of proliferation. Frontiers in immunology, 9, 3106.

Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A., & Teichmann, S. A. (2017). The Human Cell Atlas: from vision to reality. Nature, 550(7677), 451-453.

Ryu, Y., Han, G. H., Jung, E., & Hwang, D. (2023). Integration of single-cell RNA-seq datasets: a review of computational methods. Molecules and Cells, 46(2), 106.

Sharifi, S., da Costa, H. F. R., & Bierhoff, H. (2020). The circuitry between ribosome biogenesis and translation in stem cell function and ageing. Mechanisms of Ageing and Development, 189, 111282.

Sharma, P., Alsharif, S., Fallatah, A., & Chung, B. M. (2019). Intermediate filaments as effectors of cancer development and metastasis: a focus on keratins, vimentin, and nestin. Cells, 8(5), 497.

Shimizu, K., Nakajima, A., Sudo, K., Liu, Y., Mizoroki, A., Ikarashi, T., ... & Iwakura, Y. (2015). IL-1 receptor type 2 suppresses collagen-induced arthritis by inhibiting IL-1 signal on macrophages. The Journal of Immunology, 194(7), 3156-3168.

Schroeder Jr, H. W., Imboden, J. B., & Torres, R. M. (2019). Antigen receptor genes, gene products, and coreceptors. In Clinical Immunology (pp. 55-77). Elsevier.

Scheid, J. F., Eraslan, B., Hudak, A., Brown, E. M., Sergio, D., Delorey, T. M., ... & Xavier, R. J. (2023). Remodeling of colon plasma cell repertoire within ulcerative colitis patients. Journal of Experimental Medicine, 220(4), e20220538.

Slovin, S., Carissimo, A., Panariello, F., Grimaldi, A., Bouché, V., Gambardella, G., & Cacchiarelli, D. (2021). Single-cell RNA sequencing analysis: a step-by-step overview. RNA Bioinformatics, 343-365.

Suzuki, K., Vogelzang, A., & Fagarasan, S. (2019). MZB1 folding and unfolding the role of IgA. Proceedings of the National Academy of Sciences, 116(27), 13163-13165.

Szulc-Dąbrowska, L., Bossowska-Nowicka, M., Struzik, J., & Toka, F. N. (2020). Cathepsins in bacteria-macrophage interaction: defenders or victims of circumstance?. Frontiers in cellular and infection microbiology, 10, 601072.

Tharmapalan, P., Mahendralingam, M., Berman, H. K., & Khokha, R. (2019). Mammary stem cells and progenitors: targeting the roots of breast cancer for prevention. The EMBO journal, 38(14), e100852.

Trzpis, M., McLaughlin, P. M., de Leij, L. M., & Harmsen, M. C. (2007). Epithelial cell adhesion molecule: more than a carcinoma marker and adhesion molecule. The American journal of pathology, 171(2), 386-395.

Venkatesan, S., & Swanton, C. (2016). Tumor evolutionary principles: how intratumor heterogeneity influences cancer treatment and outcome. American Society of Clinical Oncology Educational Book, 36, e141-e149.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., ... & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The innovation, 2(3).

Wu, T., Wang, Y., Jiang, R., Lu, X., & Tian, J. (2017). A pathways-based prediction model for classifying breast cancer subtypes. Oncotarget, 8(35), 58809.

Yeo, J. C., Wall, A. A., Luo, L., & Stow, J. L. (2016). Sequential recruitment of Rab GTPases during early stages of phagocytosis. Cellular logistics, 6(1), e1140615.

Yun, Y. R., Won, J. E., Jeon, E., Lee, S., Kang, W., Jo, H., ... & Kim, H. W. (2010). Fibroblast growth factors: biology, function, and application for tissue regeneration. Journal of tissue engineering, 1(1), 218142.

Zhao, H., & Chappell, J. C. (2019). Microvascular bioengineering: a focus on pericytes. Journal of Biological Engineering, 13(1), 1-12.

Anexo

Características de los tipos celulares

Células epiteliales:

En un primer nivel, tienen función barrera de compartimentalización. Forman una lámina continua que reviste las superficies internas y externas de órganos que permite la protección contra daños físicos, químicos o biológicos y sirve para formar un gradiente para secreción/absorción de sustancias (Kurn & Daly., 2023). En línea con esta función, todas las células epiteliales en su función de barrera tienen dos características en común: la polaridad apico-basal y el alto empaquetamiento celular y ambas se deben, en parte, a la composición de su citoesqueleto y a las uniones intercelulares que establecen, muchos genes de proteínas relacionadas con estas estructuras sirven de marcadores para identificar estas células (Alberts *et al*., 2002).

