

UNIVERSIDAD AUTONOMA DE MADRID

FACULTAD DE MEDICINA



TRABAJO FIN DE MÁSTER

# CARACTERIZACIÓN TRANSCRIPTÓMICA DEL MICROAMBIENTE TUMORAL Y DEL REMODELADOR DE CROMATINA BPTF EN EL ADENOCARCINOMA DUCTAL PANCREÁTICO

Máster Universitario en Bioinformática y Biología Computacional

Autor: **VALLEJO PALMA, GERMÁN**

Director: **SÁNCHEZ ARÉVALO, VÍCTOR JAVIER**

Tutor: **REDREJO RODRÍGUEZ, MODESTO**

Departamento de Bioquímica

CURSO: 2023-2024

FECHA: Enero, 2024



## AGRADECIMIENTOS

En primer lugar, quiero agradecer a mi tutor, Víctor Javier Sánchez Arévalo, la oportunidad de trabajar en su laboratorio y su guía a lo largo de todo el proyecto. También me gustaría agradecer al resto de compañeros del grupo de investigación su ayuda y el buen ambiente que generan. Además, quiero hacer mención a mi amiga Bea, que me acompañó en la biblioteca durante el proceso de escritura y aguantó todas mis teorías sobre el cáncer. Por último, quiero mostrar mi agradecimiento a mi madre, sin cuyo apoyo habría sido infinitamente más difícil terminar este trabajo.

# ÍNDICE

## 1. RESUMEN

## 2. INTRODUCCIÓN

- 2.1 Adenocarcinoma ductal de páncreas
- 2.2 Importancia y aplicaciones de las NGS en la investigación del ADP

## 3. OBJETIVOS

## 4. MATERIALES Y MÉTODOS

- 4.1 Obtención y procesamiento de conjuntos de datos poblacionales
- 4.2 Obtención y procesamiento de datos de single cell RNA-Seq
- 4.3 Preparación y análisis del experimento de RNA Seq
- 4.4 Análisis de expresión diferencial y análisis funcional
- 4.5 Firmas moleculares
- 4.6 Actividad de vías moleculares y factores de transcripción
- 4.7 Reducción dimensional y clustering
- 4.9 Análisis WGCNA y deconvolución de células inmunes
- 4.10 Obtención y validación de una firma de riesgo
- 4.11 Información complementaria y código

## 5. RESULTADOS

- 5.1 Análisis exploratorio de una cohorte de ADP del TCGA
- 5.2 Caracterización del microambiente tumoral en el ADP mediante sc-RNA-Seq
- 5.3 Efecto a nivel transcriptómico del silenciamiento de BPTF en conjunción con el tratamiento con TNFa
- 5.4 Desarrollo y validación de una firma de riesgo a partir de genes regulados por BPTF

## 6. DISCUSIÓN

## 7. CONCLUSIONES

## 8. BIBLIOGRAFÍA

## 1. RESUMEN

El adenocarcinoma ductal de páncreas (ADP) es una enfermedad de creciente incidencia y baja tasa de supervivencia (11%), debido a las dificultades en su diagnóstico y a la falta de tratamientos efectivos. El microambiente tumoral (TME) es clave en el desarrollo de dicha enfermedad y en su resistencia a la quimioterapia e inmunoterapia. En este proyecto se han usado bases de datos públicas con información transcriptómica y clínica de pacientes de ADP para llevar a cabo una descripción del TME. Se ha aplicado una aproximación multi-ómica con un enfoque dual. Desde un punto de vista poblacional se han analizado datos de RNA-Seq de cohortes de pacientes de ADP mediante herramientas de aprendizaje automático, pipelines bioinformáticos clásicos y análisis de redes. Por otra parte, se alcanzó una resolución de célula única mediante el análisis de datos de sc-RNA-Seq. Gracias a esta caracterización, se asoció la infiltración y activación de células del sistema inmune a una mayor supervivencia. También permitió identificar tipos celulares y patrones de expresión que provocan un estado de inmunosupresión en el TME. De forma adicional se estudió el potencial del gen BPTF como diana terapéutica, comprobando que su silenciamiento causa un arresto proliferativo mediado por los ejes BPTF/MYC y KRAS/PI3K/AKT/MTORC1, que regulan la reprogramación del metabolismo tumoral. Se propone también que el silenciamiento de BPTF puede reducir la respuesta a la inflamación en células tumorales, modulando la expresión de los genes inducidos por TNF $\alpha$ . Además, se emplearon los genes dependientes de BPTF para generar una firma de riesgo con valor pronóstico; entrenada y validada con datos de expresión de cuatro cohortes de pacientes independientes. Esta aproximación también reveló el valor pronóstico de la lisozima y su posible papel como marcador y diana terapéutica en el ADP, que abre la puerta a una nueva línea de investigación.

**Palabras clave:** ADP, páncreas, microambiente tumoral, sc-RNA-Seq, BPTF, RNA-Seq, firma de riesgo.

## 1. ABSTRACT

Pancreatic ductal adenocarcinoma (PDAC) is a disease of increasing incidence and low survival rate (11%), due to the difficulties in its diagnosis and the lack of effective treatments. The tumour microenvironment (TME) is key in the development of this disease and in its resistance to chemotherapy and immunotherapy. In this project we have used public databases with transcriptomic and clinical information of PDAC patients to carry out a description of the TME. A multi-omics strategy with a dual approach has been applied. From a population point of view, RNA-Seq data from cohorts of PDAC

patients have been analysed using machine learning tools, classical bioinformatics pipelines and network analysis. Moreover, single cell resolution was achieved by analysis of sc-RNA-Seq data. Thanks to this characterization, infiltration and activation of immune system cells were associated with increased survival. It also allowed the identification of cell types and expression patterns that lead to a state of immunosuppression in TME. Additionally, the potential of the BPTF gene as a therapeutic target was studied, proving that its silencing causes a proliferative arrest mediated by the BPTF/MYC and KRAS/PI3K/AKT/MTORC1 axes, which regulate the reprogramming of tumour metabolism. It is also proposed that BPTF silencing can reduce the inflammatory response in tumour cells by modulating the expression of TNFa-induced genes. Furthermore, BPTF-dependent genes were used to generate a risk signature with prognostic value; trained and validated with expression data from four independent patient cohorts. This approach also revealed the prognostic value of lysozyme and its possible role as a marker and therapeutic target in PDAC, which opens the door to a new line of research.

**Keywords:** PDAC, pancreas, tumour microenvironment, sc-RNA-Seq, BPTF, RNA-Seq, risk signature.

## 2. INTRODUCCIÓN

### 2.1. Adenocarcinoma ductal de páncreas

El adenocarcinoma ductal de páncreas (ADP) se posiciona como uno de los tipos de cáncer más letales, con una tasa global de supervivencia del 11%. Su incidencia aumenta anualmente en un 1%, situándose como la tercera causa principal de muerte por cáncer en países occidentales, después del cáncer de pulmón y colorrectal [1], [2].

El ADP se caracteriza por un difícil pronóstico debido a su diagnóstico tardío y a la resistencia a los tratamientos actuales de quimioterapia e inmunoterapia [3]. Esta resistencia a la quimioterapia está fuertemente vinculada al denso estroma que envuelve a las células tumorales. Este estroma, compuesto principalmente por fibroblastos asociados al tumor (CAF), representa una barrera que dificulta el acceso del agente quimioterapéutico a su objetivo específico. Por otra parte, estos tumores presentan una significativa infiltración de células del sistema inmunológico con propiedades protumorales, como las células T reguladoras y los macrófagos M2, contribuyendo así a la resistencia [4].

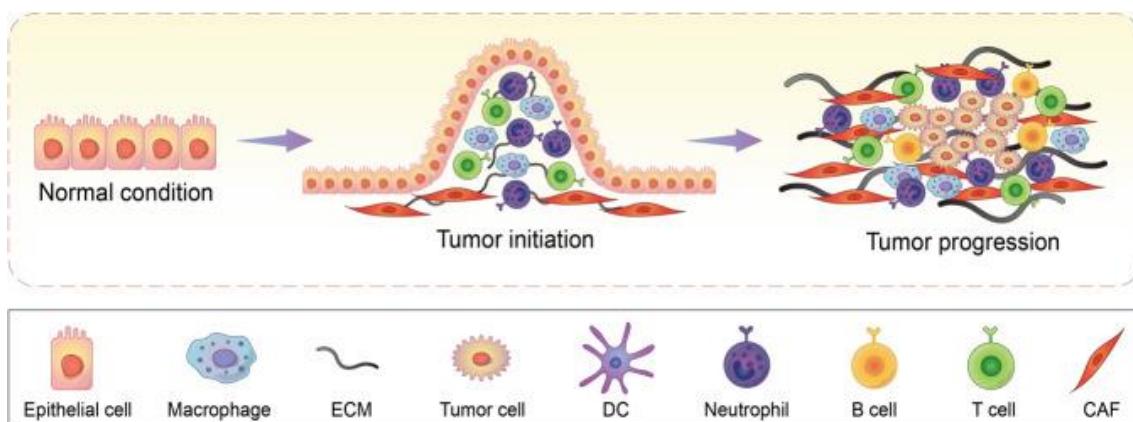
Dentro de los subtipos de cáncer de páncreas, el ADP destaca como el más prevalente. Múltiples lesiones precursoras desencadenan su progresión, siendo las neoplasias

intraepiteliales pancreáticas (PanINs) las más frecuentes, desempeñando un papel fundamental en la progresión hacia el adenocarcinoma invasivo. Además, se distinguen por ser las mejor caracterizadas tanto a nivel molecular como anatómico [5], [6]. En más del 90% de los casos de ADP, la mutación en el gen KRAS se identifica como la mutación iniciadora, dando lugar a las PanIN-1A y 1B en combinación con cambios histológicos específicos. Estas lesiones también se asocian con la inactivación del gen supresor de tumores CDKN2A. La progresión continúa con la inactivación del gen INK4A, dando origen a las lesiones PanIN-2, que presentan displasia y morfología papilar. La última etapa, PanIN-3, se relaciona con la pérdida de los genes supresores de tumores TP53 y SMAD4, exhibiendo un alto grado de displasia y contribuyendo al establecimiento del carcinoma *in situ* [7].

Este modelo, que es común en el adenocarcinoma ductal de páncreas (ADP) derivado de células acinares, también se ha observado en otras lesiones iniciadoras, como las neoplasias mucinosas papilares intraductales (IPMNs) o las neoplasias quísticas mucinosas. Sin embargo, se han identificado variaciones en las vías moleculares involucradas en estas lesiones [8].

#### 2.1.1 Microambiente tumoral del Adenocarcinoma Ductal de Páncreas

Una de las consecuencias del desarrollo de ADP es la inducción de cambio en el estroma que rodea las células cancerosas. Una función clave del estroma de cualquier tejido no transformado es la de proporcionar una respuesta a la aparición de lesiones mediante sus componentes vascular, inmune y conectivo. Estas características son aprovechadas por las células cancerígenas para generar un microambiente tumoral (TME) favorable para su desarrollo [9].



**Figura 1: Representación del microambiente tumoral en la progresión del cáncer de páncreas.** Figura adaptada de Tang et al. (2021) Signal Transduction and Targeted Therapy 6(1):72 [11].

Una de las células más importantes en el TME son los fibroblastos, ya que cuando las células transformadas comienzan a multiplicarse, en las fases iniciales de la neoplasia, los fibroblastos asociados al tumor (CAFs) empiezan a producir proteínas y enzimas remodeladoras de la matriz extracelular de forma anormal, generando así un soporte que facilita la proliferación de las células cancerosas (**Figura 1**) [10].

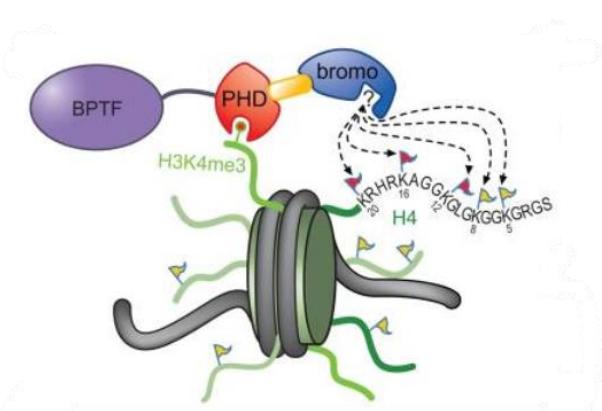
La respuesta efectiva del sistema inmune frente a las células cancerosas involucra células como los linfocitos T CD8+ citotóxicos, que eliminan células malignas, en cuya activación colaboran linfocitos T CD4+ colaboradores (Th1 principalmente) [12]. También son importantes los macrófagos polarizados a M1 (activados clásicamente), que participan en la presentación de antígenos y liberan al medio citoquinas proinflamatorias como el factor de necrosis tumoral alfa (TNF $\alpha$ ), la interleucina-6 (IL6), la interleucina-1 $\beta$  (IL1B) o CXCL1 [13].

Sin embargo, el ADP es un modelo de "tumor frío", un tipo de tumores que se consideran inmunológicamente "excluidos" o "abandonados", lo que significa que el sistema inmunitario no reconoce el tumor como una amenaza o que está siendo suprimido activamente por el microambiente del tumor [11]. En estos tumores, las células del TME liberan citoquinas que favorecen la inmunoevasión, incluyendo interleucinas (IL-8 o IL-10), factor estimulante de colonias de macrófagos (M-CSF) y factor de crecimiento endotelial vascular (VEGF) [14], [15]. Esto es en parte causa y en parte resultado de una infiltración de células del sistema inmune en la que hay un desbalance hacia células inmunosupresoras, como las células T reguladoras o los macrófagos activados alternativamente (M2) [16].

El microambiente tumoral del ADP suele presentar una baja vascularización, lo que, combinado con el rápido crecimiento celular, hace que este sea un ambiente hipódico. Esto promueve la expresión de factores como HIF1, que propician la angiogénesis, la inmunoevasión y la progresión tumoral [17].

### 2.1.2 Papel de BPTF en Adenocarcinoma Ductal de Páncreas

BPTF, o Bromodomain PHD Finger Transcription Factor, ocupa un lugar central en la familia de factores remodeladores de nucleosomas dependientes de ATP (NURF). Esta familia de remodeladores de cromatina despliega dominios lectores (PHD) capaces de reconocer marcas epigenéticas (H3K4me3 y H4K16ac), permitiéndoles remodelar la estructura de la cromatina para facilitar el acceso y reclutamiento de factores de transcripción (**Figura 2**) [18].



**Figura 2: Representación del remodelador de la cromatina BPTF.** El dominio PHD de BPTD reconoce la marca H3K4me3 y el bromodomino realiza las modificaciones epigenéticas. Figura adaptada de Ruthenburg *et al.* (2011). Cell 145(5):692-706 [18].

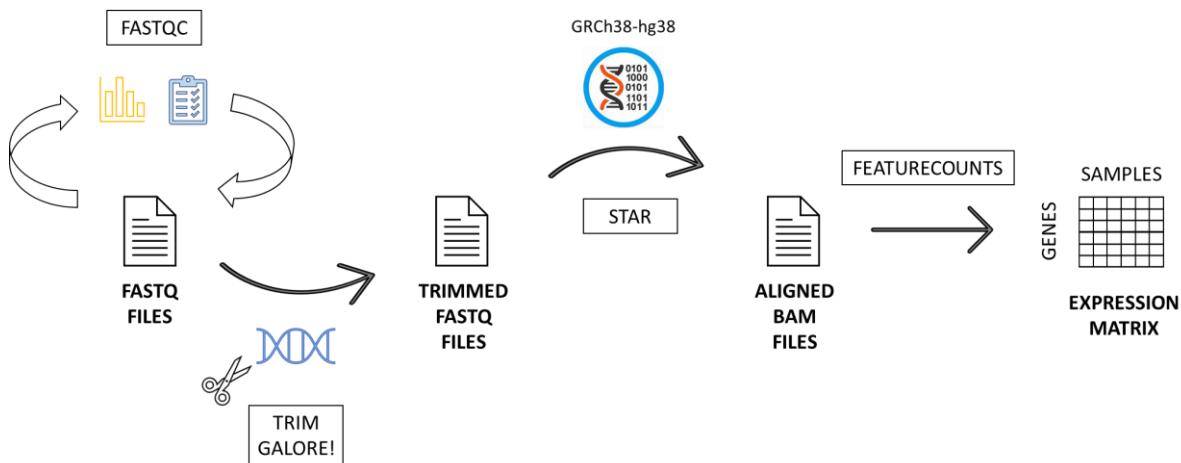
Este remodelador de cromatina es de gran importancia en la actividad transcripcional y tumorogénesis, permitiendo la sobreexpresión de oncogenes como c-MYC [19]. Estudios previos de nuestro laboratorio ya han demostrado que la inhibición de BPTF tiene un impacto en la viabilidad de células tumorales en el cáncer de páncreas, y que potencia el tratamiento por quimioterapia convencional (gemcitabina) [20]. Además, se ha observado que una mayor expresión del eje BPTF/MYC promueve la migración, la proliferación y la glucólisis en el ADP [21].

## 2.4 Importancia y aplicaciones de las NGS en la investigación del ADP.

Comprender los patrones de expresión génica asociados al cáncer es esencial para desentrañar los mecanismos moleculares subyacentes a esta enfermedad. La transcriptómica, que se centra en el estudio de la expresión de todos los transcritos de RNA, es fundamental en este proceso. La información derivada de los patrones de expresión proporciona detalles sobre qué genes están activados o desactivados en un momento dado, revelando así las vías biológicas afectadas y las alteraciones genéticas específicas que contribuyen al desarrollo y progresión del cáncer.

### 2.4.1 La secuenciación del RNA completo, RNA-Seq.

Las técnicas de secuenciación de próxima generación (NGS), como el RNA-Seq, han emergido como herramientas cruciales para desentrañar la complejidad de la expresión génica en el cáncer [22]. Este enfoque no solo caracteriza los perfiles de expresión en diversos tipos de tumores, sino que también facilita la identificación de subtipos moleculares y la predicción de respuestas a tratamientos específicos, contribuyendo así a avanzar hacia una medicina más precisa y personalizada.



**Figura 3: Pipeline de análisis de datos de RNA-Seq.** Se muestra un control de calidad de las lecturas en formato FASTQ mediante FASTQC, después de un trimming con TRIM GALORE, una alineación con STAR de las lecturas a un genoma de referencia, que produce archivos BAM, usados para dar lugar a las matrices de expresión con FEATURECOUNTS.

Desde un punto de vista técnico, el RNA-Seq es un proceso que involucra la conversión de RNA en secuencias de DNA complementarias (cDNA), seguido de la secuenciación masiva paralelizada de estas moléculas [23]. La información generada en este proceso se traduce en archivos de lecturas crudas (FASTQ), los cuales son sometidos a un procesamiento posterior en un pipeline bioinformático.

En un pipeline típico de RNA-Seq (**Figura 3**), se llevan a cabo procesos como el control de la calidad de las lecturas, el alineamiento de secuencias a un genoma de referencia o la cuantificación de la expresión génica. Un resultado común de este pipeline es la obtención de una tabla de recuentos, la cual refleja la abundancia de transcritos para cada gen en cada muestra. El análisis “downstream” típico de RNA-Seq incluye la identificación de genes diferencialmente expresados (DEGs) entre condiciones experimentales o grupos de muestras [24]. Además, se pueden realizar análisis funcionales como los análisis de sobrerepresentación (ORA, del inglés Over Representation Análisis) o los análisis de enriquecimiento de sets de genes (GSEA, del inglés Gene Set Enrichment Análisis) que utilizan firmas moleculares previamente establecidas para identificar los procesos biológicos afectados.

Cuando emergieron las tecnologías de NGS, surgieron también proyectos públicos que comenzaron a recopilar amplios conjuntos de datos de RNA-Seq, vinculados además a variables clínicas y de supervivencia [25]. Para su análisis cobra importancia la aplicación de herramientas avanzadas de bioinformática, como el aprendizaje automático o el análisis de redes. Gracias a la combinación de estas bases de datos y

métodos de análisis se ha posibilitado la identificación de biomarcadores y la correlación entre perfiles de expresión y resultados clínicos [26], ofreciendo una visión más completa de la relación entre la expresión génica y la evolución de enfermedades como el ADP.

En resumen, la tecnología RNA-Seq es un recurso de gran importancia y ampliamente implantado en la investigación biosanitaria actual, que permite, tanto realizar experimentos a pequeña escala, como proyectos de mayor envergadura. A pesar de la utilidad de esta técnica, los rápidos avances en el campo de la transcriptómica han sacado a relucir las limitaciones de esta tecnología, con el surgimiento de nuevas herramientas como la secuenciación a nivel de célula única o sc-RNA-Seq.

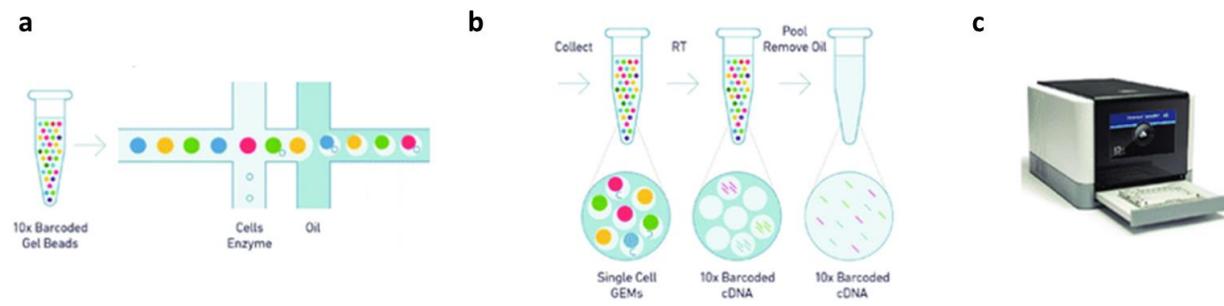
#### 2.4.2 La importancia de la resolución, sc-RNA-Seq.