Su citoesqueleto contiene filamentos intermedios formados de citoqueratina s que les proporcionan un soporte estructural y sirven de anclaje para las uniones intercelular y uniones célula sustrato haciéndolas especialmente resistentes al estrés mecánico. Sus diferentes subtipos se utilizan como marcadores, ya que se ha visto que diferentes tipos de epitelios presentan distintos perfiles de citoqueratina. Por ejemplo, KRT18 y KRT19 se encuentran en células epiteliales luminales, mientras que KRT5 y KRT14 se expresan en células epiteliales basales (Sharma, Alsharif, Fallatah & Chung.,2019).

Por otro lado, las células epiteliales expresan una serie de proteínas de adhesión celular que le permite establecer diferentes tipos de uniones intercelulares entre ellas y con otros tipos de células como las células inmunitarias. Entre las primeras, se encuentran ocludinas y claudinas que participan en las uniones tipo estrechas (les permiten sellar selectivamente el espacio intercelular), proteínas de uniones adherentes y desmosomas que establecen uniones con los citoesqueletos lo que les brindan resistencia mecánica (Alberts *et al*., 2002). En las uniones adherentes moléculas de adhesión llamadas caderinas, se unen a la actina del citoesqueleto. En los desmosomas, se unen los filamentos intermedios de citoqueratina por medio de proteínas desmogleínas, desmocolinas, placoglobina y desmoplaquina se caracterizan por una mejor adhesión siendo particularmente resistentes a estrés mecánico (Alberts *et al*., 2002). Mientras las uniones estrechas y adherentes son compartidas con las células endoteliales, los desmosomas son específicas de las células epiteliales. Otras proteínas de adhesión participan en la unión con otras células y moléculas mediando procesos como la proliferación celular, migración, diferenciación y la señalización celular, respuesta inmunitaria. Un ejemplo es la proteína EPCAM (un marcador clásico) (Trzpis, McLaughlin, de Leij & Harmsen.,2007).

Además de estas características comunes, las células epiteliales pueden estar altamente especializadas y adaptadas para realizar funciones específicas. Por ejemplo, pueden participar en la absorción y secreción (intestino) o facilitar el intercambio de gases (epitelio alveolar en los pulmones). En la glándula mamaria se encuentran dos tipos de células epiteliales, las luminales que recubren la luz de los alvéolos y ductos y las basales/mioepitelial que cubren las luminales y están en contacto directo con la membrana basal. Ambas poblaciones tienen asociadas células madre y progenitoras sirven como fuente para la regeneración de tejido (Deugnier, Teulière, Faraldo, Thiery & Glukhova., 2002)

- **Células luminales**

Tienen la función de producción y secreción de componentes de la leche, como lactosa, lípidos y proteínas (expresando genes de caseina) durante la lactancia así como componentes accesorios que faciliten su liberación. Por lo que tienen una alta actividad de vesículas secretoras en su citoplasma, donde se empaquetan los componentes de la leche antes de la secreción. Fuera del periodo de lactancia expresan a niveles bajos proteínas como PIP (proteína inducida por prolactina) que es inducida por la prolactina y está involucrada en la producción de glóbulos de leche y gotitas de lípidos en las células epiteliales mamarias. Por otro lado, expresa mucinas que forman una barrera física que brinda protección a las células epiteliales que recubren los tractos respiratorio y gastrointestinal y forman las superficies ductales de órganos como el hígado, las mamas, el páncreas y los riñones, la proteína similar a mucina Mucin 1 (MUCL1) es una molécula que se ha visto consistentemente expresada en tejido mamario en estudios bulk RNA-seq y scRNA-seq (Conley, Bosco ,Hollingsworth, Herbst & Xiao.,2016., Bhat-Nakshatri *et al*., 2021. Pal *et al*., 2021, Abdulla., 2022, Kumar et al.,2023).

La actividad de estas células tiene que estar coordinadas con etapas especificas del desarrollo, del ciclo menstrual y embarazo, por lo que son altamente sensibles a las señales hormonales, particularmente al estrógeno y la progesterona, que regulan su crecimiento, diferenciación y producción de leche por lo que expresan diferentes receptores , incluidos los receptores de estrógeno (ER), progesterona (PR), prolactina (PRLR) y andrógenos (AR) así como mediadores de su señalización como el factor de transcripción FOXA1 que interactúan con los receptores de estrógeno y andrógeno para regular la expresión de genes involucrados en el desarrollo, la diferenciación y la producción de leche de las glándulas mamarias (Pal *et al*., 2021., Kumar et al.,2023)

- **Células progenitoras luminales**

Son un subconjunto especializado de células que tienen la capacidad de diferenciarse en células epiteliales luminales formándose repetidamente en la mama adulta y su actividad es necesaria para la homeostasis mamaria normal contribuyendo a la regeneración de tejidos en condiciones de traumatismos, reacciones a patógenos u otras enfermedades asociadas con daño epitelial así como permitir el crecimiento de la glándula mamaria durante la lactancia (Tharmapalan, Mahendralingam,Berman & Khokha., 2019).

Algunos genes relacionados con esta función son los componentes del procesamiento de ácido retinoico que es una molécula de señalización que regula la autorrenovación y la pluripotencia y la especificación en las células madre y otros progenitores, al inducir modificaciones de la cromatina en las regiones reguladoras de los genes (Gandhi *et al*., 2013). Entre ellas el gen ALDH1A3 que participa en la síntesis del ácido retinoico. Se han encontrado otros genes como la metaloproteinasa MMP7 que puede descomponer varios componentes de la matriz extracelular, como el colágeno y la elastina, lo que permite la reestructuración y renovación de los tejidos (Colacino*, et al.,*2018).

- **Células basales**.

Se denominan a las células basales que se ubican adyacentes a la membrana basal y/o al estroma ,sin embargo, en el caso de las glándulas (como las glándulas mamarias, sudoríparas y salivales) los epitelios participan en el transporte de fluidos expresando proteínas similares al músculo liso como actina y miosina lo que, en el caso de la glándula mamaria, les permite contraerse para la expulsión de la leche desde los alvéolos hacia los conductos durante la lactancia. Debido este fenotipo se les denomina células mioepiteliales. Para la regulación de la contracción, expresan receptores hormonas, incluida la oxitocina que media la contracción durante la lactancia (Gieniec & Davis., 2022).

Por otra parte, eedia en gran parte la interacción entre las células luminales y la matriz extracelular sirviendo de anclaje y medio de comunicación, este anclaje se debe a estructuras llamadas hemidesmosomas formadas por proteínas integrinas que unen las citoqueratinas de las células mioepiteliales con las lamininas de la membrana basal (Haaksma,Schwartz & Tomasek, 2011).

Células estromales

Fibroblastos

Son células del tejido conectivo que participan en la mantención y remodelación de la matriz en respuesta a procesos como la reparación, expansión de tejido y la respuesta inmune. Sintetizan y secretan componentes de la matriz extracelular como colágeno, elastina y fibronectina. Expresan genes de proteínas que participan en su remodelación como metaloproteinasas de matriz (MMP) que son endopeptidasas que degradan proteínas de la matriz extracelular, reguladores de metaloproteinasas como los inhibidores tisulares de metaloproteinasas (TIMPs) y para coordinar su actividad con los procesos de reparación y respuesta, tiene una comunicación bidireccional con el epitelio que establece mediante la producción de factores de crecimiento al mismo tiempo que expresan los receptores. Ejemplo de estos son los receptores de factores de crecimiento derivado de plaquetas (PDGFR) y factor de crecimiento de fibroblastos (FGFR) que estimulan la proliferación de fibroblastos y la síntesis de componentes de la matriz extracelular (Inman,Robertson, Mott & Bissell.,2015). Por otro lado, los fibroblastos pueden producir mediadores inflamatorios y contribuir a la inflamación local en respuesta a una lesión o infección interactuando con las células inmunitarias y liberando quimiocinas CXC conocidas por sus propiedades quimioatrayentes de neutrófilos (Bernard *et al*., 2018).