El sc-RNA-Seq es una moderna técnica de transcriptómica que puede detectar y asignar transcritos a células individuales, lo que permite caracterizar perfiles de expresión génica en tejidos complejos. Desde su introducción en 2009 [27], esta técnica ha supuesto un avance en múltiples áreas de la biología, permitiendo el análisis de tejidos con un detalle sin precedentes, abordando su inherente complejidad.

La heterogeneidad celular en el TME desempeña un papel importante en la progresión del cáncer, por lo que es crucial comprender a fondo los patrones de expresión génica de las distintas clases de células. Métodos como el “bulk” RNA-Seq combinan la información de miles de células de distintos tipos para la secuenciación. Por lo tanto, los clones celulares raros, que pueden desempeñar un papel importante en la progresión tumoral, quedan ocultos [28]. La resolución a nivel de célula única del sc-RNA-Seq supera las limitaciones de las formas comunes de secuenciación de ARN y permite una comprensión más clara de los mecanismos moleculares que promueven la aparición del tumor.

Uno de los métodos más económicos y robustos para generar datos de sc-RNA-Seq es el de 10X Chromium Controller. Esta técnica logra la asociación de transcritos a células individuales mediante la aplicación de un oligómero que etiqueta cada célula de manera específica. En primera instancia, se generan una serie de pequeñas gotas de gel que contienen códigos de barras con una fuerte afinidad por los transcritos [29]. Utilizando canales de microfluidos, cada célula individual de la muestra es aislada junto con una única gota de gel y enzimas listas para romper las células, resultando en cuentas de gel en emulsión (GEMs) que quedan separadas unas de otras en un medio oleoso (**Figura 4.a**). La transcripción inversa tiene lugar dentro de cada gota y los cDNAs con código

de barras se amplifican en masa (**Figura 4.b**). En última instancia, las bibliotecas resultantes se someten a la secuenciación de lectura corta de Illumina (**Figura 4.c**).



**Figura 4: Protocolo de un experimento de sc-RNA-Seq por el método de 10X Chromium Controller.** Figura adaptada de Werba *et al.* (2023) Nature Communications 14(1):797 [30].

La investigación mediante tecnologías de sc-RNA-Seq ha experimentado un aumento significativo en los últimos años, proporcionando nuevas oportunidades y enfoques en el tratamiento clínico del cáncer. Con el desarrollo de la tecnología de secuenciación, la sensibilidad y precisión de la técnica están mejorando gradualmente, a la vez que disminuye su coste. Aparejado a este desarrollo, se ha dado un esfuerzo por parte de la comunidad bioinformática de scverse (<https://scverse.org/>) para generar un ecosistema unificado de análisis de datos procedentes de las tecnologías NGS en Python. Scverse abarca estructuras para contener este tipo de datos y herramientas altamente escalables y eficientes con las que ejecutar desde un pipeline básico de sc-RNA-Seq hasta análisis “downstring” que permitan explotar todo el potencial de este tipo de experimentos [31].

### 3. OBJETIVOS

Los objetivos planteados por el presente proyecto fueron:

1. Descripción del microambiente tumoral mediante datos de RNA-Seq y de supervivencia de pacientes de adenocarcinoma ductal pancreático.
2. Caracterización del microambiente tumoral del adenocarcinoma ductal pancreático mediante un experimento de sc-RNA-Seq.
3. Análisis del efecto a nivel transcriptómico del silenciamiento del gen BPTF.
4. Obtención y validación de una firma de riesgo molecular a partir de genes dependientes de BPTF.

## 4. MATERIALES Y MÉTODOS

### 4.1 Obtención y procesamiento de conjuntos de datos poblacionales

En este estudio se emplearon cuatro conjuntos de datos de tumores primarios de ADP (recolectados y preprocesados por el Dr. Ramón González), siendo tres de ellos obtenidos del repositorio público del Consorcio Internacional del Genoma del Cáncer (ICGC): 1) ICGC-AU2 de microarrays ( $n=267$ ), 2) ICGC-AU de RNA-Seq ( $n=59$ ); y 3) ICGC-CA de RNA-Seq ( $n=178$ ). Bajando los datos de expresión y supervivencia disponibles en el portal del ICGC (<https://dcc.icgc.org/>). El último procede del Atlas del Genoma del Cáncer (TCGA,  $n=141$ ). En este caso, los datos de expresión provienen de Firebrowse (<http://firebrowse.org/>) del Broad Institute, mientras que los datos clínicos y de supervivencia se descargaron del portal "<https://portal.gdc.cancer.gov/>".

Se obtuvieron, en la medida de lo posible, los datos brutos como punto de partida, a excepción del conjunto de microarrays ICGC-AU2, donde los únicos datos de expresión disponibles eran datos normalizados (normalización cuantil) y ya transformados logarítmicamente en base 2. Para las otras dos series del ICGC y la serie de TCGA (de RNA-Seq), se partió de tablas de cuentas “crudas”. Los conteos brutos de expresión de las tres series se convirtieron a transcritos por millón (TPM) y posteriormente fueron transformados logarítmicamente:  $\log_2(\text{TPM} + 1)$ .

### 4.2 Obtención y procesamiento de datos de single cell RNA-Seq

Los datos utilizados para el análisis de sc-RNA-Seq proceden de un conjunto de datos generado y preprocesado por Peng *et al.* (2019) [32], que contiene muestras de 24 pacientes con tumores primarios de ADP y 11 de tejido pancreático normal. Están a disposición pública, bajo el número de acceso GSA: CRA001160. En el reanálisis se partió de la matiz de expresión bruta, disponible en la misma localización.

Para la ejecución del análisis, se implementaron principalmente los paquetes ScanPy (versión 1.9.5) y scvi-tools (versión 0.6.8) de Python. Se aplicó un filtrado de células en función de las cuentas totales y mitocondriales (criterio de desviaciones absolutas de la media). También se predijeron y eliminaron los dobletes. Después se normalizaron los datos a profundidad de 10000 cuentas y se transformaron logarítmicamente. Se seleccionaron los 2000 genes más variables para seguir con el análisis y se integraron las muestras eliminando el efecto de las cuentas totales y de la muestra. Se obtuvo una representación latente (menor dimensionalidad) a partir de la cual se calcularon los vecinos en el grafo generado por una matriz de distancias entre las células, y estos a su vez se usaron para obtener la representación UMAP. La anotación de los tipos celulares

se realizó por transferencia de etiquetas, desde un set de datos de referencia, y se refinó mediante la el clustering (Leiden) y la búsqueda de marcadores conocidos de los tipos celulares

#### **4.3 Preparación y análisis del experimento de RNA Seq**

Dos compañeros del grupo de investigación (María Ferrer y Raúl Muñoz) llevaron a cabo la preparación de 16 muestras, cuatro por cada condición, siendo estas condiciones: shBPTF TNF $\alpha$ - / shBPTF TNF $\alpha$ + / shSCR TNF $\alpha$ - / shSCR TNF $\alpha$ +

Se introdujo un plásmido portador del sistema de shRNA/shSCR para silenciar el gen de BPTF en células humanas de la línea T3M4 (modelo de ADP) mediante infección con vectores virales (lentivirus), producidos previamente en la línea celular empaquetadora HEK293T. Tras su infección con los shRNA contra BPTF (shBPTF) y el control (shSCR), la mitad de las muestras fueron tratadas con el Factor de Necrosis Tumoral Alfa (TNF $\alpha$ ) (R&DSystems) a una concentración de 20 ng/ $\mu$ l 24 horas después de ser plantadas. Tras 24 horas de tratamiento, todas las células en cultivo (de muestras tratadas y no tratadas) fueron levantadas con tripsina (Corning,05322010) y el RNA fue extraído y purificado usando el kit RNeasy Mini Kit (50) (QIAGEN, REF: 74104) de acuerdo con el protocolo. Se cuantificó la concentración de RNA resultante usando el espectrofotómetro NanoDrop (ThermoFisher) y se analizó el RNA integrity number (RIN) para la determinación de la calidad del RNA. Tras esto, las muestras se enviaron a Beijing Genomics Institute's (BGI) donde prepararon librerías por el método de cDNA y se secuenció utilizando la tecnología “DNA nanoball sequencing” (DNB-seq). Como resultado se obtuvieron lecturas pareadas en archivos FASTQ, que fueron el punto de partida para el análisis subsiguiente.

Las lecturas en FASTQ pasaron por un control de calidad (FASTQC, versión 0.12.0), seguido por un timming que eliminó los adaptadores (TRIM GALORE, versión 0.6.8). Después se alinearon frente al genoma de referencia GRch38 mediante STAR (versión 2.7.11a). Los archivos BAM resultantes se usaron para cuantificar la expresión genética mediante featureCounts (algoritmo implementado en el paquete Subread, versión 2.0.6). Se aplicó la herramienta multiQC (versión 1.19) como control de calidad de todos los pasos antes mencionados.

#### 4.4 Análisis de expresión diferencial y análisis funcional.

Para el análisis de expresión diferencial se utilizó la herramienta pyDeseq2 (versión 0.4.4) de Python. En el caso del análisis de sc-RNA-Seq se hizo de dos formas: mediante la función implementada en Scanpy con el método de Wilcoxon, y mediante la generación previa de un pseudo-bulk y la herramienta pyDeseq2. Se seleccionaron como genes diferencialmente expresados (DEGs) aquellos que presentaban  $|Log2 Fold Change| > 1$ , y un p-valor ajustado menor a 0.05.

Para el análisis funcional se ejecutaron análisis ORA (“Over Representation Analysis”), en los que testaba la relevancia funcional de listas de genes de interés. También se llevaron a cabo análisis GSEA (“Gene Set Enrichment Analysis”), ranqueando las matrices de expresión normalizadas mediante el método de la correlación. Para comparar más de dos condiciones se usó el análisis GSVA (Gene Set Variation Analysis). Se consideraron significativas aquellas firmas sobrerepresentadas o enriquecidas con un FDR < 0.05. Estos métodos se ejecutaron en Python mediante el paquete GSEAp (versión 1.1.0).

#### 4.5 Firmas moleculares

En distintas secciones del presente trabajo se usan firmas moleculares ya sea para hacer un análisis funcional o para calcular una puntuación del conjunto de los genes de la firma por célula en el experimento de sc-RNA-Seq (función `sc.tl.score_genes` de Scanpy). Las fuentes de las firmas moleculares utilizadas han sido las siguientes:

- Firmas de polarización de macrófagos a fenotipo M1 o M2, se extrajeron de la literatura [33].
- Firmas de la colección Hallmarks [34], consta de 50 firmas curadas a partir de las numerosas firmas de MSigDB (Molecular Signatures Database) [35].
- Firmas de la base de datos Gene Ontology(GO) [36]).
- Firmas de la base de datos Kyoto Encyclopedia of Genes and Genomes (KEGG) [37]
- Firmas de la base de datos Reactome [38].

Todas las firmas mencionadas, a excepción de las obtenidas en la literatura se obtuvieron a partir del portal de librerías de Enrichr:

➤ <https://maayanlab.cloud/Enrichr/#libraries>

#### **4.6 Actividad de vías moleculares y factores de transcripción**

Se usó el módulo decoupleR (versión 1.5.0) para inferir la actividad de rutas moleculares y factores de transcripción en células únicas (análisis downstring sc-RNA-Seq). Para ello se usaron las funciones run\_mlm y run\_ulp respectivamente junto con matrices con conocimiento previo de interacción de factores de transcripción. Los resultados se representan en UMAPs.

#### **4.7 Reducción dimensional y clústering.**

La reducción dimensional se llevó a cabo por análisis de componentes principales (PCA) implementado en el paquete de Python scikit-learn (versión 1.3.2), previo escalado de los datos (StandardScaler); a excepción del análisis de sc-RNA-Seq, en el que se usó el UMAP. Para el clústering se usaron las dos primeras componentes de los datos como entrada y el método gaussiano implementado en scikit-learn (GaussianMixture).

#### **4.9 Análisis WGCNA y deconvolución de células inmunes**

La deconvolución de la infiltración inmune del set de datos TCGA se realizó mediante dos métodos (SingScore y DeconRNASeq) del paquete TumorDecon (versión 1.1.1), en Python. Se compararán la infiltración de cada tipo celular entre los dos clústeres antes calculados con un t-test (o Wilcoxon, en caso de no normalidad), representando las diferencias significativas ( $p < 0.05$ ) en boxplots y heatmaps, ordenando los datos por clustering jerárquico e incluyendo la pertenencia al clúster en estos últimos.

También realizamos un análisis de co-expresión utilizando el módulo PyWGCNA (versión 1.20.4) (**Figura 1, anexo II**), para agrupar módulos de genes con patrones de expresión similares. Se realizó un preprocessamiento previo, clasificando las muestras con un clústering jerárquico para detectar outliers. Después, se seleccionó el “soft power” de  $\beta = 10$  ( $R^2$  libre de escala = 0.9) y se construyó la matriz de adyacencia. A continuación, los genes con patrones de coexpresión similares en las muestras se agruparon jerárquicamente en diferentes módulos. Los módulos se caracterizaron mediante un ORA. Finalmente, se calculó la correlación entre cada módulo y las variables clínicas de las muestras (además de la pertenencia al Clúster).

#### **4.10 Obtención y validación de una firma de riesgo**

Para la construcción de la firma de riesgo, se utilizó un conjunto de datos de entrenamiento (ICGC-AU2) y tres conjuntos de datos de validación (ICGC-AU, ICGC-CA y TCGA). De estos se utilizó tan solo la expresión de los genes dependientes de BPTF en tratamiento con TNF $\alpha$  (DEGs). En el dataset de entrenamiento se realizó una regresión Cox Lasso (ajustando el parámetro alfa por validación cruzada). Los genes con coeficiente distinto de 0 se usaron para una nueva regresión Cox (no regularizada) y se definió la firma de riesgo con los genes significativos ( $p<0.05$ ) y sus coeficientes. Se calculó una puntuación de riesgo para cada paciente, mediante la combinación lineal de los valores de expresión de los genes multiplicados por sus coeficientes, y se estratificaron los sets de datos por el valor de la mediana de esta puntuación (alto y bajo riesgo). Se ajustaron curvas de Kaplan-Meier y se comprobaron las diferencias significativas ( $p<0.05$ ) de supervivencia entre ambos grupos (test de rangos logarítmicos). Esto se hizo para el set de entrenamiento y los de validación.

Los genes de la firma de riesgo se estudiaron de forma individualizada, usando su valor de expresión para estratificar los datasets y comprobar la diferencia en la supervivencia. Aquellos que tenían valor pronóstico en al menos 3 de los 4 sets de datos, se estudiaron más a fondo empleando los datos de sc-RNA-Seq para comprobar su expresión en el microambiente tumoral. Por último, si aparecían en el tejido tumoral, se comprobaba la expresión diferencial del gen en tumores de ADP frente a tejido pancreático normal mediante una búsqueda en GEPIA (<http://gepia.cancer-pku.cn/>)

#### **4.11 Información complementaria y código**

El presente documento viene asociado a dos anexos, en el primero se hace una explicación detallada de los materiales y métodos, en la que se citan todos los paquetes informáticos utilizados. En el segundo se añaden figuras complementarias. Además, todo el código desarrollado para la obtención de los resultados de este proyecto se encuentra disponible en GitHub:

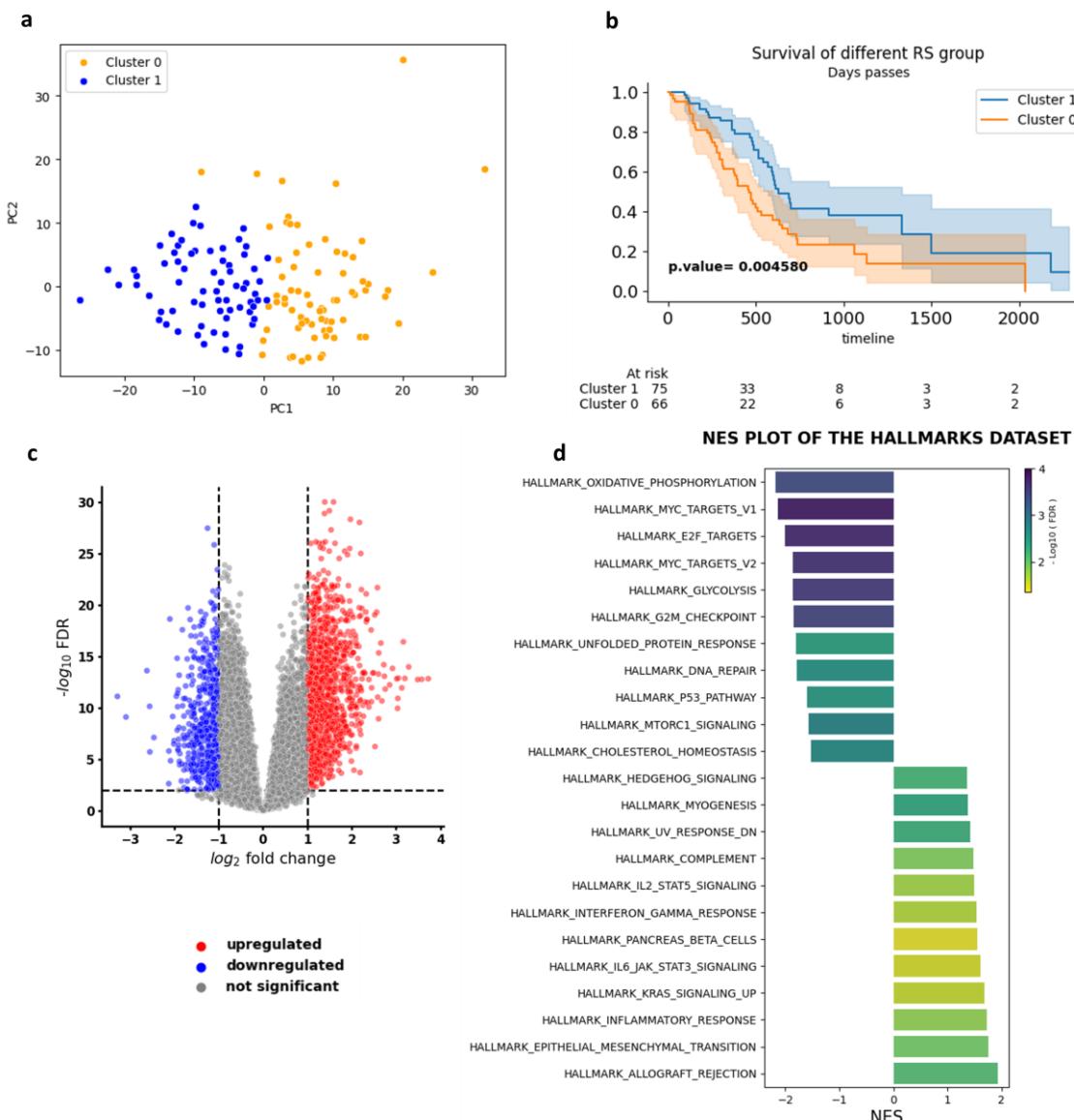
➤ [https://github.com/GERMAN00VP/TFM\\_GERMAN\\_VALLEJO](https://github.com/GERMAN00VP/TFM_GERMAN_VALLEJO)

### **5. RESULTADOS**

#### **5.1 Análisis exploratorio de una cohorte de ADP del TCGA**

A partir de la cohorte del TCGA (mencionada en el apartado 4.1 de este trabajo), el profesor Luis Bote, en calidad de colaborador del grupo de investigación, realizó una

clasificación no supervisada (apartado 4.7). Así se encontraron dos grupos de pacientes, en adelante clúster 0 y clúster 1 (**Figura 5.a**), cuyas curvas de supervivencia mostraron que este último era un grupo de menor riesgo (**Figura 5.b**)



**Figura 5: Clusterización, curvas de Kaplan-Meier y análisis transcriptómico de la cohorte del TCGA.** Muestras del data set TCGA (n=141) por sus dos primeras componentes principales, coloreadas por la pertenencia al clúster (**5.a**). Curvas de Kaplan-Meier de ambos clústeres del TCGA, p-valor de la prueba de rangos logarítmicos (**5.b**). Volcano plot, mostrando genes diferencialmente expresados ( $|Log2\text{ Fold Change}|>1$  &  $FDR < 0.05$ ) (**5.c**). Puntuación de enriquecimiento normalizada para las firmas moleculares “hallmarks” coloreadas por  $-\text{Log}10\text{ FDR}$ , se muestran solo resultados significativos ( $FDR<0.05$ ) (**5.d**).