Células endoteliales

Forman una barrera que recubre la superficie interior de los vasos sanguíneos/linfáticos y controlan el paso de sustancias, como nutrientes, gases y glóbulos blancos entre el torrente sanguíneo y los tejidos circundantes. Al igual que en las células epiteliales su función depende de sus uniones intercelulares compartiendo muchas proteínas con ellas. Expresan ocludinas y claudinas, siendo la Claudina-5 (CLDN5) específica del endotelio, proteínas de uniones adherentes como la VE-cadherina (CDH5) y conexinas que forman las uniones comunicantes. Expresan moléculas de adhehsion que participa en la unión con otras células como selectina (codificada por SELE), molécula de adhesión intercelular-I (ICAM-1) y PECAM1 que estás implicado en la adhesión y el tráfico de leucocitos durante la inflamación o el factor de von Willebrand (vWF) que participa en el reclutamiento inicial de plaquetas y la formación de trombos cuando el flujo sanguíneo es elevado, la molécula de adhesión de células vasculares-I (VCAM-1) (Alberts *et al*., 2002, Félétou.,2011).

Las células endoteliales participan activamente en la hemostasis y en las reacciones inmunitarias. Regulan los procesos de angiogénesis a través de la producción de la angiopoyetina 2 (por el gen ANGT2) (Mofarrahi, Hussain.,2011) y factores de crecimiento como el factor de crecimiento B derivado de plaquetas (Pdgfb) y el factor de crecimiento endotelial vascular (Vegfa) (Galley & Webster., 2004, Maezawa, Cina & Quaggin.,2013.). Regulan el tono vascular a través de la producción de óxido nítrico, endotelina y prostaglandinas. Producen y reaccionan a varias citocinas y moléculas de adhesión que participan en las reacciones inflamatorias. Expresan genes para la regulación de la coagulación sanguínea como TFPI (inhibidor de la vía del factor tisular) que codifica una proteína que inhibe la vía del factor tisular,

un paso fundamental en la formación de coágulos sanguíneos (Kato.,2002), la trombomodulina (codificado por THBD) cofactor de la trombina y receptores de trombina.

Pericitos

Los pericitos son células similares a fibroblastos que se encuentran en estrecha asociación con las células endoteliales brindando soporte estructural a los vasos sanguíneos y apoyando su funcionamiento. Regulan procesos como la angiogénesis, reparación del endotelio y el tono vascular por lo que expresan receptores de factores de crecimiento como PDGFRB, VEGFR2 y Ang1 que codifica un factor de crecimiento implicado en el reclutamiento de pericitos y la maduración vascular (Eilken *et al* .,2017). Y como la angiogénesis comienza con la degradación de la membrana basal por células endoteliales activadas que migran y proliferan (Yun *et al*.,2010), su función como regulador de la angiogénesis implica varias proteínas en común con los fibroblastos como componentes de la matriz extracelular (colágeno, laminina y fibronectina (FN1)) y expresa así como genes relacionados con remodelación de la matriz extracelular como metaloproteinasas de matriz y desintegrinas y metaloproteinasas A con motivo de trombospondina (ADAMTS) (Malemud.,2019) e inhibidores de metaloproteinasas de la matriz Inhibidores tisulares de metaloproteinasas (TIMPs) (Fan & Kassiri.,2020). En su función de regulación del tono vascular, tiene propiedad contráctiles por lo que expresan actinas y miosinas. Esto a su vez podría estar mediado por canales de Kv afectar el potencial de membrana de los pericitos y, en consecuencia, su estado contráctil. La actividad de los canales Kv contribuye a establecer el potencial de membrana que determina si los pericitos se contraerán o relajarán, lo que influirá en el tono vascular y el flujo sanguíneo. Además, responden activamente a estímulos proinflamatorios (principalmente IFN-γ, IL-1β y TNF-α) mediante la secreción de diversas citoquinas y quimiocinas (Dabravolski *et al*., 2023).

Células inmunitarias

El tejido mamario normal contiene células de los sistemas inmunitario innato y adaptativo, tanto de linaje mieloide (monocitos, macrófagos, células dendríticas) como linfoide (linfocitos T y linfocitos B) y se localizan predominantemente en los lóbulos en lugar del estroma proporcionando vigilancia inmunológica y eliminación de células epiteliales con cambios mutacionales. Las células inmunitarias predominantes son linfocitos, incluidos linfocitos T CD8+ y CD4+ y macrófagos (Goff & Danforth 2021).

Macrófagos

Los macrófagos son un tipo de glóbulos blancos parte de la inmunidad innata, reconoce de manera inespecífica patrones asociados a moléculas típicas de microorganismos como los lipopolisacáridos que componen las paredes celulares bacterianas. Llevan a cabo distintas funciones, entre ellas, engullir y digerir microorganismos, eliminar de desechos celulares y células muertas, inducir reparación de tejidos y estimular otras células implicadas en la función inmunitaria.