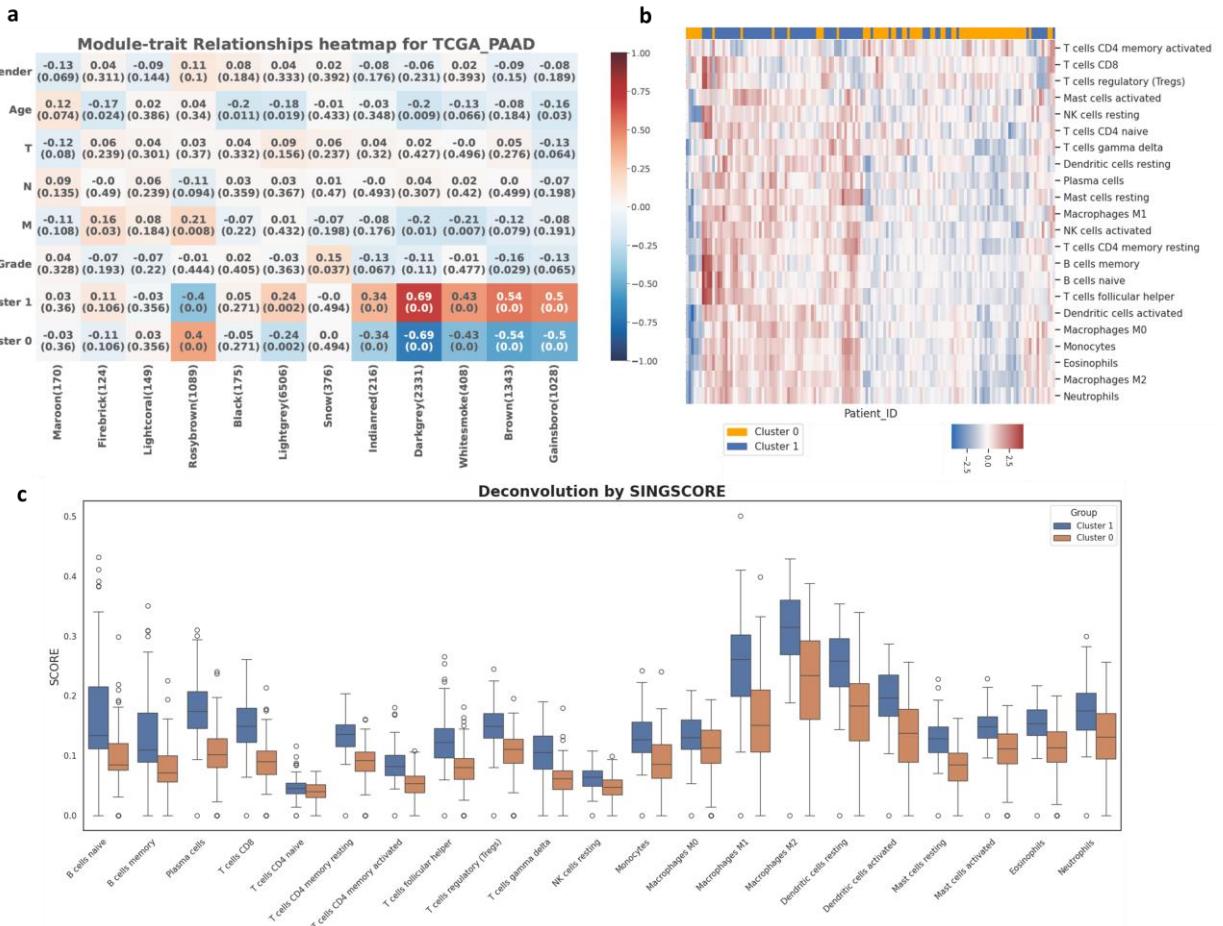
Para entender en qué radicaba la diferencia entre ambos grupos de pacientes se llevó a cabo un análisis de expresión diferencial, comparando la condición clúster 1 contra clúster 0. De esta forma hallamos un total de 1797 genes diferencialmente expresados (DEGs), 1224 de los cuales estaban sobreexpresados, como se puede observar en la

**Figura 5.c.** Para entender el significado funcional de esta diferencia en la expresión genética entre condiciones, se realizó un análisis funcional GSEA y se testando la colección Hallmarks. Este análisis revela que en el grupo de pacientes de alto riesgo (clúster 0) aparecen enriquecidas significativamente firmas relacionadas con proliferación celular y la progresión tumoral, como “G2M checkpoint” o “MYC targets” entre otras. Mientras que el grupo de bajo riesgo guarda relación con firmas de respuesta inmunológica como “interferon gamma response” o “inflammatory response” (**Figura 5.d**). Cabe destacar que en este grupo también aparecen enriquecidas firmas relacionadas con la transición epitelio mesénquima (“epitelial mesenchimal transition”) y vías de señalización como “IL6-JAK-STAT3 signaling” o “IL2-STAT5 signaling”.

Siguiendo con el análisis de esta cohorte, se aplicó el método de WGCNA aplicando el preprocesado explicado en el apartado 4.9 de este trabajo y obteniendo una red libre de escala subdividida en 12 módulos, como se muestra en las **Figuras 1.a, 1.b y 1.c del anexo II**. Al estudiar la correlación de estos módulos con las variables clínicas y con la pertenencia a uno u otro clúster, se observó que por lo general los módulos guardaban mayor correlación con los clústeres que con otras variables (**Figura 6.a**).

Destacaron dos módulos (“brown” y “darkgrey”), por tener mayor correlación con el clúster de mayor supervivencia. Para entender en qué procesos estaban involucrados los genes de estos módulos se hizo un análisis funcional ORA, testando las firmas de Hallmarks, GO, KEGG y Reactome. Esto mostró que el módulo “brown” tenía relación con firmas de respuesta inmune e inflamatoria, mientras que el módulo “darkgrey” guardaba relación con procesos de organización de la matriz extracelular (**Figuras 2.a y 2.b, anexo II**).

Por último, se llevó a cabo una deconvolución del perfil de infiltración de células del sistema inmune en los pacientes de esta cohorte usando los dos métodos explicados en el apartado 4.9 (SingScore y DeconRNASeq). En la **Figura 6.b** (y **2.d del anexo II**) se aprecia como las muestras pertenecientes a los mismos clústeres también se clasifican juntas (de forma aproximada) en base a su perfil de infiltración inmune. Además, todos los tipos de células inmunes que presentaban diferencias significativas tenían una mayor puntuación (y frecuencia) en el clúster de mayor supervivencia (**Figura 6.c y Figura 2.c del anexo II**).



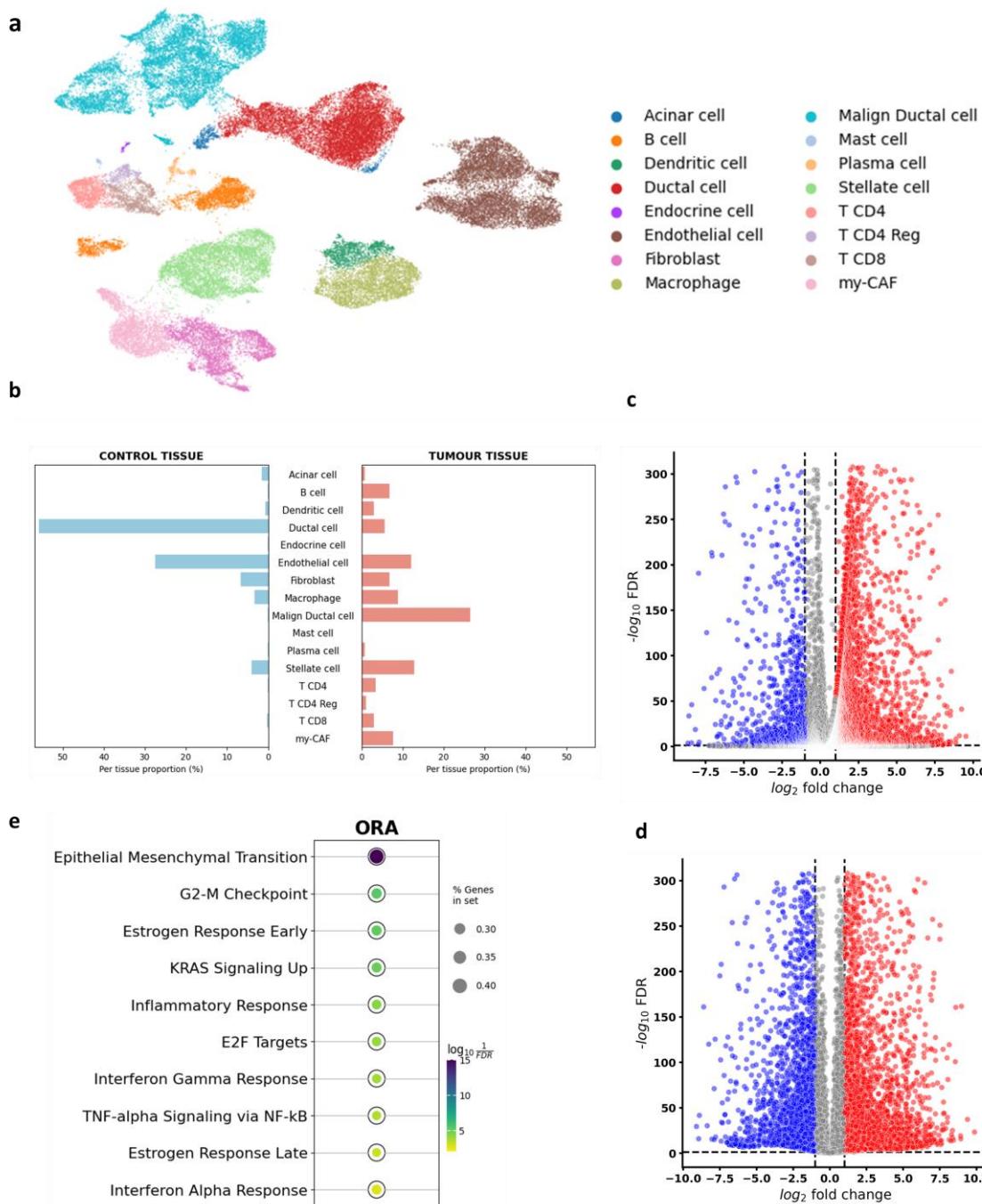
**Figura 6: Análisis WGCNA y deconvolución de la infiltración inmune de la cohorte del TCGA.** Matriz de correlación de módulos de genes WGCNA con variables clínicas y clusters, sobre el gráfico se anota el valor de correlación y la significancia (entre paréntesis, **6.a**). Heatmap mostrando la puntuación de SingScore para los distintos tipos celulares, ordenando en base a un clustering jerárquico (en la parte superior se muestra la pertenencia al clúster gaussiano, **6.b**). Boxplot mostrando diferencias significativas (t-test,  $p<0.05$ ) en las puntuaciones para cada tipo celular en el clúster 1 ( $n=75$ ) frente al clúster 0 ( $n=66$ ) (**6.c**).

## 5.2 Caracterización del microambiente tumoral en el ADP mediante sc-RNA-Seq

Con el propósito de estudiar el microambiente tumoral en profundidad, se reanalizó el experimento de sc-RNA-Seq (con 24 muestras de ADP y 11 controles) mencionado en el apartado 4.2 de este trabajo (se muestran hitos intermedios de este procedimiento en la **Figura 3, anexo II**).

Gracias a este análisis se obtuvieron 49966 células, lo que permitió una caracterización detallada de la heterogeneidad transcriptómica presente en la muestra y facilitó la identificación y clasificación precisa de 16 poblaciones celulares, incluyendo las células ductales tumorales (**Figura 7.a**). Como se puede apreciar en la **Figura 7.b**, la diversidad

y heterogeneidad de tipos celulares en el microambiente tumoral superan a las observadas en el tejido control.



**Figura 7: Caracterización del microambiente tumoral en el ADP mediante sc-RNA-Seq.**  
 UMAP mostrando la anotación de tipos celulares (7.a). Gráfico de barras, se muestra el porcentaje de células de cada tipo celular por tejido (tumoral o control, 7.b). Volcano plots mostrando genes diferencialmente expresados ( $|\log_2 \text{Fold Change}| > 1$  &  $\text{FDR} < 0.05$ ) con método de Scanpy (7.c), o pseudobulk (7.d). Firmas sobrerepresentadas en análisis ORA (firmas “hallmarks”), de los DEGs comunes, el tamaño del punto es el número de genes encontrados en el set y el color representa la significancia ( $-\log_{10}(\text{FDR})$ ) (7.e).

Las células ductales predominan en el tejido sano, a diferencia del tejido tumoral, donde las células ductales (no malignas) representan una pequeña fracción de las células totales. Las células endoteliales también aparecen en mayor proporción en el tejido normal, sin embargo, un análisis funcional de los genes sobreexpresados en tejido endotelial tumoral muestra sobrerepresentación de firmas como “regulation of angiogenesis” y “positive regulation of cell population proliferation” (**Figura 4.a, anexo II**), mientras que los genes regulados a la baja se relacionaban con firmas como “regulation of lymphocyte differentiation” y “regulation of T cell activation” (**Figura 4.b, anexo II**). Uno de los genes sobreexpresados en células endoteliales tumorales fue HIF1A, cuya expresión se muestra en la **Figura 4.c del anexo II**. En menores proporciones encontramos fibroblastos, células estrelladas y macrófagos; siendo las células acinares y dendríticas las de menor proporción en tejido normal.

En cambio, en el tejido tumoral, las células ductales malignas son predominantes, y se observa una mayor infiltración de células inmunes, incluyendo células B, células plasmáticas, T CD8, TCD4 y T CD4 reguladoras. Además, se identificaron células my-CAF (miofibroblastos asociados a cáncer), exclusivas del tejido tumoral. También se detectaron mayores niveles de macrófagos en el tejido tumoral, sin embargo, no fue posible encontrar subclústeres que expresaran únicamente marcadores típicos de M1 o M2 (CD86 y CD163 respectivamente). Sin embargo, al puntuar firmas específicas para macrófagos M1 y M2 se observa un espectro de polarización desde macrófagos poco polarizados (a la izquierda del UMAP), pasando por una región que concuerda más con el fenotipo M1 y finalizando con una zona (a la derecha del UMAP) que concuerda con una polarización hacia M2 (**Figura 5.a, anexo II**).

En cambio, en el tejido tumoral, las células ductales malignas son predominantes, y se observa una mayor infiltración de células inmunes, incluyendo células B, células plasmáticas, T CD8, TCD4 y T CD4 reguladoras. Además, se identificaron células my-CAF (miofibroblastos asociados a cáncer), exclusivas del tejido tumoral. También se detectaron mayores niveles de macrófagos en el tejido tumoral, sin embargo, no fue posible encontrar subclústeres que expresaran únicamente marcadores típicos de M1 o M2 (CD86 y CD163 respectivamente). Sin embargo, al puntuar firmas específicas para macrófagos M1 y M2 se observa un espectro de polarización desde macrófagos poco polarizados (a la izquierda del UMAP), pasando por una región que concuerda más con el fenotipo M1 y finalizando con una zona (a la derecha del UMAP) que concuerda con una polarización hacia M2 (**Figura 5.a, anexo II**).

En una segunda fase del análisis, nos centramos en las células ductales y células ductales malignas (UMAP ampliado de estos tipos celulares en la **Figura 5.b, anexo II**).

Para caracterizar los cambios transcripcionales que ocurren tras la transformación en células malignas, se llevó a cabo un análisis de expresión diferencial. Se aplicaron dos métodos según se describe en el apartado 4.4 del presente trabajo. El método de Wilcoxon por Scanpy permitió encontrar 5869 DEGs (**Figura 7.c**). El segundo método, basado en la construcción de un pseudobulk, reveló 5698 DEGs (**Figura 7.d**). Para consolidar los resultados de ambos análisis, se tomó la intersección entre ambos conjuntos de genes, obteniendo 3268 DEGs, de los cuales 2354 estaban regulados al alza y 914 a la baja.

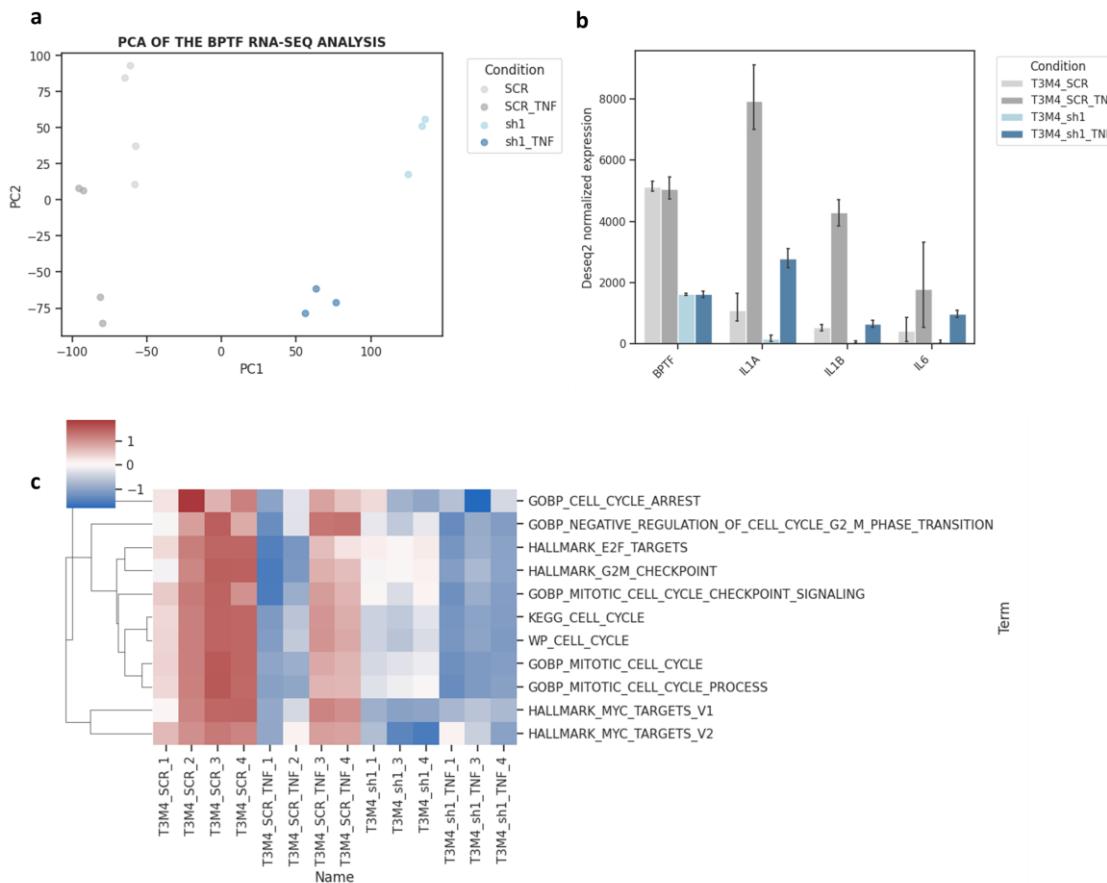
Posteriormente, se realizó un análisis funcional de ORA con la lista de DEGs y testando las firmas de Hallmarks. Esto reveló que procesos como la transición epitelio-mesénquima, la interacción con la matriz extracelular, la síntesis de colágeno, las vías de respuesta a estrógenos, la diabetes tipo II, la señalización de TNF-alfa por la vía de NFKB y la activación de la ruta de KRAS (**Figura 7.e**) aparecían significativamente sobrerepresentados.

Finalmente, se estimó de la actividad de distintas rutas moleculares y factores de transcripción en las células ductales y cancerosas a nivel de célula única. Este análisis reveló que ciertas rutas presentan mayor actividad en las células ductales malignas, entre las que encontramos TNF alfa, EGFR, Hipoxia, JAK/STAT, MAPK, TGF beta, PI3K o VEGF como se muestra en los UMAPs de la **Figura 5.c del anexo II**. Por otra parte, también se detectaron factores de transcripción con mayor actividad en células ductales malignas frente a las ductales normales; como JUN, FOS, AP1, NFKB1, RELA, o STAT3 (**anexo II, Figura 5.d**).

### **5.3 Efecto a nivel transcriptómico del silenciamiento de BPTF en conjunción con el tratamiento con TNF $\alpha$**

Para evaluar las disparidades en la expresión de BPTF en pacientes, empleamos el portal GEPIA, el cual alberga datos transcriptómicos de la cohorte de pacientes del TCGA. Llevamos a cabo un análisis de expresión diferencial de genes entre muestras de tejido tumoral y tejido sano. Este análisis arrojó resultados que indican una significativa sobreexpresión de BPTF en las muestras de tejido tumoral, como se puede apreciar en la **Figura 7.a del anexo II**. Para explorar la relevancia de este gen y sus implicaciones en la regulación transcripcional en el contexto del ADP, se llevó a cabo el experimento de RNA-Seq con una línea celular humana de ADP descrito en el apartado 4.3 del presente trabajo.

El análisis bioinformático del experimento de RNA-Seq permitió comprobar que se contaba con lecturas de buena calidad en los archivos FASTQ de los que se partió, las cuales fueron alineadas y asignadas a sus genes correspondientes con éxito, como se muestra en la **Figura 6 del anexo II**. Así se obtuvo la matriz de cuentas crudas utilizadas en el análisis posterior.



**Figura 8: Análisis preliminar del experimento de RNA-Seq.** Representación de las dos componentes principales de la matriz de expresión de RNA-Seq (muestras coloreadas por condición) (8.a). Gráfico de barras de niveles de expresión de cuatro genes normalizados por pyDeseq2, en las distintas condiciones (8.b). Heatmap mostrando la puntuación de enriquecimiento de cada muestra para firmas del ciclo celular (8.c).