Su actividad es gatillada por la interacción con componentes microbianos (p. ej., lipopolisacárido) por medio del receptor CD14 (receptor de lipopolisacárido,) patrones moleculares asociados a patógenos (PAMP) por medio de receptores tipo Toll (TLR), interacción con citoquinas como interferones, interleucinas (regulan la activación, diferenciación y proliferación en células inmunitarias) y quimocinas (moléculas que promueven la quimiotaxis de leucocitos a los sitios de inflamación), que son liberadas por otras células inmunitarias como linfocitos T auxiliares, linfocitos B y monocitos induciendo la migración a los ganglios linfáticos para la presentación de antígenos y la vigilancia inmunitaria(Cavaillon., 1994,

Garlanda & Jaillon.,2016, Meli, Veerasubramanian, Atcha, Reitz, Downing & Liu., 2019). Así como interaccion con las convertasas del sistema de complemento.

Tras la activación, los macrófagos a su vez sintetizan y liberan una gran variedad de citoquinas (IL-1, IL-6, IL-8, TNFA, MIP-3, MCP-3) que reclutan células inmunitarias como los neutrófilos y las células T al lugar de infección o lesión, amplificando la respuesta inflamatoria (Cavaillon., 1994 ,Iwaszko, Biały & Bogunia-Kubik., 2021) y activan su actividad fagocítica en la que se engulle el patógeno o célula dañada, formando un fagosoma, el cual se fusiona con los lisosomas donde las enzimas lisosomales descomponen el objetivo en fragmentos más pequeños, que luego pueden ser metabolizados por el macrófago para que puedan ser presentadas por sus moléculas MHC de clase II, que participan en la activación de diferentes tipos de células T (Shimizu *et al*., 2015).

Expresa glicoproteínas de membrana asociadas a lisosomas/endosomas (LAMP) como CD68, genes asociados con la maduración del fagosoma, como Rab GTPasas (Rab5, Rab7) que participan desde la incorporación del patógeno hasta la fusión de los fagosomas con los lisosomas (Yeo, Wall, Luo & Stow., 2016). Genes relacionados con la actividad fagocítica, como genes de lisozimas, genes de citocromos componentes de la oxidasa unida a la membrana de los fagocitos que genera superóxido (Hipolito, Ospina-Escobar & Botelho., 2018) y genes de cistatina que regulan la actividad de las catepsinas que son uno de los mayores componentes del sistema proteolítico de las lisozimas (Szulc-Dąbrowska, Bossowska-Nowicka,Struzik & Toka., 2020).

Linfocitos T

Los linfocitos T son glóbulos blancos que participan en la respuesta inmune adaptativa, que se encuentran en la circulación, el bazo y los ganglios linfáticos y desencadenan procesos como inducción de apoptosis en células infectadas o anormales, activación de macrófagos, reclutamiento de neutrófilos y activación de las células B. Reconocen antígenos específicos que sean presentados a por medio de los complejos mayor de histocompatibilidad, y en primera instancia se clasifican según el complejo mayor de histocompatibilidad que reconozcan en: CD8+ (reconocen MHC-I presentes en todas las células nucleadas) o CD4+ (MCH-II presentes en células dendríticas como macrófagos). Al encontrar un antígeno compatible, un linfocito CD8+ o CD4+ se activa para diferenciación. Estas células activadas experimentan una expansión clonal, (lo que da como resultado una población más grande de células T con la misma especificidad antigénica) y se diferencian en efectores (los CD8+ en citotóxicos y los CD4+ en auxiliares), linfocitos de memoria o reguladores. Estos linfocitos diferenciados viajan al lugar de infección en donde se van a activar para la respuesta al encontrar el mismo antígeno y van a efectuar la eliminación de la fuente del antígeno (células infectadas o tumores) (Caminero,Iqbal & Tadi., 2020) .

Para el reconocimiento del antígeno las células T expresan en su superficie el receptor de célula T (TCR) que consta de una cadena alfa y una beta (Codificada por TRA y TRB, respectivamente). A su vez el TCR forma un el complejo con la molécula CD3 y una molécula dimérica conocida como cadena zeta, este complejo transmite las señales al unirse el antígenos por medio de proteínas tirosina quinasas (PTK) de la familia Src, como Lck y Fyn y la inducción de vías PI3K/Akt/mTOR que inducen la actividad de diferentes factores de transcripción celular que determinan la re activación del ciclo celular, inicio de síntesis de citocinas, factores citotoxicos o enzimas líticas (Chen., 2021).