Tras el procesamiento inicial se representó cada muestra por las dos componentes principales de su PCA gracias al cual se detectaron dos outliers que fueron eliminados del análisis para evitar distorsiones en los análisis posteriores. Así obtuvimos las 14 muestras que se representan en la **Figura 8.a**. Después, se llevó a cabo la normalización de la tabla de cuentas por pyDeseq2 y se comprobó que la expresión de BPTF en las muestras tratadas con shSCR era mucho mayor que en el resto,

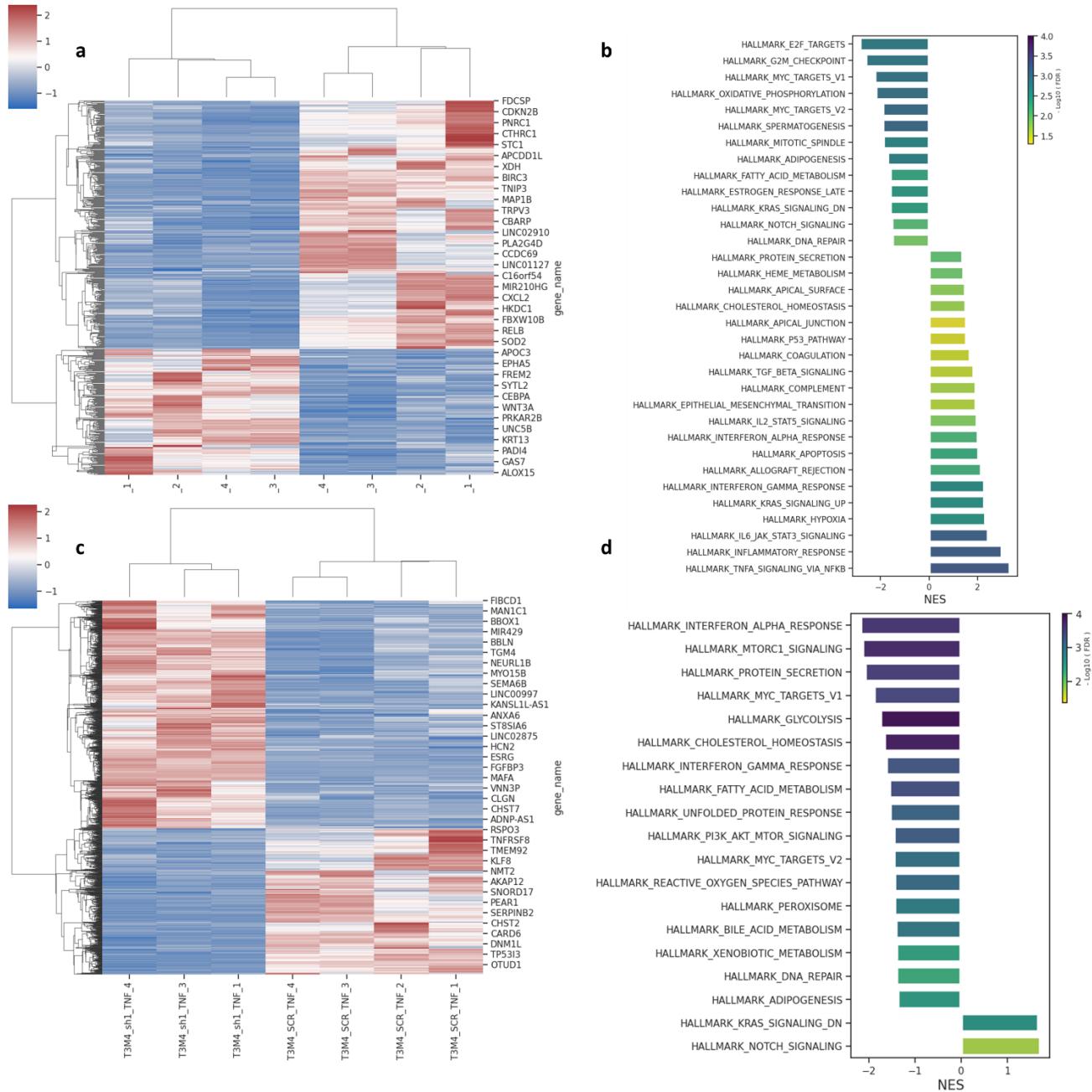
confirmando así el éxito del silenciamiento (**Figura 8.b**). Además, para ratificar que el tratamiento por TNF $\alpha$  había surtido efecto, se evaluó la expresión de genes regulados por esta citoquina como IL-1 $\alpha$ , IL-1 $\beta$  e IL-6. Se pudo observar una clara subida en la expresión de las tres interleuquinas tras el tratamiento con TNF $\alpha$  (**Figura 8.b**), lo cual indica que el tratamiento funcionaba correctamente, tanto en las líneas con BPTF silenciado como en el control. Siendo menor el aumento de expresión de estas citoquinas en ausencia de BPTF.

Como primera aproximación global al análisis del experimento de RNA-Seq se realizó un análisis funcional GSVA con firmas moleculares relacionadas con el ciclo celular y la regulación del sistema inmune. Esto muestra que las muestras en las que se silenció BPTF tenían un menor enriquecimiento de vías relacionadas con la proliferación, fenómeno que se acentúa en las muestras tratadas con TNF $\alpha$  (**Figura 8.c**). Por otra parte, la **Figura 7.b del anexo II** evidencia como las muestras tratadas con TNF $\alpha$  ven incrementado su enriquecimiento para firmas relacionadas con la regulación del sistema inmune.

Para estudiar más en detalle estas diferencias se llevaron a cabo distintos análisis de expresión diferencial, variando las condiciones testadas en cada caso. En primer lugar, para comprobar la respuesta al tratamiento con TNF $\alpha$ , se realizó el contraste shSCR TNF $\alpha$ + contra shSCR TNF $\alpha$ - . De esta forma se obtuvieron 614 DEGs, de los cuales la mayoría (407) estaban sobreexpresados al tratar con TNF $\alpha$ , como se muestra en el heatmap (**Figura 9.a**) y el volcano plot (**Figura 7.c del anexo II**). Tras esto, se llevó a cabo un análisis funcional GSEA, testando las firmas curadas de “Hallmarks” (**Figura 9.b**). El resultado mostró numerosas firmas enriquecidas para la condición sin TNF $\alpha$  (NES negativo) como “MYC targets”, “E2F targets” o “G2M checkpoint”. Además, otras tantas aparecían enriquecidas para la condición tratada con TNF $\alpha$  entre las que encontramos: “Interferon gamma response”, “Inflammatory response”, “TNF signaling via NFKB”, “Epithelial Mesenquimal Transition”, “IL6/JAK/STAT3 signaling”, “KRAS signaling up” y “TGF $\beta$  signaling”.

Para entender el efecto del silenciamiento de BPTF se analizó la expresión diferencial entre las muestras shBPTF TNF $\alpha$ - frente a shSCR TNF $\alpha$ - . Esto dio como resultado 2367 DEGs, de los cuales 1539 estaban sobreexpresados y 828 infraexpresados, como se ve en el volcano plot (**Figura 7.d, anexo II**). Siendo mayor la regulación al alza que a la baja tras silenciar BPTF, como se observa en el heatmap (**Figura 7.f, anexo II**). También en este caso se llevó a cabo un análisis GSEA, (**Figura 7.g, anexo II**) que reveló el enriquecimiento de numerosas rutas en la condición no silenciada, como: “G2M checkpoint”, “MYC targets”, “E2F targets”, “interferon gamma response”, “MTORC1

signaling” o “Protein secretion”. Las células silenciadas por otra parte presentaban enriquecimiento de las firmas de “Apical surface” y “Hedgehog signaling”.



**Figura 9: Análisis de expresión diferencial y funcional del experimento de RNA-Seq.**

Heatmaps mostrando el valor de expresión normalizada por pyDeseq2 de los DEGs y gráficos de barras mostrando el NES para cada firma (coloreados por  $-\log_{10}(\text{FDR})$ ); de los contrastes: shSCR TNF $\alpha$ + vs shSCR TNF $\alpha$ - (**9.a** y **9.b**) y shBPTF (sh1) TNF $\alpha$ + vs SCR TNF $\alpha$ + (**9.c** y **9.d**).

Por último, pasamos a comparar las muestras shBPTF (sh1) TNF $\alpha$ + frente a shSCR TNF $\alpha$ - para estudiar el efecto que ejerce el silenciamiento de BPTF sobre células tumorales ductales en un ambiente en el que está presente TNF $\alpha$ . La expresión

diferencial reveló que había 1869 DEGs, 1139 regulados al alza y 730 a la baja como se ve en el volcán plot (**Figura 7.e del anexo II**). Es decir, el silenciamiento de BPTF en conjunción con el tratamiento con TNF $\alpha$  da lugar a una mayor cantidad de genes sobreexpresados (heatmap, **Figura 9.c**), pero el número de DEGs es menor del que se encuentra en ausencia del tratamiento con TNF $\alpha$ . Un último análisis GSEA (**Figura 9.d**) revela un enriquecimiento de firmas moleculares similar a lo encontrado en el contraste sin tratamiento con TNF $\alpha$ . Hay numerosas diferencias a nivel de puntuación de enriquecimiento (NES) pero de forma cualitativa tan solo cambia que la firma “G2M checkpoint” no aparece ya enriquecida en las células no silenciadas, siendo sustituida por “bile acid metabolism”. Y que las firmas enriquecidas en las células silenciadas en este caso son “KRAS signaling down” y “NOTCH signaling”.

#### **5.4 Desarrollo y validación de una firma de riesgo a partir de genes regulados por BPTF**

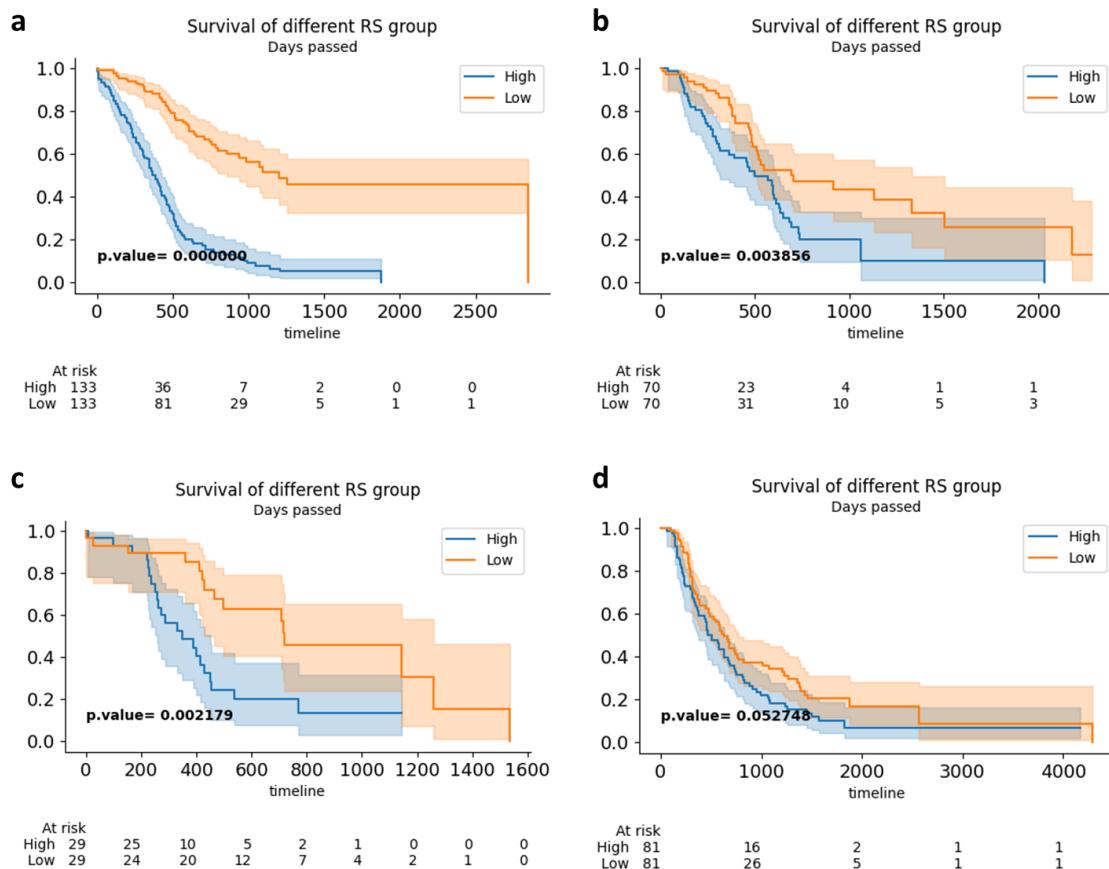
Por último, se tomó la lista de genes diferencialmente expresados en respuesta al silenciamiento de BPTF en muestras tratadas con TNF $\alpha$ , la cual serviría de punto de partida para extraer una firma que permita clasificar muestras de pacientes de alto y bajo riesgo. Para ello se contaba con datos una cohorte de entrenamiento (ICGC-AU2) y otras tres usadas para la validación de la firma (ICGC-AU, ICGC-CA y TCGA) cuyo origen y procesamiento se detalla en el apartado 4.1 de este trabajo.

| Genes    | Coeficientes | P-valor |
|----------|--------------|---------|
| ADAMTS17 | 0.612        | 0.011   |
| ST3GAL1  | 0.372        | 0.014   |
| ETFDH    | 0.46         | 0.009   |
| PCDH11X  | 0.585        | 0.033   |
| BRSK2    | 0.583        | >0.001  |
| RRAS2    | 0.307        | 0.005   |
| LYZ      | -0.125       | 0.046   |
| SLC43A3  | -0.575       | 0.016   |
| FZD10    | 0.477        | 0.006   |
| DKK1     | 0.147        | 0.042   |
| GSN      | -0.482       | 0.009   |
| FGF13    | -0.363       | 0.02    |

**Tabla 1: Genes, coeficientes y valores p de los genes constituyentes de la firma de riesgo.**

Siguiendo con el procedimiento para la obtención de la firma de, se realizó un filtrado de los genes que se usarían para obtener la firma quedando así una lista de 1243 genes. Después se aplicó una regresión Cox Lasso (**Figuras 8.a y 8.b, anexo II**) que permitió

eliminar aquellos genes poco relevantes para la predicción de la supervivencia y reducir la lista inicial de genes a tan solo 37. Estos fueron los usados posteriormente para llevar a cabo una regresión Cox multivariante, que permitió eliminar aquellos genes que no se relacionaron significativamente con la supervivencia y crear una lista final de 12 genes con sus coeficientes de la regresión Cox (**Tabla 1**), que constituyen la firma de riesgo presentada en este trabajo.



**Figura 10: Construcción y validación de una firma de riesgo basada en genes dependientes de BPTF.** Curvas de Kaplan-Meier de grupos de alto (azul) y bajo (naranja) riesgo en los sets de datos de entrenamiento (ICGC-AU2, **10.a**), y de validación: TCGA (**10.b**), ICGC-AU (**10.c**) e ICGC-CA (**10.d**).

En base a esta firma, se calculó un "risk score" para cada muestra de la cohorte de entrenamiento, estratificando sus pacientes en alto o bajo riesgo según se indica en el apartado 4.11. Tras esto, ambos grupos se ajustaron a curvas de Kaplan-Meier y comprobamos que los pacientes de alto riesgo presentaban una menor probabilidad de supervivencia acumulada de forma significativa, según la prueba de rangos logarítmicos

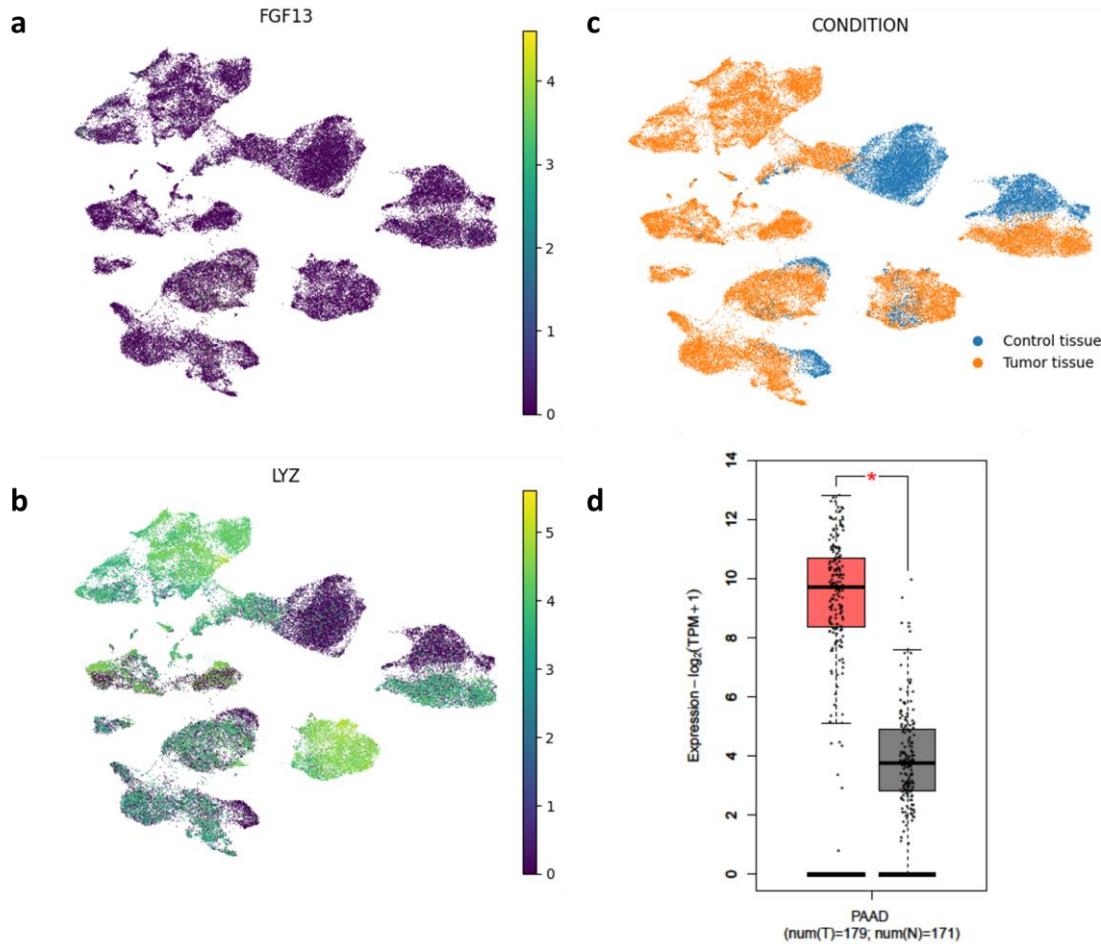
(**Figura 10.a**). Este mismo procedimiento se aplicó a los datasets de validación, usando la firma calculada en el dataset de entrenamiento. Y nuevamente se observó su capacidad de estratificar a los pacientes en grupos de alta y baja supervivencia en las cohortes TCGA y ICGC-AU de forma significativa (**Figuras 10.b y 10.c**). En la cohorte de ICGC-CA las diferencias no fueron significativas, sin embargo, sigue la misma tendencia que las otras con un valor p de 0.052 (**Figura 10.d**).

Finalmente se llevó a cabo un estudio del valor pronóstico individual de cada uno de los genes. Se encontró que tan solo la estratificación de los pacientes en función de la expresión de los genes LYZ y FGF13 generaba dos grupos con diferente distribución de probabilidad de supervivencia acumulada de forma significativa (en al menos tres de los cuatro sets de datos), siendo mayor en ambos casos en los pacientes con altos niveles de estos genes (**Tabla 2 y Figura 9 del anexo II**).

|          | Survival High Expression - Survival Low Expression |          |         |         | Log Rank Test P-value |          |         |        |
|----------|--|----------|---------|---------|-----------------------|----------|---------|--------|
|          | ICGC-AU  | ICGC-AU2 | ICGC-CA | TCGA    | ICGC-AU               | ICGC-AU2 | ICGC-CA | TCGA   |
| ADAMTS17 | -0.0937  | 0.0473   | -0.0121 | -0.1244 | 0.2441                | 0.2502   | 0.9444  | 0.018  |
| BRSK2    | 0.0844   | 0.0306   | 0.0307  | -0.016  | 0.2964                | 0.5164   | 0.5275  | 0.6808 |
| DKK1     | 0.272  | 0.0475   | 0.1448  | 0.0648  | 0.0005                | 0.4062   | 0.0028  | 0.3104 |
| ETFDH    | -0.059   | 0.0575   | -0.0265 | -0.0293 | 0.5576                | 0.139    | 0.7418  | 0.8191 |
| FGF13    | -0.1985  | -0.1374  | -0.0786 | -0.0948 | 0.0236                | 0.0009   | 0.1123  | 0.0381 |
| FZD10    | -0.088   | 0.1      | 0.0143  | -0.0963 | 0.5056                | 0.0063   | 0.9555  | 0.1208 |
| GSN      | -0.1865  | -0.098   | 0.0089  | -0.1176 | 0.1251                | 0.0749   | 0.8226  | 0.0706 |
| LYZ      | -0.2524  | -0.1075  | -0.1474 | -0.1712 | 0.0037                | 0.0079   | 0.0074  | 0.0035 |
| PCDH11X  | 0.0146   | 0.0335   | 0.2647  | -0.1287 | 0.9095                | 0.3342   | 0.2114  | 0.0358 |
| RRAS2    | 0.1946   | 0.1772   | 0.0398  | 0.1025  | 0.0087                | 0.0001   | 0.4538  | 0.1824 |
| SLC43A3  | 0.1139   | -0.107   | 0.0962  | -0.0448 | 0.204                 | 0.0101   | 0.1276  | 0.6151 |
| ST3GAL1  | 0.1729   | 0.1043   | 0.0094  | 0       | 0.112                 | 0.0145   | 0.9472  | 0.7501 |

**Tabla 2: Valor pronóstico individual de los genes constituyentes de la firma de riesgo.**

Por lo tanto, se procedió a visualizar la expresión de estos genes en el dataset de sc-RNA-Seq previamente analizado. En la **Figura 11.c** se muestra la procedencia de las células del dataset (tejido pancreático tumoral o no tumoral) lo que permite distinguir como el gen de la lisozima (LYZ) se expresa casi exclusivamente en las células de origen tumoral (**Figura 11.b**). Por otra parte, no se detectó expresión de FGF13 en el dataset (**Figura 11.a**). Debido a esto, se realizó una búsqueda en GEPIA del gen LYZ y esto reafirmó lo encontrado anteriormente, mostrando las muestras de pacientes de ADP (n=179) mayores niveles de expresión de lisozima que las muestras de páncreas normal (n=171) (**Figura 11.d**) de forma significativa (p<0.05).



**Figura 11: Expresión de dos genes de la firma de riesgo en el dataset de sc-RNA-Seq y de LYZ en muestras de pacientes (GEPIA).** Datos de sc-RNA-Seq mediante la representación UMAP, valores de expresión de los genes normalizados (profundidad 10000) y transformados logarítmicamente ( $\log_{10}(p)$ ) (11.a, 11.b y 11.c). Boxplot con expresión ( $\log_2(\text{TPM}+1)$ ) de LYZ en tejido tumoral ( $n=179$ ) frente a normal ( $n=171$ ), diferencias significativas ( $p<0.05$ ) marcadas con un asterisco (11.d).

## 6. DISCUSIÓN

El primer objetivo de este trabajo consistió en estudiar el adenocarcinoma ductal pancreático (ADP) mediante el análisis de datos de expresión génica y supervivencia de tejidos de pacientes de ADP disponible en bases de datos públicas (TCGA). Para ello se realizó una clasificación no supervisada de los pacientes (clustering gaussiano), hallando así un grupo de pacientes de mayor supervivencia (**Figuras 1.a y 1.b**).

Para entender los procesos subyacentes a esta supervivencia diferencial, se analizaron las redes de genes coexpresados, se estudió la infiltración de células del sistema inmune y se realizó un análisis funcional con los datos de RNA-Seq de ambos grupos

(**Figuras 1.d y 2**). De esta forma, se encontró un módulo de genes relacionados con la activación de células inmunes y con señalización por citoquinas, cuya expresión correlaciona positivamente con el clúster de mayor supervivencia. La deconvolución mostró que efectivamente hay una mayor infiltración de células inmunes, y el análisis funcional corroboró estos resultados, apareciendo enriquecidas firmas de activación del complemento, respuesta a interferón gamma o de respuesta inflamatoria. Esto viene a confirmar lo que ha sido ampliamente descrito en la literatura [11], que la infiltración y activación de células del sistema inmune es un indicador de buen pronóstico en el ADP.

En este grupo se ha detectado también correlación positiva con un módulo relacionado con la remodelación de la matriz extracelular. También se han encontrado firmas relacionadas con la transición epitelio mesénquima. Esto podría indicar una mayor presencia de células del estroma (fibroblastos o células estrelladas) en estos pacientes, aunque este hecho debería ser confirmado mediante otro tipo de herramientas como ESTIMATE [39]. Sin embargo, estos resultados van en la línea de otras publicaciones, que han detectado que un estroma más denso se relaciona con un mejor pronóstico [40].

El clúster de menor supervivencia presenta un escenario en el que hay una menor respuesta e infiltración del sistema inmune, permitiendo quizás un mayor crecimiento celular (enriquecimiento de dianas de E2F y puntos de control G2M). Que podría estar soportado por la reprogramación del metabolismo (enriquecimiento de firmas de glucólisis y fosforilación oxidativa), activado posiblemente, gracias a la acción del oncogén MYC (enriquecimiento de dianas de MYC), un mecanismo ampliamente descrito en la literatura [21].

El segundo objetivo de este proyecto consistía en caracterizar el microambiente tumoral del ADP mediante un análisis de sc-RNA-Seq de tejidos pancreáticos tumorales y no tumorales. Gracias a la ejecución exitosa del pipeline desarrollado en este trabajo, pudimos diferenciar 16 tipos celulares en el set de datos, que se distribuían de forma heterogénea entre las muestras tumorales y no tumorales (**Figura 3.b**). Una de las principales diferencias es que las primeras presentaban poblaciones de células plasmáticas, B y T a diferencia de las segundas. Esto denota la mayor infiltración de células inmunes en el microambiente tumoral. Además, se detectó un subtipo de células T llamadas reguladoras, que son capaces de suprimir la respuesta inmune, favoreciendo la progresión tumoral [16].

También se da un aumento en el número de células estrelladas y la aparición de un subgrupo de miofibroblastos asociados al cancer (my-CAF), lo que numerosos autores han asociado a la formación del estroma desmoplásico, la promoción del crecimiento

tumoral y de un ambiente inmunosupresor [10]. La población de células endoteliales tumorales parece estar teniendo un mayor crecimiento (angiogénesis) que el tejido control, como muestran las rutas sobrerepresentadas del análisis ORA de los genes sobreexpresados. Este proceso puede ser resultado de la mayor expresión de factores de respuesta a hipoxia (HIF1) observado en células endoteliales tumorales, que promueven la angiogénesis [17]. A pesar de este crecimiento, hay mayor proporción de células endoteliales en el tejido control.

La población de macrófagos, por otra parte, es superior en el microambiente tumoral, presentando algunos polarización a M1 y otros a M2, esto es típico del microambiente tumoral del PDAC y nuevamente cuadra con un ambiente inmunosupresor [16]. Al centrarnos en las células ductales tumorales (**Figura 3.e**), se observó una mayor proliferación celular (G2M checkpoints, activación de la vía de las MAPK y de las dianas de E2F) acompañada por el aumento de la señalización por KRAS típica de los tumores de ADP [6]. Además, se observó que las rutas de respuesta inflamatoria y en concreto a TNF se veían activadas en las células ductales malignas, mostrando además una mayor actividad de factores de transcripción implicados en la respuesta a la inflamación como NFKB, AP1 o STAT3.

En una tercera fase del proyecto, se trató de estudiar el mecanismo por el cual BPTF podría promover la progresión del tumoral, y de esclarecer su valor como diana terapéutica en el cáncer de páncreas. En primer lugar, se comprobó su expresión diferencial entre tejidos tumorales y normales de la cohorte del TCGA (GEPIA), descubriendo que BPTF presenta una mayor expresión en los primeros. Además, como hemos visto anteriormente, es común observar un ambiente inflamatorio, con altos niveles de TNF $\alpha$ , en el ADP. Por tanto, se incluyó este factor entre las condiciones del análisis de RNA-Seq.

El experimento de RNA-Seq permitió, en primer lugar, comprobar el efecto que ejercía TNF $\alpha$  sobre líneas celulares T3M4 (modelo de ADP). Gracias al análisis funcional (**Figura 5.b**) se observó una reducción de la proliferación celular (Puntuación de enriquecimiento normalizada (NES) negativa para dianas de E2F y MYC, puntos de control G2M, etc.) y posiblemente la apoptosis (ruta de p53 y firma de apoptosis con NES positiva) en las células tratadas con TNF $\alpha$ , lo que concuerda con lo descrito por la literatura [41]. Además, se da una activación de la respuesta inflamatoria (NES positivo de firmas de respuesta inflamatoria, a interferón gamma/alfa y a TNF $\alpha$  vía NFKB) y de la secreción de citoquinas proinflamatorias como IL1A, IL1B o IL6. Por último, aparece enriquecida la ruta de transición epitelio mesénquima en las células tratadas (EMT). TNF $\alpha$  podría estar aumentando la señalización por TGF $\beta$  (ruta también enriquecida en

células tratadas) y de esta forma promover la EMT, algo que ya se ha descrito en la literatura [42].

Posteriormente, se analizó el efecto que producía el silenciamiento de la expresión de BPTF en células tratadas con TNF $\alpha$  (**Figura 5.d**). Se encontró una reducción de la proliferación celular en las células silenciadas, promovida probablemente por la regulación a la baja de las dianas de MYC (NES negativo en el análisis funcional). Esto concuerda con la literatura, donde BPTF se ha asociado a proliferación mediante activación de la expresión de MYC [21]. Además, al silenciar BPTF se da una reprogramación metabólica, que puede ser causa o consecuencia de la reducción en la proliferación celular y que es orquestada probablemente por el eje KRAS/PI3K/AKT/mTORC1, ya que la señalización de todas estas rutas aparece enriquecida en las células no silenciadas y esta es una vía ampliamente descrita en la literatura [43], [44]. Además, mTORC1 estimula la biosíntesis de proteínas, lípidos y ácidos nucleicos, promoviendo la actividad anabólica crucial para el crecimiento celular, y algunos de estos procesos aparecen también asociados a células no silenciadas en el análisis funcional como la síntesis de ácidos grasos o la adipogénesis.

Por otra parte, es posible que la inhibición de BPTF pueda estar modulando a la respuesta inflamatoria de las células tumorales. En primer lugar, porque se observa que al silenciar BPTF se regulan a la baja las firmas relacionadas con la regulación de la respuesta inmune (**Figura 7, anexo II**), y también porque cuando se trata con TNF $\alpha$  a células no silenciadas, aumenta la expresión de citoquinas proinflamatorias (como hemos comentado antes: IL1A, IL1B e IL6), un aumento que merma cuando BPTF está silenciado (**Figura 4.b**). Además, en las células no silenciadas se encontró el enriquecimiento de la respuesta a interferón alfa y gamma, unos factores cuya expresión es regulada en parte por TNF $\alpha$ , y que pueden promover una respuesta antinflamatoria [45], [46]. Estudios previos han demostrado que BPTF coopera directamente con NFKB (factor de transcripción mediador de los efectos de TNF $\alpha$ ) en la regulación de la ciclooxygenasa 2 en el cáncer pulmonar [47]. Estos resultados apoyan por tanto la hipótesis de que BPTF pueda estar regulando la respuesta a TNF $\alpha$  vía NFKB, dando lugar su silenciamiento a un efecto antiinflamatorio.

Estos hallazgos arrojan luz sobre los mecanismos por los cuales se reprime la proliferación celular tras la inhibición de BPTF y reafirman a este gen como una diana terapéutica prometedora en el tratamiento del ADP. Para clarificar los resultados obtenidos por este proyecto, sería recomendable el estudio del mecanismo que relaciona el silenciamiento de BPTF con la represión de la señalización por KRAS, o una confirmación experimental de mecanismo que lo relaciona la inducción de los genes

de respuesta a interferón alfa/beta. Además, futuras líneas de investigación podrían investigar la capacidad de los inhibidores químicos de BPTF [48] de emular estos resultados, como un primer paso de su translación a la clínica.

En una cuarta y última fase del proyecto, se aprovechó el hecho de que, mediante el experimento de RNA-Seq, se había obtenido una lista de genes (DEGs) de gran relevancia en la progresión tumoral, para el desarrollo y validación de una firma de riesgo. La búsqueda de firmas moleculares es una aproximación de gran utilidad que puede aportar un valor diagnóstico, pronóstico y terapéutico, y que ya ha sido abordada en numerosas ocasiones [49], [50]. Aplicando técnicas de aprendizaje automático (como la regresión Lasso) combinadas con análisis de supervivencia (Cox, Kaplan-Meier), se encontró una firma molecular compuesta por 12 genes dependientes de BPTF. Esta permite estratificar a los pacientes en grupos de alto y bajo riesgo, y su translación a la clínica podría reportar numerosos beneficios. Un estudio más profundo de esta firma es necesario para continuar en esta dirección, por ejemplo, probando si conserva su valor pronóstico al hacer una regresión Cox multivariante incluyendo variables clínicas de los pacientes.

Por otra parte, un análisis independiente de cada uno de los genes que constituyen la firma reveló que la expresión del gen de la lisozima (LYZ), por sí misma, era capaz de estratificar a los pacientes de las cuatro cohortes testadas en grupos de alto y bajo riesgo. Siendo los pacientes con altos niveles de expresión los de mayor supervivencia. Esto es coherente con los efectos anti proliferativos de esta proteína encontrados en el cáncer de mama [51]. De forma aparentemente contradictoria, este gen presenta una mayor expresión en tejidos de tumores de ADP que en tejidos de páncreas sano (**Figura 7.d**), además, se pudo comprobar en el dataset de sc-RNA-Seq que su expresión se limitaba a las células del microambiente tumoral (**Figuras 7.b y 7.c**). Esta aparente incongruencia podría explicarse por un efecto bifásico en la respuesta de las células tumorales a distintas concentraciones de lisozima. Este fenómeno fue observado por un estudio en el que el tratamiento con bajos niveles de lisozima promovía el crecimiento de células de cáncer gástrico mientras que altas dosis inhibían su proliferación [52]. Sin embargo, el mecanismo que da lugar a la sobreexpresión de la lisozima en el ADP y mediante el cual provoca el arresto proliferativo es aún incierto, y abre la puerta a una prometedora línea de investigación en este campo.

Los hallazgos realizados en el presente proyecto ponen de manifiesto la relevancia y creciente importancia de las herramientas bioinformáticas y de aprendizaje automático como motor que impulsa la investigación y expande nuestras fronteras de conocimiento en el ámbito biosanitario.

## 7. CONCLUSIONES

En base a los resultados obtenidos en el presente trabajo se ha llegado a las siguientes conclusiones:

1. Pacientes de ADP con muy poca infiltración inmune tienen menor supervivencia.
2. El microambiente tumoral en el ADP presenta una infiltración inmunosupresora, mayor cantidad de estroma y menor vascularización, debido a la reacción desmoplásica.
3. EL silenciamiento de BPTF promueve el arresto proliferativo de forma dependiente de MYC y de rutas downstring de KRAS como PI3K/AKT/MTORC1, que provocan una reprogramación metabólica.
4. El silenciamiento de BPTF está implicado en la respuesta de las células tumorales a TNF $\alpha$ .
5. Los genes dependientes de BPTF permiten generar una firma de riesgo con valor pronóstico en el ADP.
6. La lisozima emerge como potencial diana terapéutica o biomarcador con valor pronóstico.

## 8. BIBLIOGRAFIA

- [1] M. Malvezzi *et al.*, «European cancer mortality predictions for the year 2023 with focus on lung cancer», *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, vol. 34, n.º 4, pp. 410-419, abr. 2023, doi: 10.1016/j.annonc.2023.01.010.
- [2] R. L. Siegel, K. D. Miller, N. S. Wagle, y A. Jemal, «Cancer statistics, 2023», *CA. Cancer J. Clin.*, vol. 73, n.º 1, pp. 17-48, ene. 2023, doi: 10.3322/caac.21763.
- [3] R. Vera *et al.*, «SEOM Clinical Guideline for the treatment of pancreatic cancer (2016)», *Clin. Transl. Oncol. Off. Publ. Fed. Span. Oncol. Soc. Natl. Cancer Inst. Mex.*, vol. 18, n.º 12, pp. 1172-1178, dic. 2016, doi: 10.1007/s12094-016-1586-x.
- [4] H.-X. Zhan *et al.*, «Crosstalk between stromal cells and cancer cells in pancreatic cancer: New insights into stromal biology», *Cancer Lett.*, vol. 392, pp. 83-93, abr. 2017, doi: 10.1016/j.canlet.2017.01.041.
- [5] C. Guerra *et al.*, «Chronic Pancreatitis Is Essential for Induction of Pancreatic Ductal Adenocarcinoma by K-Ras Oncogenes in Adult Mice», *Cancer Cell*, vol. 11, n.º 3, pp. 291-302, mar. 2007, doi: 10.1016/j.ccr.2007.01.012.
- [6] K. Singh *et al.*, «Kras mutation rate precisely orchestrates ductal derived pancreatic intraepithelial neoplasia and pancreatic cancer», *Lab. Invest.*, vol. 101, n.º 2, pp. 177-192, feb. 2021, doi: 10.1038/s41374-020-00490-5.
- [7] F. Notta, S. A. Hahn, y F. X. Real, «A genetic roadmap of pancreatic cancer: still evolving», *Gut*, vol. 66, n.º 12, pp. 2170-2178, dic. 2017, doi: 10.1136/gutjnl-2016-313317.
- [8] F. V. Opitz, L. Haeberle, A. Daum, y I. Esposito, «Tumor Microenvironment in Pancreatic Intraepithelial Neoplasia», *Cancers*, vol. 13, n.º 24, p. 6188, dic. 2021, doi: 10.3390/cancers13246188.
- [9] D. S. Foster, R. E. Jones, R. C. Ransom, M. T. Longaker, y J. A. Norton, «The evolving relationship of wound healing and tumor stroma», *JCI Insight*, vol. 3, n.º 18, pp. e99911, 99911, sep. 2018, doi: 10.1172/jci.insight.99911.
- [10] P. Lu, V. M. Weaver, y Z. Werb, «The extracellular matrix: a dynamic niche in cancer progression», *J. Cell Biol.*, vol. 196, n.º 4, pp. 395-406, feb. 2012, doi: 10.1083/jcb.201102147.
- [11] T. Tang, X. Huang, G. Zhang, Z. Hong, X. Bai, y T. Liang, «Advantages of targeting the tumor immune microenvironment over blocking immune checkpoint in cancer immunotherapy», *Signal Transduct. Target. Ther.*, vol. 6, n.º 1, p. 72, feb. 2021, doi: 10.1038/s41392-020-00449-4.

- [12] D. Sakellariou-Thompson *et al.*, «4-1BB Agonist Focuses CD8+ Tumor-Infiltrating T-Cell Growth into a Distinct Repertoire Capable of Tumor Recognition in Pancreatic Cancer», *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, vol. 23, n.º 23, pp. 7263-7275, dic. 2017, doi: 10.1158/1078-0432.CCR-17-0831.
- [13] M. Schumacher, I. Dennis, C. Liu, C. Robinson, y M. Frey, «ErbB4 regulates interferon signaling in classically activated macrophages», *FASEB J.*, vol. 35, n.º S1, 2021, doi: 10.1096/fasebj.2021.35.S1.03743.
- [14] L. J. Bayne *et al.*, «Tumor-derived granulocyte-macrophage colony-stimulating factor regulates myeloid inflammation and T cell immunity in pancreatic cancer», *Cancer Cell*, vol. 21, n.º 6, pp. 822-835, jun. 2012, doi: 10.1016/j.ccr.2012.04.025.
- [15] M. Jung *et al.*, «IL-10 improves cardiac remodeling after myocardial infarction by stimulating M2 macrophage polarization and fibroblast activation», *Basic Res. Cardiol.*, vol. 112, n.º 3, p. 33, may 2017, doi: 10.1007/s00395-017-0622-5.
- [16] J. L. Carstens *et al.*, «Spatial computation of intratumoral T cells correlates with survival of patients with pancreatic cancer», *Nat. Commun.*, vol. 8, p. 15095, abr. 2017, doi: 10.1038/ncomms15095.
- [17] A. Yamasaki, K. Yanai, y H. Onishi, «Hypoxia and pancreatic ductal adenocarcinoma», *Cancer Lett.*, vol. 484, pp. 9-15, ago. 2020, doi: 10.1016/j.canlet.2020.04.018.
- [18] A. J. Ruthenburg *et al.*, «Recognition of a mononucleosomal histone modification pattern by BPTF via multivalent interactions», *Cell*, vol. 145, n.º 5, pp. 692-706, may 2011, doi: 10.1016/j.cell.2011.03.053.
- [19] L. Richart *et al.*, «BPTF is required for c-MYC transcriptional activity and in vivo tumorigenesis», *Nat. Commun.*, vol. 7, p. 10153, ene. 2016, doi: 10.1038/ncomms10153.
- [20] R. Muñoz Velasco *et al.*, «Targeting BPTF Sensitizes Pancreatic Ductal Adenocarcinoma to Chemotherapy by Repressing ABC-Transporters and Impairing Multidrug Resistance (MDR)», *Cancers*, vol. 14, n.º 6, p. 1518, mar. 2022, doi: 10.3390/cancers14061518.
- [21] J. Lin *et al.*, «Hypoxia-induced exosomal circPDK1 promotes pancreatic cancer glycolysis via c-myc activation by modulating miR-628-3p/BPTF axis and degrading BIN1», *J. Hematol. Oncol. J Hematol Oncol*, vol. 15, n.º 1, p. 128, sep. 2022, doi: 10.1186/s13045-022-01348-7.
- [22] A. C. Culhane y J. Howlin, «Molecular profiling of breast cancer: transcriptomic studies and beyond», *Cell. Mol. Life Sci. CMLS*, vol. 64, n.º 24, pp. 3185-3200, dic. 2007, doi: 10.1007/s00018-007-7387-1.

- [23] J. Podnar, H. Deiderick, G. Huerta, y S. Hunicke-Smith, «Next-Generation Sequencing RNA-Seq Library Construction», *Curr. Protoc. Mol. Biol.*, vol. 106, p. 4.21.1-4.21.19, abr. 2014, doi: 10.1002/0471142727.mb0421s106.
- [24] M. H. H. Withanage, H. Liang, y E. Zeng, «RNA-Seq Experiment and Data Analysis», *Methods Mol. Biol. Clifton NJ*, vol. 2418, pp. 405-424, 2022, doi: 10.1007/978-1-0716-1920-9\_22.
- [25] Y. Zoabi y N. Shomron, «Processing and Analysis of RNA-seq Data from Public Resources», *Methods Mol. Biol. Clifton NJ*, vol. 2243, pp. 81-94, 2021, doi: 10.1007/978-1-0716-1103-6\_4.
- [26] L. Wang *et al.*, «Multi-omics landscape and clinical significance of a SMAD4-driven immune signature: Implications for risk stratification and frontline therapies in pancreatic cancer», *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 1154-1167, 2022, doi: 10.1016/j.csbj.2022.02.031.
- [27] F. Tang *et al.*, «mRNA-Seq whole-transcriptome analysis of a single cell», *Nat. Methods*, vol. 6, n.º 5, pp. 377-382, may 2009, doi: 10.1038/nmeth.1315.
- [28] Y. Zhang *et al.*, «Single-cell RNA sequencing in cancer research», *J. Exp. Clin. Cancer Res. CR*, vol. 40, n.º 1, p. 81, mar. 2021, doi: 10.1186/s13046-021-01874-1.
- [29] E. Z. Macosko *et al.*, «Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets», *Cell*, vol. 161, n.º 5, pp. 1202-1214, may 2015, doi: 10.1016/j.cell.2015.05.002.
- [30] G. Werba *et al.*, «Single-cell RNA sequencing reveals the effects of chemotherapy on human pancreatic adenocarcinoma and its tumor microenvironment», *Nat. Commun.*, vol. 14, n.º 1, p. 797, feb. 2023, doi: 10.1038/s41467-023-36296-4.
- [31] I. Virshup *et al.*, «The scverse project provides a computational ecosystem for single-cell omics data analysis», *Nat. Biotechnol.*, vol. 41, n.º 5, pp. 604-606, may 2023, doi: 10.1038/s41587-023-01733-8.
- [32] J. Peng *et al.*, «Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma», *Cell Res.*, vol. 29, n.º 9, pp. 725-738, sep. 2019, doi: 10.1038/s41422-019-0195-y.
- [33] E. Azizi *et al.*, «Single-cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment», *Cell*, vol. 174, n.º 5, pp. 1293-1308.e36, ago. 2018, doi: 10.1016/j.cell.2018.05.060.
- [34] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, y P. Tamayo, «The Molecular Signatures Database Hallmark Gene Set Collection», *Cell Syst.*, vol. 1, n.º 6, pp. 417-425, dic. 2015, doi: 10.1016/j.cels.2015.12.004.