Como se mencionó antes hay dos grandes grupos de células T, cada uno cumple funciones específicas en la respuesta inmunitaria:

Citotóxicos (o linfocitos CD8+): Destruyen directamente a células del cuerpo infectadas con el patógeno. Al ser presentadas con el antígeno por medio MHC-I, los linfocitos secretan las proteínas perforina y granzima. La perforina crea poros en las membranas celulares de los patógenos, permitiendo que las granzimas entren en ellas e induzcan la apoptosis (Caminero,Iqbal & Tadi., 2020).

Auxiliares (o linfocitos CD4+ o helper): su activación depende de MHC-II y una señal coestimuladora mediada por el receptor CD28 que interactúa con proteínas B7 en las células presentadoras de antígenos. Estas proteínas brindan información sobre el antígeno capturado, ya que se sobreexpresan cuando el antígeno corresponde a un agente infeccioso proporcionando un medio de prevenir la autoinmunidad por activación innecesaria de respuestas inmunes (Beyersdorf, Kerkau & Hünig., 2015). Tras la acitvacion, reclutan y activan otras células inmunitarias al secretar citocinas que estimulan la activación de los linfocitos B para que produzcan anticuerpos contra el patógeno, las células T citotóxicas, las células B y macrófago (Caminero,Iqbal & Tadi., 2020).

Linfocitos B y Células plasmáticas

Los linfocitos B son glóbulos blancos parte de la respuesta adaptativa. Se encargan de reconocer antígenos directamente de los patógenos y producir anticuerpos contra ellos, estos no necesitan de una célula presentadora de antígenos a diferencia de los linfocitos T. Cuando una célula B encuentra un antígeno que coincide con su receptor, se activa y pasa por un proceso llamado expansión clonal. Esto da como resultado la producción de células plasmáticas y células B de memoria (Althwaiqeb & Bordoni.,2023)

El receptor de células B está compuesto por una molécula inmunoglobulina (Ig) con una cadena H y una L, y un heterodímero transmembrana llamado Ig-α/Ig-β (codificado por CD79A y CD79B), que transduce las señales (se piensa que de manera análoga al complejo TCR/CD3 en los linfocitos T) por medio de proteínas tirosina quinasas de la familia Src, como Lck, Fyn y Blk y las Syk tirosina quinasas. (Schroeder Jr, Imboden & Torres., 2019).

Su diferenciación de las células B es desencadenado por la interacción con las células presentadoras de antígenos y las células T auxiliares y es mediado en gran parte por el factor de transcripción Blimp-1 (codificado por PRDM1), que induce la expresión de genes productores de anticuerpos y suprime los genes definen el linaje de células B que se requieren para la proliferación/metabolismo (Radtke & Bannard., 2019) y el gen IRF4 (Interferon regulator factor 4) que interviene en la diferenciación y el mantenimiento de las células plasmáticas.

Las células plasmáticas son los efectores de los linfocitos B cuya función es producir y secretar grandes cantidades de anticuerpos llamados inmunoglobulinas.

Por su función productora, las células plasmáticas se caracterizan por una serie de procesos relacionados con la producción de proteínas como una alta actividad de retículo plasmático que les permite sintetizar una gran cantidad de anticuerpo, necesario para una respuesta inmunitaria rápida y eficiente. La síntesis de anticuerpos involucra polimerización de inmunoglobulinas por proteínas como J chain que conecta moléculas de IgM individuales para formar anticuerpos IgM pentaméricos. Por otro lados, las células con una alta tasa de síntesis requieren una mayor capacidad de plegamiento de proteínas, lo que implica la detección del estrés y una mayor expresión de chaperonas de plegamiento de proteínas. En concordancia con esto, las células plasmáticas expresan proteínas como MZB1 y proteínas de la respuesta de proteínas no plegadas (UPR) como DERL3 que está involucrada en la degradación de proteínas mal plegadas o no

ensambladas y proteínas chaperonas contribuye al ensamblaje y plegamiento y de las inmunoglobulinas dentro del retículo endoplásmico antes de su secreción (Suzuki, Vogelzang & Fagarasan.,2019, Scheid *et al*., 2023).