- [35] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, y J. P. Mesirov, «Molecular signatures database (MSigDB) 3.0», *Bioinforma. Oxf. Engl.*, vol. 27, n.º 12, pp. 1739-1740, jun. 2011, doi: 10.1093/bioinformatics/btr260.
- [36] M. Ashburner *et al.*, «Gene ontology: tool for the unification of biology. The Gene Ontology Consortium», *Nat. Genet.*, vol. 25, n.º 1, pp. 25-29, may 2000, doi: 10.1038/75556.
- [37] M. Kanehisa y S. Goto, «KEGG: kyoto encyclopedia of genes and genomes», *Nucleic Acids Res.*, vol. 28, n.º 1, pp. 27-30, ene. 2000, doi: 10.1093/nar/28.1.27.
- [38] M. Milacic *et al.*, «The Reactome Pathway Knowledgebase 2024», *Nucleic Acids Res.*, vol. 52, n.º D1, pp. D672-D678, ene. 2024, doi: 10.1093/nar/gkad1025.
- [39] K. Yoshihara *et al.*, «Inferring tumour purity and stromal and immune cell admixture from expression data», *Nat. Commun.*, vol. 4, p. 2612, 2013, doi: 10.1038/ncomms3612.
- [40] L. M. Wang *et al.*, «The prognostic role of desmoplastic stroma in pancreatic ductal adenocarcinoma», *Oncotarget*, vol. 7, n.º 4, pp. 4183-4194, ene. 2016, doi: 10.18632/oncotarget.6770.
- [41] C. Tekin, H. L. Aberson, M. F. Bijlsma, y C. A. Spek, «Early macrophage infiltrates impair pancreatic cancer cell growth by TNF- $\alpha$  secretion», *BMC Cancer*, vol. 20, n.º 1, p. 1183, dic. 2020, doi: 10.1186/s12885-020-07697-1.
- [42] Y. Yoshimatsu *et al.*, «TNF- $\alpha$  enhances TGF- $\beta$ -induced endothelial-to-mesenchymal transition via TGF- $\beta$  signal augmentation», *Cancer Sci.*, vol. 111, n.º 7, pp. 2385-2399, jul. 2020, doi: 10.1111/cas.14455.
- [43] C.-H. Chan *et al.*, «PAK and PI3K pathway activation confers resistance to KRASG12C inhibitor sotorasib», *Br. J. Cancer*, vol. 128, n.º 1, pp. 148-159, ene. 2023, doi: 10.1038/s41416-022-02032-w.
- [44] C. C. Dibble y L. C. Cantley, «Regulation of mTORC1 by PI3K signaling», *Trends Cell Biol.*, vol. 25, n.º 9, pp. 545-555, sep. 2015, doi: 10.1016/j.tcb.2015.06.002.
- [45] W. Wang *et al.*, «Convergent Transcription of Interferon-stimulated Genes by TNF- $\alpha$  and IFN- $\alpha$  Augments Antiviral Activity against HCV and HEV», *Sci. Rep.*, vol. 6, p. 25482, may 2016, doi: 10.1038/srep25482.
- [46] C. Xu *et al.*, «TNF $\alpha$  and IFN $\gamma$  rapidly activate PI3K-AKT signaling to drive glycolysis that confers mesenchymal stem cells enhanced anti-inflammatory property», *Stem Cell Res. Ther.*, vol. 13, n.º 1, p. 491, oct. 2022, doi: 10.1186/s13287-022-03178-3.
- [47] M. Dai *et al.*, «BPTF cooperates with p50 NF- $\kappa$ B to promote COX-2 expression and tumor cell growth in lung cancer», *Am. J. Transl. Res.*, vol. 11, n.º 12, pp. 7398-7409, 2019.

- [48] H. Zahid *et al.*, «New Design Rules for Developing Potent Cell-Active Inhibitors of the Nucleosome Remodeling Factor (NURF) via BPTF Bromodomain Inhibition», *J. Med. Chem.*, vol. 64, n.º 18, pp. 13902-13917, sep. 2021, doi: 10.1021/acs.jmedchem.1c01294.
- [49] W. Song *et al.*, «Glycolysis-Related Gene Expression Profiling Screen for Prognostic Risk Signature of Pancreatic Ductal Adenocarcinoma», *Front. Genet.*, vol. 12, p. 639246, 2021, doi: 10.3389/fgene.2021.639246.
- [50] C. Wei *et al.*, «Bioinformatics profiling utilized a nine immune-related long noncoding RNA signature as a prognostic target for pancreatic cancer», *J. Cell. Biochem.*, vol. 120, n.º 9, pp. 14916-14927, sep. 2019, doi: 10.1002/jcb.28754.
- [51] S. Mahanta, S. Paul, A. Srivastava, A. Pastor, B. Kundu, y T. K. Chaudhuri, «Stable self-assembled nanostructured hen egg white lysozyme exhibits strong anti-proliferative activity against breast cancer cells», *Colloids Surf. B Biointerfaces*, vol. 130, pp. 237-245, jun. 2015, doi: 10.1016/j.colsurfb.2015.04.017.
- [52] T. K. Guo *et al.*, «The anti-proliferative effects of recombinant human lysozyme on human gastric cancer cells», *J. Int. Med. Res.*, vol. 35, n.º 3, pp. 353-360, 2007, doi: 10.1177/147323000703500310.

## **ANEXO I: EXTENSIÓN DE MATERIALES Y MÉTODOS**

### **1. Obtención y procesamiento de conjuntos de datos públicos**

En este estudio se emplearon cuatro conjuntos de datos de adenocarcinomas pancreáticos (recolectados y preprocesados por el Dr. Ramón González, siendo tres de ellos obtenidos del repositorio público del Consorcio Internacional del Genoma del Cáncer (ICGC) y uno del Atlas del Genoma del Cáncer (TCGA). Se excluyeron los pacientes con datos incompletos de supervivencia global en todos los conjuntos de datos.

Los tres conjuntos de datos del ICGC son los siguientes: 1) ICGC-AU2 de microarrays (n=267), en este caso no se realizó una selección adicional de pacientes; 2) ICGC-AU de RNA-Seq (n=59); y 3) ICGC-CA de RNA-Seq (n=178), en los cuales se seleccionaron pacientes con muestras de tumores primarios pancreáticos y con histología de adenocarcinoma ductal pancreático. El cuarto conjunto de datos utilizado fue el de TCGA (n=141), donde también se seleccionaron pacientes con muestras de tumores primarios pancreáticos y con histología de adenocarcinoma ductal pancreático.

Los datos de ICGC-AU2 (de microarrays) provienen de una serie australiana con datos de expresión génica de la plataforma illuminaHumanv4. Para las tres series del ICGC, se utilizaron datos de expresión génica y datos de supervivencia global disponibles en el portal del ICGC (<https://dcc.icgc.org/>). Los datos brutos estimados de expresión de TCGA a partir de RSEM se obtuvieron de Firebrowse (<http://firebrowse.org/>) del Broad Institute, mientras que los datos clínicos y de supervivencia global se obtuvieron del portal "<https://portal.gdc.cancer.gov/>".

Se obtuvieron, en la medida de lo posible, los datos brutos como punto de partida, a excepción del conjunto de microarrays ICGC-AU2, donde los únicos datos de expresión disponibles eran datos normalizados (normalización cuantil) y ya transformados logarítmicamente en base 2. Para las otras dos series del ICGC y la serie de TCGA (de RNA-Seq), se partió de tablas de cuentas “crudas”. Los conteos brutos de expresión de las tres series se convirtieron en transcripts per million (TPM) y posteriormente fueron transformados logarítmicamente:  $\log_2(\text{TPM} + 1)$ .

## 2. Obtención y procesamiento de datos de single cell RNA-Seq

Los datos utilizados para el análisis de sc-RNA-Seq proceden de un conjunto de datos generado y preprocesado por Peng *et al.* [1] y se encuentran públicamente disponibles en el “Genome Sequence Archive” con el número de acceso GSA: CRA001160. Este set de datos contiene muestras de 24 pacientes con tumores primarios de ADP y 11 de tejido pancreático normal. Como informan los autores del artículo de referencia, se secuenció mediante la tecnología Illumina HiSeq X Ten, generando lecturas pareadas. Estas lecturas fueron procesadas mediante el pipeline Cell Ranger 2.1.0 con los parámetros predeterminados. Los archivos FASTQ fueron alineados al genoma de referencia humano (hg19) mediante el algoritmo STAR. Luego, generaron matrices de genes y códigos de barras para cada muestra individual contando identificadores moleculares únicos (UMIs). Esto se encuentra disponible públicamente para su descarga en la localización antes indicada, y representa el punto de partida de nuestro reanálisis.

Para la ejecución del análisis, se implementaron principalmente los paquetes ScanPy (versión 1.9.5) [2] y scvi-tools (versión 0.6.8) [3] de Python. Siguiendo los pasos típicos de un análisis de sc-RNA-Seq, tras cargar los datos en una estructura AnnData (versión 0.9.2) se calcularon parámetros de calidad, como el número de cuentas totales (y mitocondriales) por gen o por código de barras (función calculate\_qc\_metrics de ScanPy). Usando estos parámetros se calcularon y filtraron outliers siguiendo el criterio MAD (Median Absolute Desviation). También se realizó una estimación y filtrado de dobletes (códigos de barras asignados a gotas que han capturado más de una célula) mediante el módulo SOLO [4] incluido en scvi-tools, que usa algoritmos de deep learning semi-supervisado para realizar esta clasificación.

Las cuentas crudas se normalizaron a una profundidad de 10000 cuentas y se transformaron logarítmicamente (funciones normalize\_total y log1p del módulo de preprocesamiento de ScanPy). Se calcularon y seleccionaron los 2000 genes más variables con la función highly\_variable\_genes, utilizados para los análisis posteriores. Además, se eliminaron los efectos de recuento total por célula y del porcentaje de genes mitocondriales expresados mediante regresión y se escalaron los datos (funciones regress\_out y scale de ScanPy).

En la fase de integración se usó la herramienta SCVI [5] de scvi-tools para modelar la variabilidad del conjunto de datos y eliminar la variabilidad asociada al paciente del que proviene la muestra. Además, de esta forma se obtuvo también una representación latente (baja dimensionalidad) de los datos modelados, que se usaría para calcular la

vecindad entre células mediante la función neighbors de ScanPy y así poder representarlos mediante un UMAP (Uniform Manifold Approximation and Projection).

En un último paso se llevó a cabo la anotación de los tipos celulares presentes en el set de datos. Esto se realizó en tres pasos, comenzando con una anotación por transferencia de etiquetas usando el módulo SCANVI [6] de scvi-tools y un set de datos anotado de referencia de tejido pancreático obtenido de Tabula Sapiens ([https://figshare.com/articles/dataset/Tabula\\_Sapiens\\_release\\_1\\_0/14267219](https://figshare.com/articles/dataset/Tabula_Sapiens_release_1_0/14267219)). En el segundo paso las células se agruparon en clústeres mediante el método de Leiden (a resolución 0.7) implementado en ScanPy y se llevó a cabo una anotación manual de los clústeres, partiendo de la anotación obtenida en el paso anterior y refinándola con la información de la expresión de marcadores de tipos celulares obtenidos de la base de datos CellMarker 2.0 [7]. Por último, se diferenciaron algunos subtipos celulares (tres subtipos de células T) mediante el subclustering de tipos celulares concretos (también por Leiden, con una resolución de 0.2) y la identificación de marcadores de estos subtipos en los subclústeres.

### **3. Preparación y análisis del experimento de RNA Seq**

Se generaron 16 muestras, cuatro por cada condición, siendo estas condiciones: shBPTF TNF $\alpha$ - / shBPTF TNF $\alpha$ + / shSCR TNF $\alpha$ - / shSCR TNF $\alpha$ +. Dos compañeros del grupo de investigación (María Ferrer y Raúl Muñoz) llevaron a cabo la preparación de las muestras. Se introdujo un plásmido portador del sistema de shRNA/shSCR para silenciar el gen de BPTF en células humanas de la línea T3M4 (modelo de ADP) mediante infección con vectores virales (lentivirus), producidos previamente en la línea celular empaquetadora HEK293T. Tras su infección con los shRNA contra BPTF (shBPTF) y el control (shSCR), la mitad de las muestras fueron tratadas con el Factor de Necrosis Tumoral Alfa (TNF $\alpha$ ) (R&DSystems) a una concentración de 20 ng/ $\mu$ l 24 horas después de ser plantadas. Tras 24 horas de tratamiento, todas las células en cultivo (de muestras tratadas y no tratadas) fueron levantadas con tripsina (Corning,05322010) y el RNA fue extraído y purificado usando el kit RNeasy Mini Kit (50) (QIAGEN, REF: 74104) de acuerdo con el protocolo. Se cuantificó la concentración de RNA resultante usando el espectrofotómetro NanoDrop (ThermoFisher). Se comprobó el silenciamiento de las líneas para BPTF, así como el tratamiento con TNF $\alpha$  mediante RT-qPCR. Además, se analizó el RNA integrity number (RIN) para la determinación de la calidad del RNA. Tras esto, las muestras se enviaron a Beijing Genomics Institute's (BGI) donde prepararon librerías por el método de cDNA y se secuenció utilizando la

tecnología “DNA nanoball sequencing” (DNB-seq). Como resultado se obtuvieron lecturas pareadas en archivos FASTQ, que fueron el punto de partida para el análisis subsiguiente.

El análisis bioinformático de estos archivos comprendió:

1. Un control de calidad mediante FASTQC (versión 0.12.0, <https://github.com/s-andrews/FastQC>)-
2. Un “trimming” mediante la herramienta Trim Galore (versión 0.6.8, [https://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)) para eliminar los adaptadores de la secuenciación (--hardtrim3 137, eliminando las 13 primeras bases).
3. Un alineamiento frente al genoma de referencia GRCh38 mediante STAR (versión 2.7.11a) [8].
4. Un procesamiento de los archivos BAM producidos en el alineamiento para generar una matriz de cuentas mediante featureCounts [9] un algoritmo implementado en el paquete Subread (versión 2.0.6).
5. Un control de calidad final, se usó la herramienta multiQC (versión 1.19) [10] para generar un reporte resumen de todos los pasos mencionados anteriormente.

Todas las herramientas se usaron con los parámetros por defecto y recomendados a excepción de lo indicado para Trim Galore.

#### **4. Análisis de expresión diferencial de genes**

Para la consecución del análisis de expresión diferencial con datos de bulk-RNA-Seq se usó la herramienta pyDeseq2 (versión 0.4.4)[11] de Python. Para realizar el análisis se parte de una matriz de expresión, con las cuentas brutas en formato genes x muestra, otra tabla de metadatos, en la que se asigna una condición a cada muestra, y una lista que especifica las condiciones comparar. Como resultado se obtiene una matriz de expresión normalizada y una tabla de resultados, en la que se muestran el Log2 Fold Change o el P-valor ajustado (corregido para el descubrimiento de falsos positivos o FDR por el método de Benjamini-Hochberg), entre otros parámetros.

Para la expresión diferencial de genes entre las condiciones “Malign ductal cell” y “Ductal cell” en el análisis de sc-RNA-Seq se aplicaron dos métodos. El primero emplea el método Wilcoxon de la función rank\_genes\_groups de Scanpy para comparar la expresión génica de las células pertenecientes a las distintas condiciones. Además, se

utilizó un segundo método en el cual se transformaban los datos de expresión de células individuales en un “pseudobulk”. Se hizo asignando las células a 10 muestras tumorales y 10 no tumorales de forma arbitraria, tras lo cual se agregó el valor de expresión (no normalizado) de las células asignadas a cada muestra aplicando el método `get_pseudobulk` del paquete `decoupleR` [12] de Python. Por último, se realiza la expresión diferencial entre ambas condiciones mediante `pyDESeq2`, como se ha descrito anteriormente en este apartado. También se analizó la expresión diferencial entre células endoteliales procedentes de muestras tumorales y no tumorales, para lo cual se aplicó tan solo la función `rank_genes_groups`.

Tras el análisis de expresión diferencial se puede determinar que genes están diferencialmente expresados de forma significativa (DEGs). Los criterios para considerar un gen como DEG fueron un cambio de al menos dos veces en la expresión ( $|Log2\ Fold\ Change| > 1$ ), y un p-valor ajustado menor a 0.05. Los resultados se representaron gráficamente usando los paquetes de Python `Matplotlib` (versión 3.8.0) y `Seaborn` (versión 0.12.2) para generar volcano plots y heatmaps. Estos últimos representan la expresión normalizada de los genes diferencialmente expresados (DEGs) en cada muestra, utilizando un z-score para escalar los valores y aplicando una clusterización jerárquica para ordenar las muestras y los genes en el gráfico (función `clustermap` de `Seaborn`).

## 5. Análisis funcional

A lo largo de este trabajo se llevaron a cabo dos tipos de análisis funcionales para tratar de entender la relevancia biológica de los resultados experimentales. El primero de estos es el análisis de sobrerepresentación (ORA, del inglés “Over Representation Analysis”). Este tipo de análisis testa el enriquecimiento de un conjunto de genes de interés en distintos sets de genes correspondientes a firmas moleculares. Para esto se utilizó el paquete `GSEApY` (versión 1.1.0) [13] de Python, que a su vez hace uso de la API de `Enrichr` [14] para llevar a cabo el análisis. Como resultado se obtiene un valor de significancia (la prueba de hipergeometría) corregido para FDR (significativo si  $p < 0.05$ ) y el número de genes de interés que aparece en la firma testada.

También se realizaron análisis de enriquecimiento de sets de genes (GSEA, también incluido en `GSEApY`), mediante los cuales se testa el enriquecimiento de firmas moleculares mediante el rankeado de los genes en base a su correlación con la condición de interés. Se parte de las matrices de expresión normalizadas por `pyDESeq2` y como resultado se obtiene un valor de puntuación de enriquecimiento normalizado

(NES), que será positivo cuando se correlacione positivamente con la condición de referencia y negativo cuando correlacione con la condición opuesta, y un valor de significancia (test de permutación) corregido por FDR (significativo si  $p < 0.05$ ).

Por último, en los casos en los que se tenían más de dos condiciones a comparar, se aplicó el análisis de variación de sets de genes (GSVA). Este análisis (incluido también en GSEAp) calcula un valor de enriquecimiento normalizado para cada muestra y firma molecular testada. Los resultados se representaron mediante heatmaps (función clustermap, Seaborn).

## 6. Actividad de vías moleculares y factores de transcripción

Para la consecución de este análisis complementario en el dataset de sc-RNA-Seq se implementaron dos funciones del paquete decoupleR. Se implementó la función run\_mlm de este paquete, que aplica un modelo linear multivariante a los datos de expresión de células individuales, para predecir la actividad de una colección curada de vías moleculares contenidas en PROGENy [15], que además contiene la información de sus genes diana y de los pesos asignados a cada interacción.

De forma similar, la función run\_ulp se usó para detectar la actividad de factores de transcripción en células únicas. Se aplica en este caso un modelo univariante para cada célula del dataset y cada factor de transcripción contenido en CollecTRI [16], una base de datos curada de regulones. La actividad por célula, tanto de las rutas moleculares como de los factores de transcripción se representó mediante UMAPs.

## 7. Reducción dimensional y clústering

En diversas fases de este trabajo se llevó a cabo una reducción dimensional de datos de expresión genética. Esto se hizo mediante un análisis de componentes principales (PCA) implementado en el paquete de Python scikit-learn (versión 1.3.2) [23], previo escalado de los datos (de forma que tuvieran media 0 y desviación estándar 1) mediante el método StandardScaler, incluido en el mismo paquete.

Para el clústering de las muestras del TCGA, realizado por el profesor Luis Bote en calidad de colaborador del grupo de investigación, se tomaron las dos primeras componentes del PCA antes calculado y se aplicó el método de modelos gaussianos mixtos, de aprendizaje automático no supervisado, con dos componentes. Para ello se usó el método GaussianMixture del módulo mixture de scikit-learn.

## **8. Análisis WGCNA**

Para el análisis WGCNA se usó el conjunto de datos de expresión de RNA del TCGA descrito anteriormente en el apartado 1 de este anexo. El análisis se llevó a cabo en Python mediante la el paquete PyWGCNA (versión 1.20.4) [24] e implicó:

1. Una fase de preprocesamiento en el que se detectaron outliers mediante un clustering jerárquico de las distancias entre las muestras, con el método preprocess de la clase WGCNA.
2. Una búsqueda de módulos de genes co-expresados con el método findModules, que primero obtiene el “soft power”, potencia a la que se eleva la matriz de correlación para obtener un grafo de co-expresión libre de escala (ajuste con un  $R^2 > 0.9$  a una la recta de distribución del grado de la red transformada logarítmicamente) y después lo subdivide en módulos de genes.
3. Un análisis de correlación del eigengene (componente principal de la expresión de los genes de un módulo) de los módulos de genes con variables asociadas a las muestras, con el método analyseWGCNA.

Las variables estudiadas fueron: la edad, el género, el estatus nodal (N, N1 si se ha diseminado a nodos linfáticos, N0 sino), tamaño del tumor (T), metástasis (M, M1 si existe, M0 sino) y grado histológico del tumor (G1-2 más diferenciados, G3 menos diferenciado). Además, se añadió la variable de pertenencia al clúster 1 o 0 calculados con anterioridad.

## **9 Deconvolución de células inmunes**

Para realizar la deconvolución de la infiltración de células del sistema inmune en los pacientes del TCGA se usando dos métodos del paquete de Python TumorDecon [25] (versión 1.1.1). Uno es SingScore [26], basado en la puntuación de muestras únicas al ranquear su expresión génica para firmas moleculares. El otro es DeconRNASeq [27] que se basa en modelos lineales y requiere de una matriz de expresión de referencia para los tipos celulares que se quiera cuantificar (la matriz usada por defecto es de poblaciones de células inmunes). Esto produce matrices con una puntuación (SingScore) o frecuencia (DeconRNASeq) para cada tipo de célula inmune por muestra.

Se compararon los niveles medios de cada tipo celular entre los dos clústeres antes calculados mediante un t-test (o Wilcoxon, en caso de no normalidad) mediante la

librería Scipy (versión 1.11.2) [28] y se representaron aquellos que presentaban diferencias significativas ( $p < 0.05$ ) con ambos algoritmos mediante un boxplot y un heatmap (clustermap) de Seaborn. En los heatmaps se incluyó también la pertenencia al clúster, y los datos aparecen ordenados según un clustering jerárquico.

## 10. Obtención y validación de una firma de riesgo.

Para la construcción de la firma de riesgo, se utilizó un conjunto de datos de entrenamiento (ICGC-AU2,  $n=267$ ) y tres conjuntos de datos de validación (ICGC-AU, ICGC-CA y TCGA), los cuales se detallan en la sección 1 del presente anexo. Se procedió a filtrar estos conjuntos, reteniendo los valores de expresión de los genes presentes en todos los conjuntos de datos y que, además, mostraron expresión diferencial al inhibir BPTF con TNF $\alpha$ .

Para investigar la asociación entre la expresión de estos genes y la supervivencia, se llevó a cabo una regresión de Cox con regularización Lasso en los datos de entrenamiento, utilizando la función CoxnetSurvivalAnalysis del paquete de Python scikit-survival (versión 0.22.2) [29]. Previamente, se escaló la información mediante StandardScaler. La determinación del parámetro de regularización ( $\alpha$ ) se realizó mediante validación cruzada con cinco pliegues, seleccionando el valor óptimo de  $\alpha$  basado en la maximización del índice de concordancia del modelo.

Se excluyeron del análisis aquellos genes cuyos coeficientes se redujeron a cero. Los genes restantes se utilizaron como entrada para la construcción de un nuevo modelo Cox mediante la función CoxPHFitter del módulo lifelines (versión 0.28.0) [30] de Python. En este caso, no se aplicó regularización; en su lugar, se seleccionaron los genes significativos ( $p < 0.05$ ) mediante la prueba de razón de verosimilitudes. Estos genes, junto con sus coeficientes asociados, constituyeron la firma de riesgo.

Para aplicar la firma sobre un set de datos, se calcula una puntuación para cada muestra mediante la combinación lineal de la expresión de los genes de la firma multiplicados por sus coeficientes antes calculados, siguiendo la expresión:

$$\text{Risk score} = \sum_{i=1}^n \beta_i \cdot x_i$$

Donde  $n$  es el número total de genes de la firma, y  $\beta_i$  y  $x_i$  son el coeficiente del análisis Cox multivariante y el valor de expresión de cada gen  $i$ . Se calcularía una puntuación de riesgo (risk score) para el conjunto de datos de entrenamiento y los de validación. Esta puntuación se utilizaría para estratificar a los pacientes en grupos de alto y bajo riesgo,

usando la mediana del risk score en el conjunto de datos como punto de corte. Posteriormente, se ajustarían los valores de supervivencia de ambos grupos a curvas de Kaplan-Meier utilizando la función KaplanMeierFitter del módulo lifelines. Se lleva a cabo la prueba de rangos logarítmicos para evaluar si existe una diferencia significativa ( $p < 0.05$ ) entre las funciones de supervivencia de ambos grupos.

### **11. Análisis de los genes constituyentes de la firma de riesgo.**

Los pacientes de las cuatro cohortes se estratificaron en grupos de alta y baja expresión para cada gen de la firma de riesgo, y se aplicaron curvas de Kaplan-Meier con pruebas de rangos logarítmicos ( $p < 0.05$ ), para comprobar el valor pronóstico de cada gen.

Se enfocó un análisis detallado en genes con valor pronóstico validado en al menos tres de las cuatro cohortes, verificando su expresión en el conjunto de datos de sc-RNA-Seq. Además, se examinó la expresión en tejido pancreático tumoral frente al normal a través del portal GEPIA (<http://gepia.cancer-pku.cn/>) [31] para genes con patrones de expresión diferencial en tejido tumoral según el sc-RNA-Seq analizado.

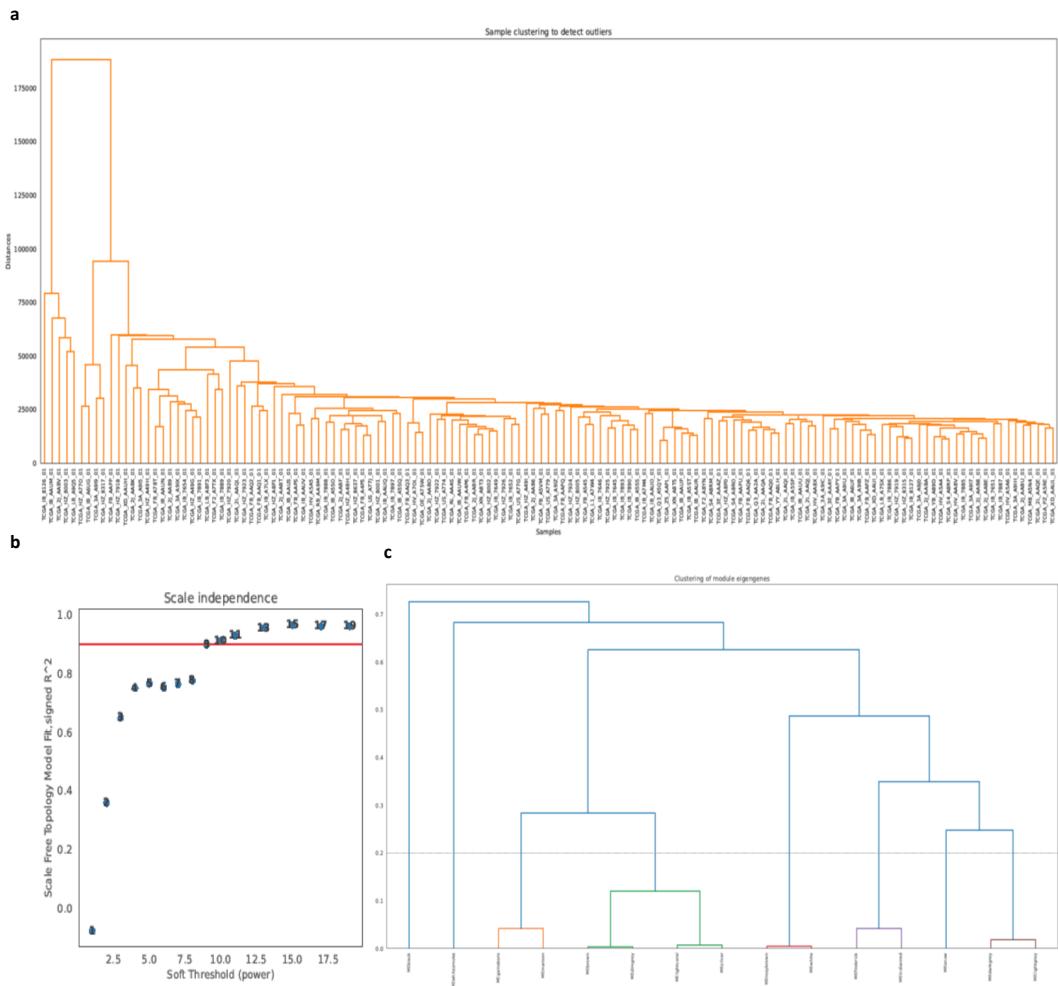
## BIBLIOGRAFÍA

- [1] J. Peng *et al.*, «Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma», *Cell Res.*, vol. 29, n.º 9, pp. 725-738, sep. 2019, doi: 10.1038/s41422-019-0195-y.
- [2] F. A. Wolf, P. Angerer, y F. J. Theis, «SCANPY: large-scale single-cell gene expression data analysis», *Genome Biol.*, vol. 19, n.º 1, p. 15, feb. 2018, doi: 10.1186/s13059-017-1382-0.
- [3] A. Gayoso *et al.*, «A Python library for probabilistic analysis of single-cell omics data», *Nat. Biotechnol.*, vol. 40, n.º 2, Art. n.º 2, feb. 2022, doi: 10.1038/s41587-021-01206-w.
- [4] N. J. Bernstein, N. L. Fong, I. Lam, M. A. Roy, D. G. Hendrickson, y D. R. Kelley, «Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning», *Cell Syst.*, vol. 11, n.º 1, pp. 95-101.e5, jul. 2020, doi: 10.1016/j.cels.2020.05.010.
- [5] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, y N. Yosef, «Deep generative modeling for single-cell transcriptomics», *Nat. Methods*, vol. 15, n.º 12, pp. 1053-1058, dic. 2018, doi: 10.1038/s41592-018-0229-2.
- [6] C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, y N. Yosef, «Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models», *Mol. Syst. Biol.*, vol. 17, n.º 1, p. e9620, ene. 2021, doi: 10.1525/msb.20209620.
- [7] C. Hu *et al.*, «CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data», *Nucleic Acids Res.*, vol. 51, n.º D1, pp. D870-D876, ene. 2023, doi: 10.1093/nar/gkac947.
- [8] A. Dobin y T. R. Gingeras, «Mapping RNA-seq Reads with STAR», *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al*, vol. 51, p. 11.14.1-11.14.19, sep. 2015, doi: 10.1002/0471250953.bi1114s51.
- [9] Y. Liao, G. K. Smyth, y W. Shi, «featureCounts: an efficient general purpose program for assigning sequence reads to genomic features», *Bioinforma. Oxf. Engl.*, vol. 30, n.º 7, pp. 923-930, abr. 2014, doi: 10.1093/bioinformatics/btt656.
- [10] P. Ewels, M. Magnusson, S. Lundin, y M. Käller, «MultiQC: summarize analysis results for multiple tools and samples in a single report», *Bioinforma. Oxf. Engl.*, vol. 32, n.º 19, pp. 3047-3048, oct. 2016, doi: 10.1093/bioinformatics/btw354.

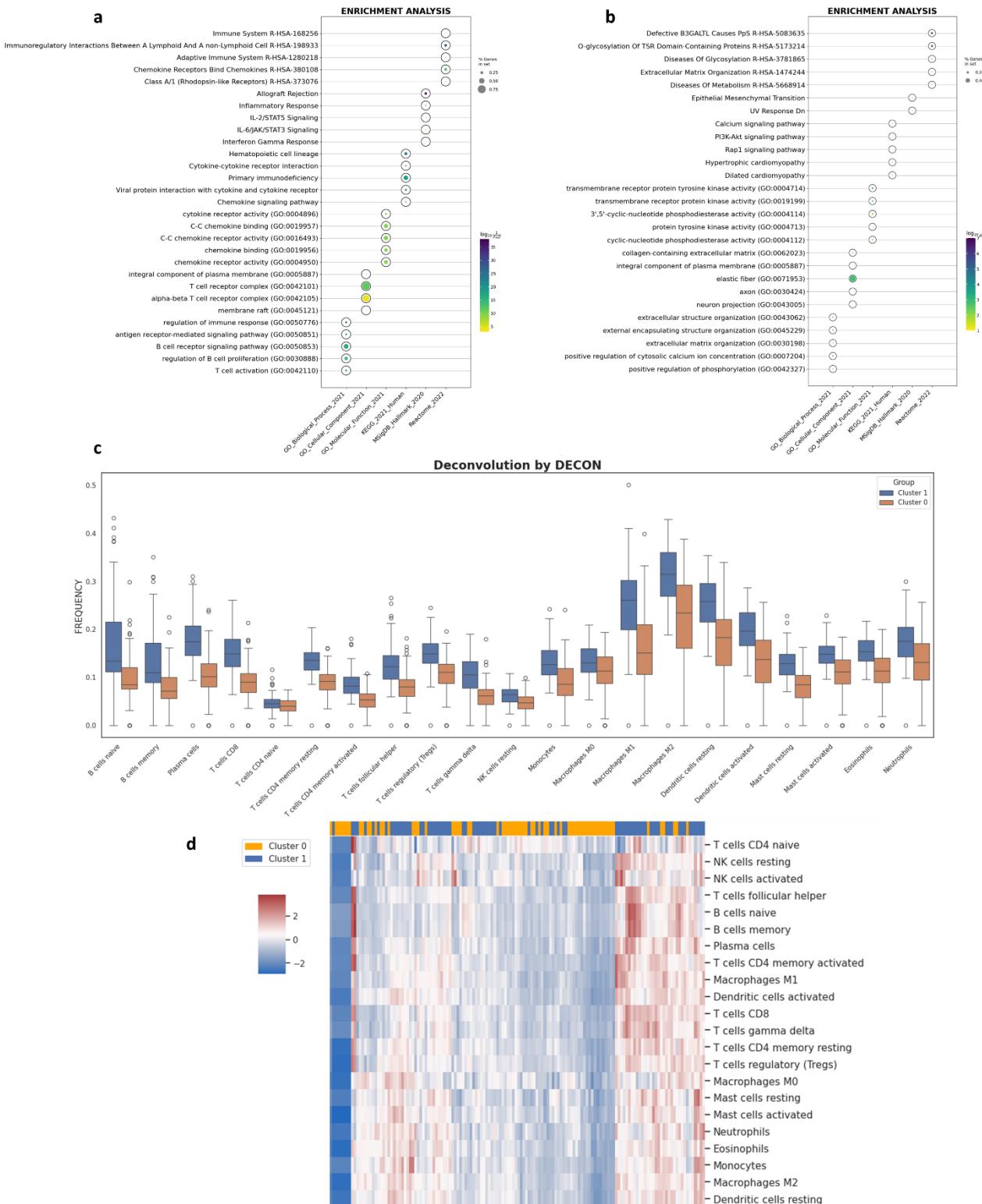
- [11] B. Muzellec, M. Teleńczuk, V. Cabeli, y M. Andreux, «PyDESeq2: a python package for bulk RNA-seq differential expression analysis», *Bioinforma. Oxf. Engl.*, vol. 39, n.º 9, p. btad547, sep. 2023, doi: 10.1093/bioinformatics/btad547.
- [12] P. Badia-I-Mompel *et al.*, «decoupleR: ensemble of computational methods to infer biological activities from omics data», *Bioinforma. Adv.*, vol. 2, n.º 1, p. vbac016, 2022, doi: 10.1093/bioadv/vbac016.
- [13] Z. Fang, X. Liu, y G. Peltz, «GSEAPy: a comprehensive package for performing gene set enrichment analysis in Python», *Bioinformatics*, vol. 39, n.º 1, p. btac757, ene. 2023, doi: 10.1093/bioinformatics/btac757.
- [14] Z. Xie *et al.*, «Gene Set Knowledge Discovery with Enrichr», *Curr. Protoc.*, vol. 1, n.º 3, p. e90, mar. 2021, doi: 10.1002/cpz1.90.
- [15] M. Schubert *et al.*, «Perturbation-response genes reveal signaling footprints in cancer gene expression», *Nat. Commun.*, vol. 9, n.º 1, p. 20, ene. 2018, doi: 10.1038/s41467-017-02391-6.
- [16] S. Müller-Dott *et al.*, «Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities», *Nucleic Acids Res.*, vol. 51, n.º 20, pp. 10934-10949, nov. 2023, doi: 10.1093/nar/gkad841.
- [17] E. Azizi *et al.*, «Single-cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment», *Cell*, vol. 174, n.º 5, pp. 1293-1308.e36, ago. 2018, doi: 10.1016/j.cell.2018.05.060.
- [18] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, y P. Tamayo, «The Molecular Signatures Database Hallmark Gene Set Collection», *Cell Syst.*, vol. 1, n.º 6, pp. 417-425, dic. 2015, doi: 10.1016/j.cels.2015.12.004.
- [19] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, y J. P. Mesirov, «Molecular signatures database (MSigDB) 3.0», *Bioinforma. Oxf. Engl.*, vol. 27, n.º 12, pp. 1739-1740, jun. 2011, doi: 10.1093/bioinformatics/btr260.
- [20] M. Ashburner *et al.*, «Gene ontology: tool for the unification of biology. The Gene Ontology Consortium», *Nat. Genet.*, vol. 25, n.º 1, pp. 25-29, may 2000, doi: 10.1038/75556.
- [21] M. Kanehisa y S. Goto, «KEGG: kyoto encyclopedia of genes and genomes», *Nucleic Acids Res.*, vol. 28, n.º 1, pp. 27-30, ene. 2000, doi: 10.1093/nar/28.1.27.
- [22] M. Milacic *et al.*, «The Reactome Pathway Knowledgebase 2024», *Nucleic Acids Res.*, vol. 52, n.º D1, pp. D672-D678, ene. 2024, doi: 10.1093/nar/gkad1025.
- [23] F. Pedregosa *et al.*, «Scikit-learn: Machine Learning in Python», *J. Mach. Learn. Res.*, vol. 12, n.º 85, pp. 2825-2830, 2011.

- [24] N. Rezaie, F. Reese, y A. Mortazavi, «PyWGCNA: a Python package for weighted gene co-expression network analysis», *Bioinforma. Oxf. Engl.*, vol. 39, n.º 7, p. btad415, jul. 2023, doi: 10.1093/bioinformatics/btad415.
- [25] R. A. Aronow, S. Akbarinejad, T. Le, S. Su, y L. Shahriyari, «TumorDecon: A digital cytometry software», *SoftwareX*, vol. 18, jun. 2022, doi: 10.1016/j.softx.2022.101072.
- [26] M. Foroutan, D. D. Bhuva, R. Lyu, K. Horan, J. Cursons, y M. J. Davis, «Single sample scoring of molecular phenotypes», *BMC Bioinformatics*, vol. 19, n.º 1, p. 404, nov. 2018, doi: 10.1186/s12859-018-2435-4.
- [27] A. R. Abbas, K. Wolslegel, D. Seshasayee, Z. Modrusan, y H. F. Clark, «Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus», *PLOS ONE*, vol. 4, n.º 7, p. e6098, jul. 2009, doi: 10.1371/journal.pone.0006098.
- [28] P. Virtanen *et al.*, «SciPy 1.0: fundamental algorithms for scientific computing in Python», *Nat. Methods*, vol. 17, n.º 3, pp. 261-272, mar. 2020, doi: 10.1038/s41592-019-0686-2.
- [29] S. Pölsterl, «scikit-survival: a library for time-to-event analysis built on top of scikit-learn», *J. Mach. Learn. Res.*, vol. 21, n.º 1, p. 212:8747-212:8752, ene. 2020.
- [30] C. Davidson-Pilon, «lifelines: survival analysis in Python», *J. Open Source Softw.*, vol. 4, n.º 40, p. 1317, ago. 2019, doi: 10.21105/joss.01317.
- [31] Z. Tang, C. Li, B. Kang, G. Gao, C. Li, y Z. Zhang, «GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses», *Nucleic Acids Res.*, vol. 45, n.º W1, pp. W98-W102, jul. 2017, doi: 10.1093/nar/gkx247.

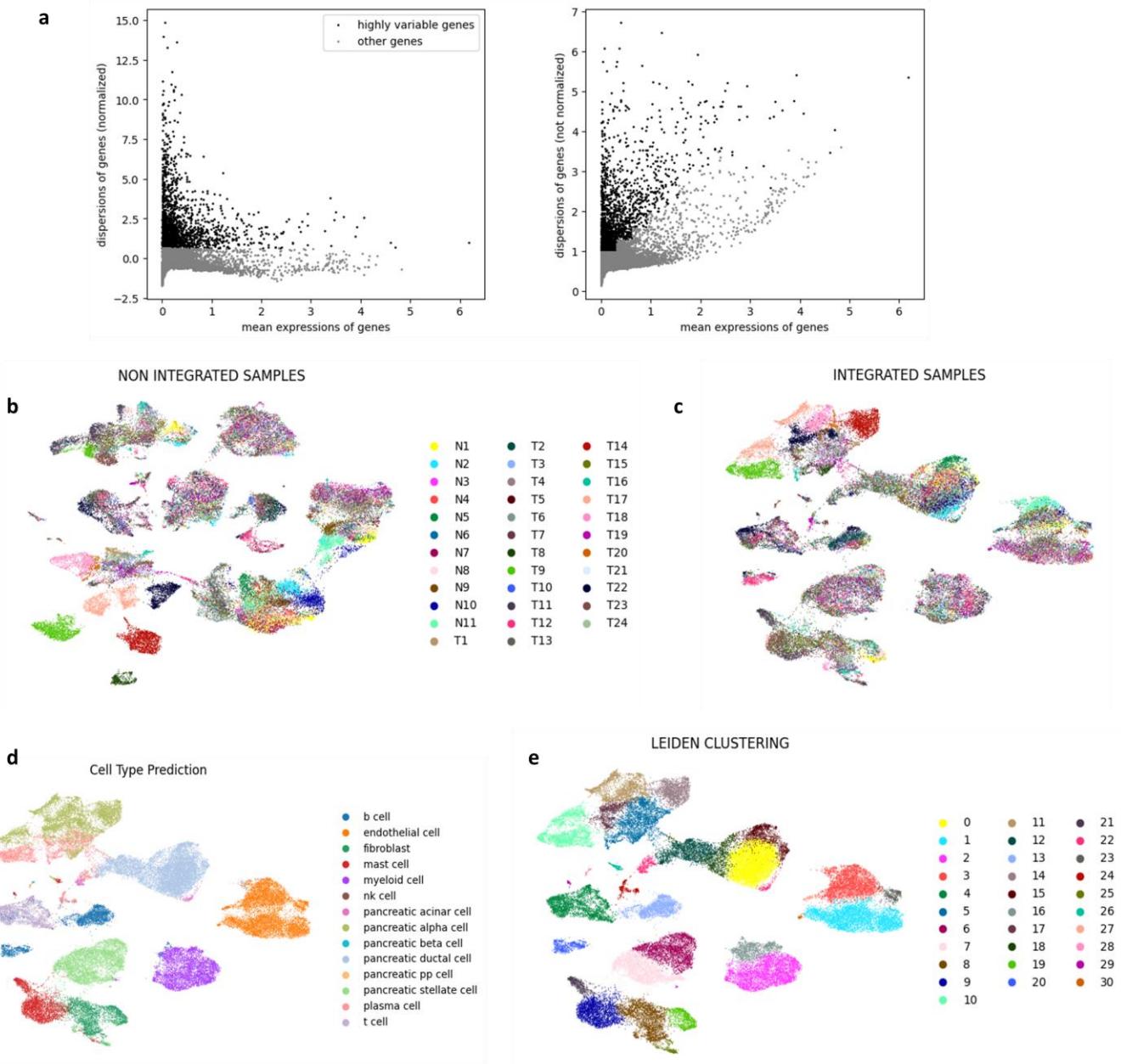
## ANEXO II: FIGURAS COMPLEMENTARIAS



**Figura 1: Análisis de redes de genes co-expresados, WGCNA.** Clustering jerárquico realizado en el preprocesamiento para la detección de outliers (**1.a**). Selección del “soft power” en base al  $R^2$  del ajuste a una red libre de escala (**1.b**). Obtención de módulos de genes co-expresados (**1.c**).

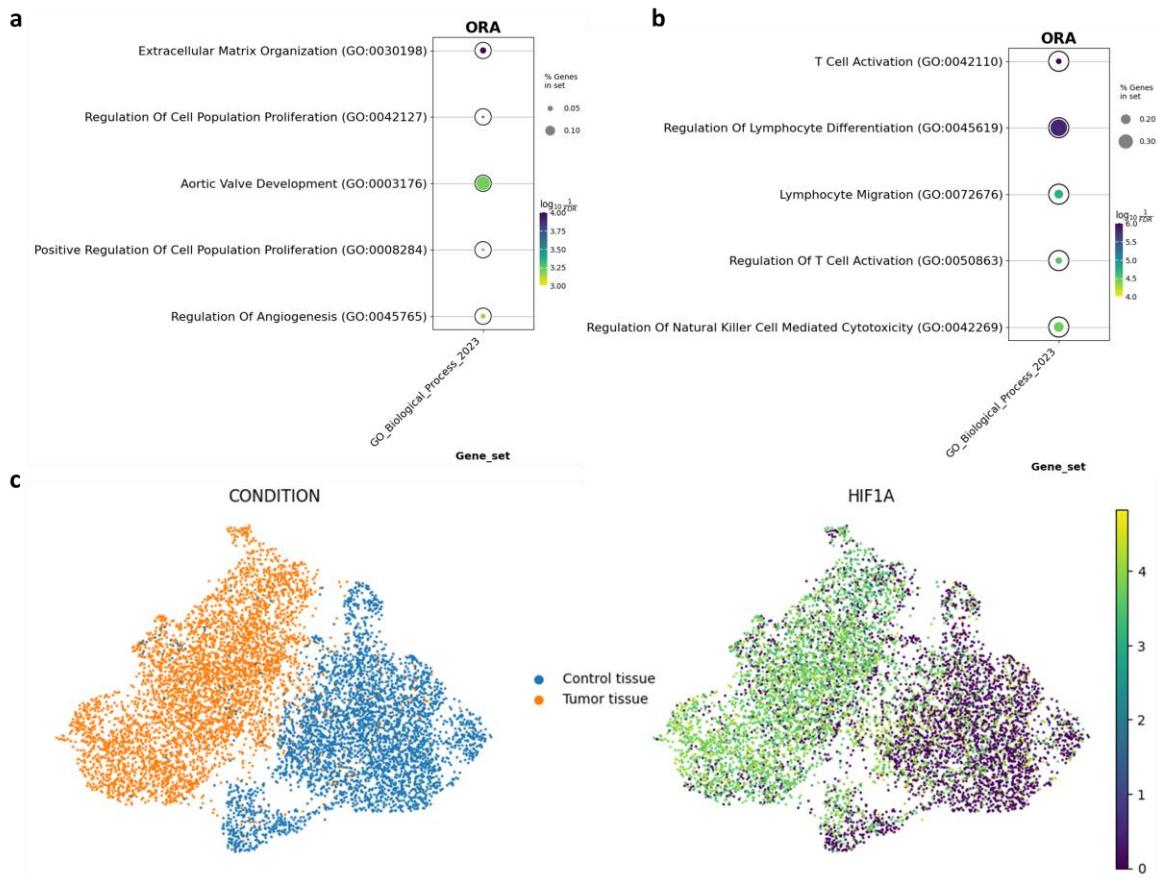


**Figura 2: ORA de módulos WGCNA y deconvolución del perfil de infiltración inmune mediante DeconRNASeq en el dataset de TCGA.** Análisis ORA de módulos “brown” (2.a) y “darkgrey” (2.b) se muestran los términos sobrerepresentados en el eje Y, la fuente de la firma molecular en el eje X, el tamaño del punto indica la proporción de genes encontrados en el set y el color la significancia (-log10(FDR)). Boxplot mostrando diferencias significativas (t-test,  $p<0.05$ ) en la frecuencia de cada tipo celular en el clúster 1 ( $n=75$ ) frente al clúster 0 ( $n=66$ ) (2.c). Heatmap mostrando la frecuencia de cada tipo celular por muestra estimada por DeconRNASeq ordenando en base a un clustering jerárquico, en la parte superior se muestra la pertenencia al clúster gaussiano (2.d)

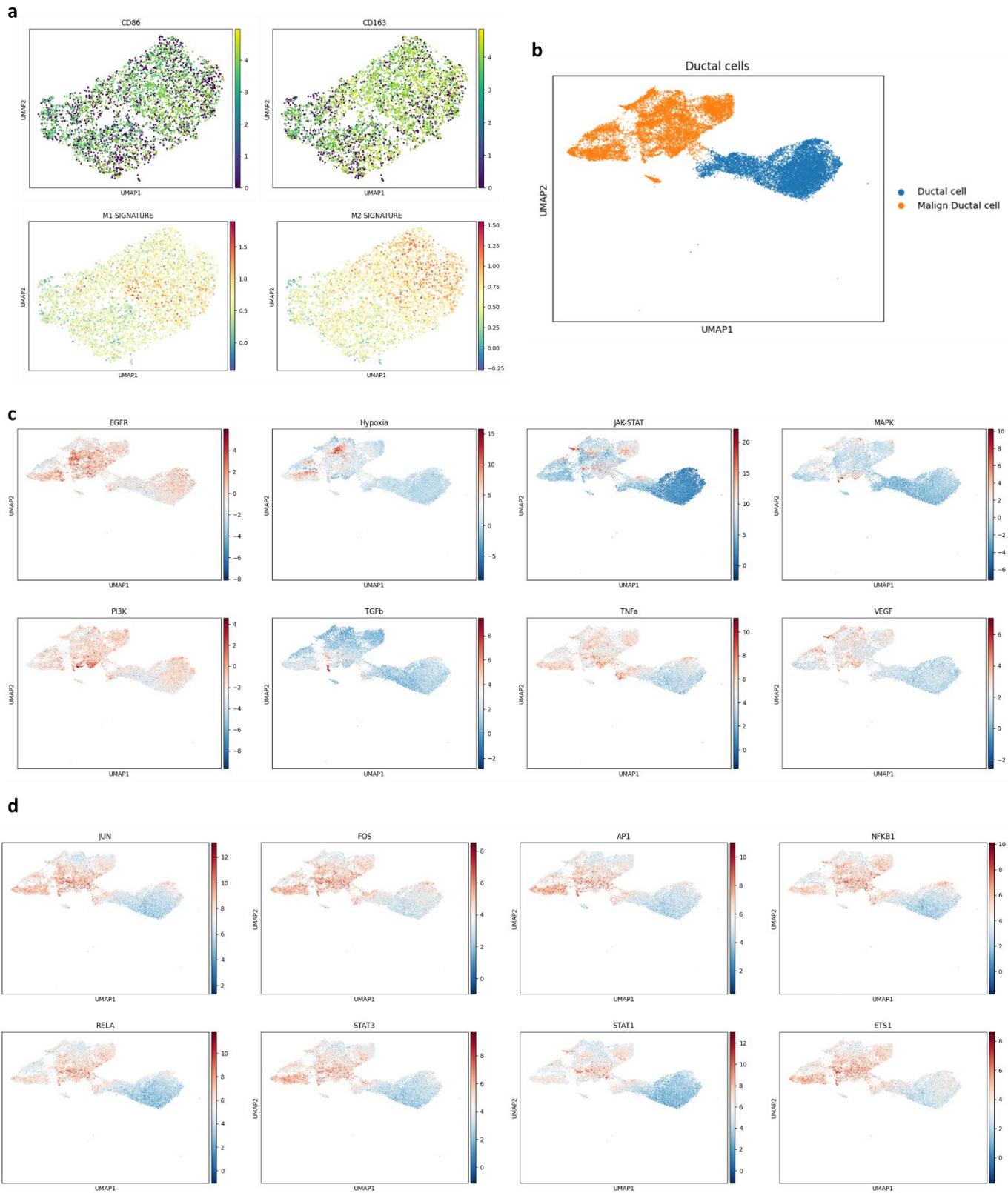


**Figura 3: Preprocesamiento, integración y anotación de tipos celulares en el análisis de sc-RNA-Seq.**

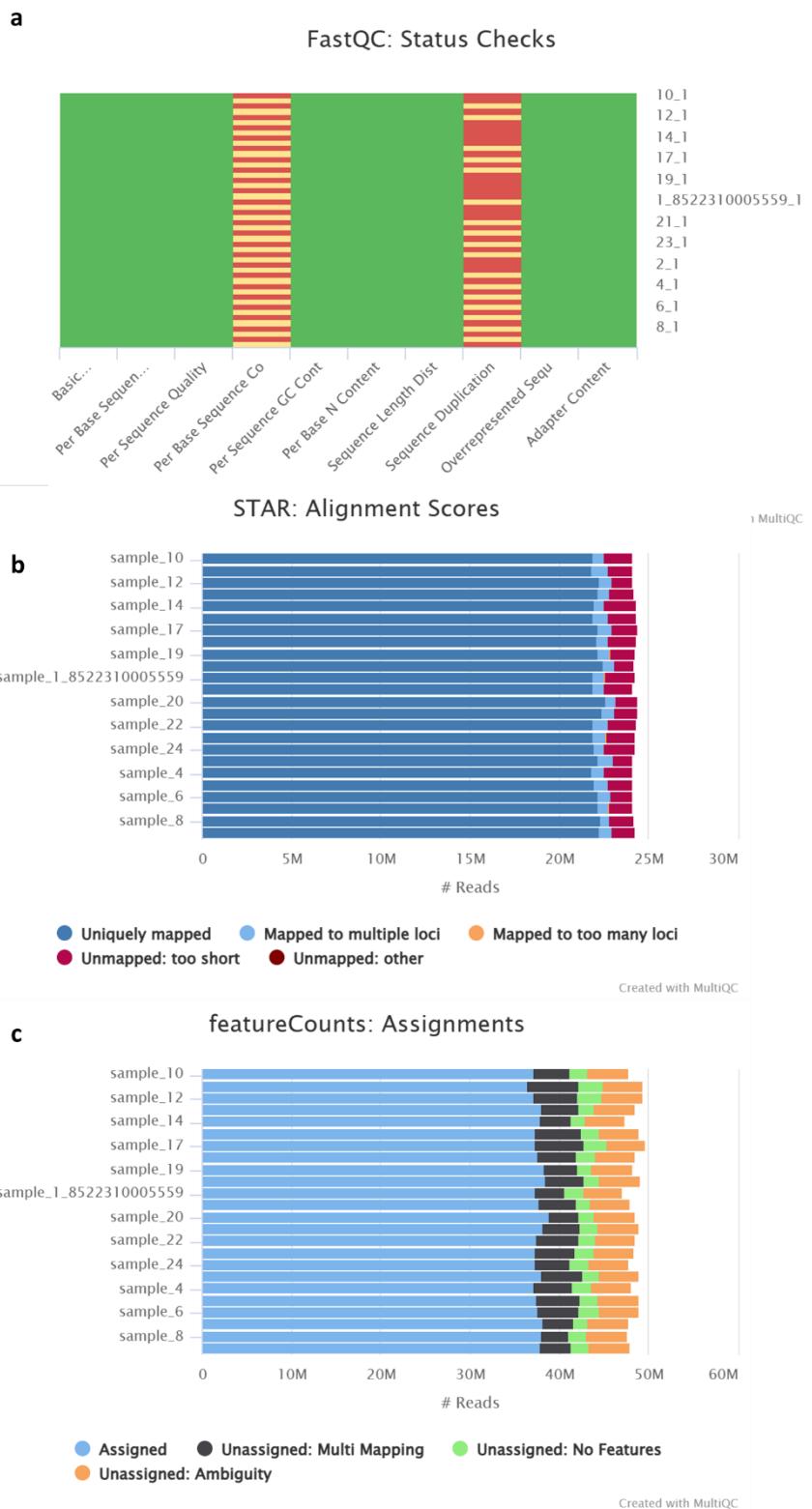
Selección de 2000 genes más variables (en negro) representando la expresión media en función de la dispersión de los genes antes y después de normalizar (**3.a**). UMAPs representando la distribución de las muestras antes (**3.b**) y después de la integración (**3.c**), el tipo celular según la anotación por transferencia de etiquetas (**3.d**) y los clústeres obtenidos por Leiden (resolución 0.7, **3.e**).



**Figura 4: Análisis de funcional y expresión de HIF1A en tejido endotelial tumoral y control.** Análisis ORA de genes diferencialmente expresados, sobre expresados (**4.a**) y regulados a la baja (**4.b**) en células endoteliales tumorales frente a no tumorales. UMAP mostrando las células endoteliales del dataset de sc-RNA-Seq coloreadas por condición (**4.c**) o por valor de expresión normalizada (profundidad 10000,  $\log_{10}(\text{expresión} + 1)$ ) del gen HIF1A.

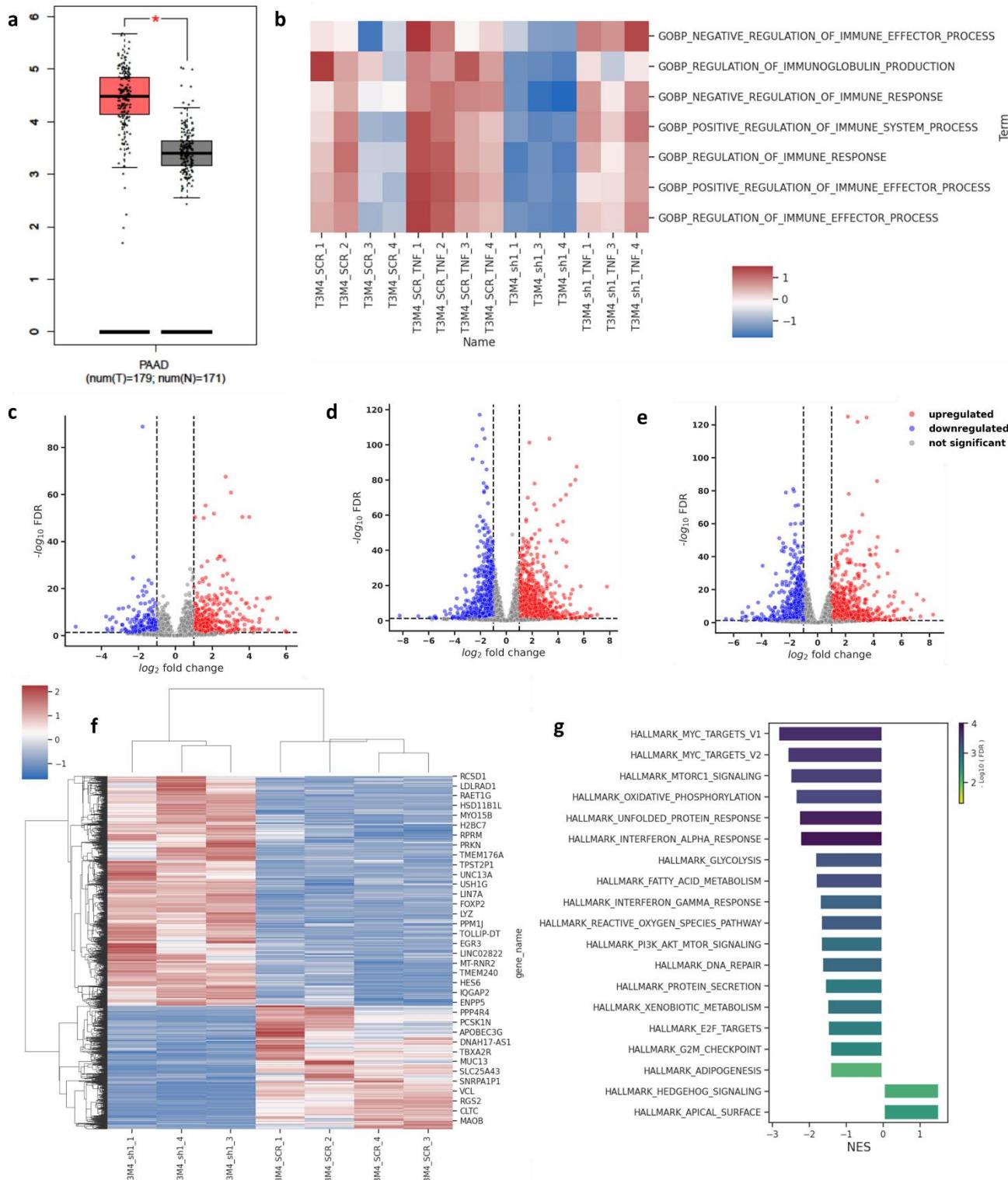


**Figura 5: Firmas de polarización de macrófagos y análisis de actividad de vías moleculares y factores de transcripción en células ductales.** UMAPs de macrófagos, se representa la expresión normalizada (profundidad 10000,  $\log_{10}(\text{expresión}+1)$ ) de marcadores (arriba) y puntuaciones para firmas (abajo) de macrófagos M1 y M2 (5.a). UMAP indicando la malignidad o no de las células ductales (5.b). UMAPs representando el valor de actividad de rutas biológicas (5.c) y factores de transcripción (5.d) estimada por decoupleR.

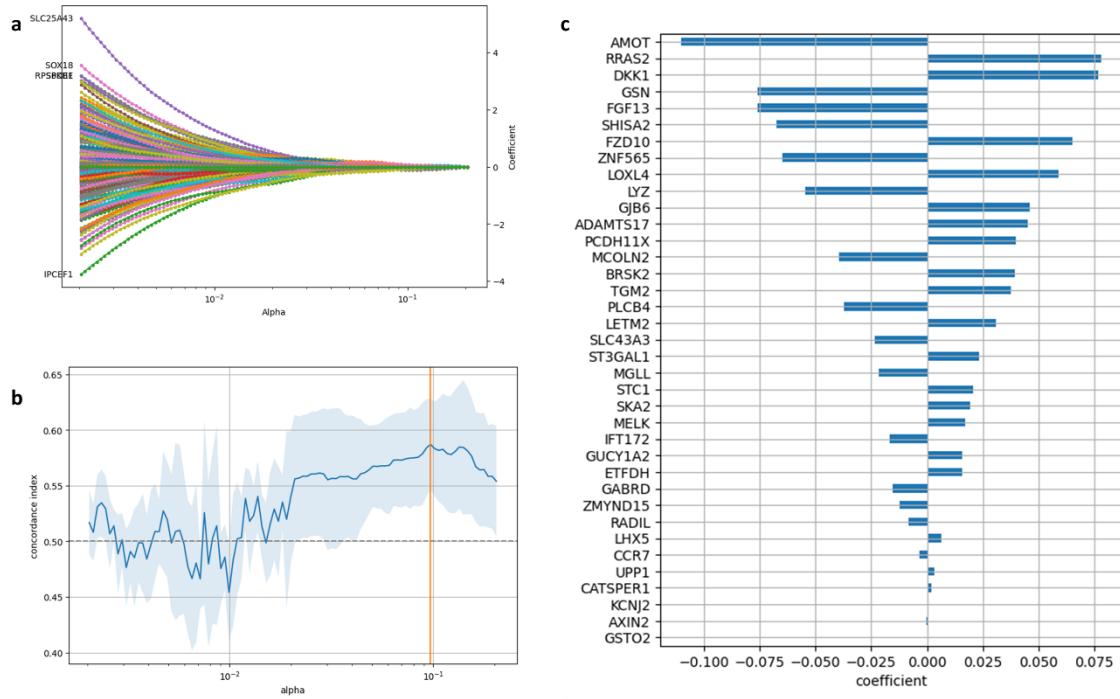


**Figura 6: Control de calidad, alineamiento y la cuantificación del análisis de RNA-Seq por MultiQC.**

Se muestra un resumen de los parámetros de calidad de FASTQC (6.a), de el alineamiento con STAR (6.b) y de la asignación de lecturas a transcritos con featurecounts (6.c)

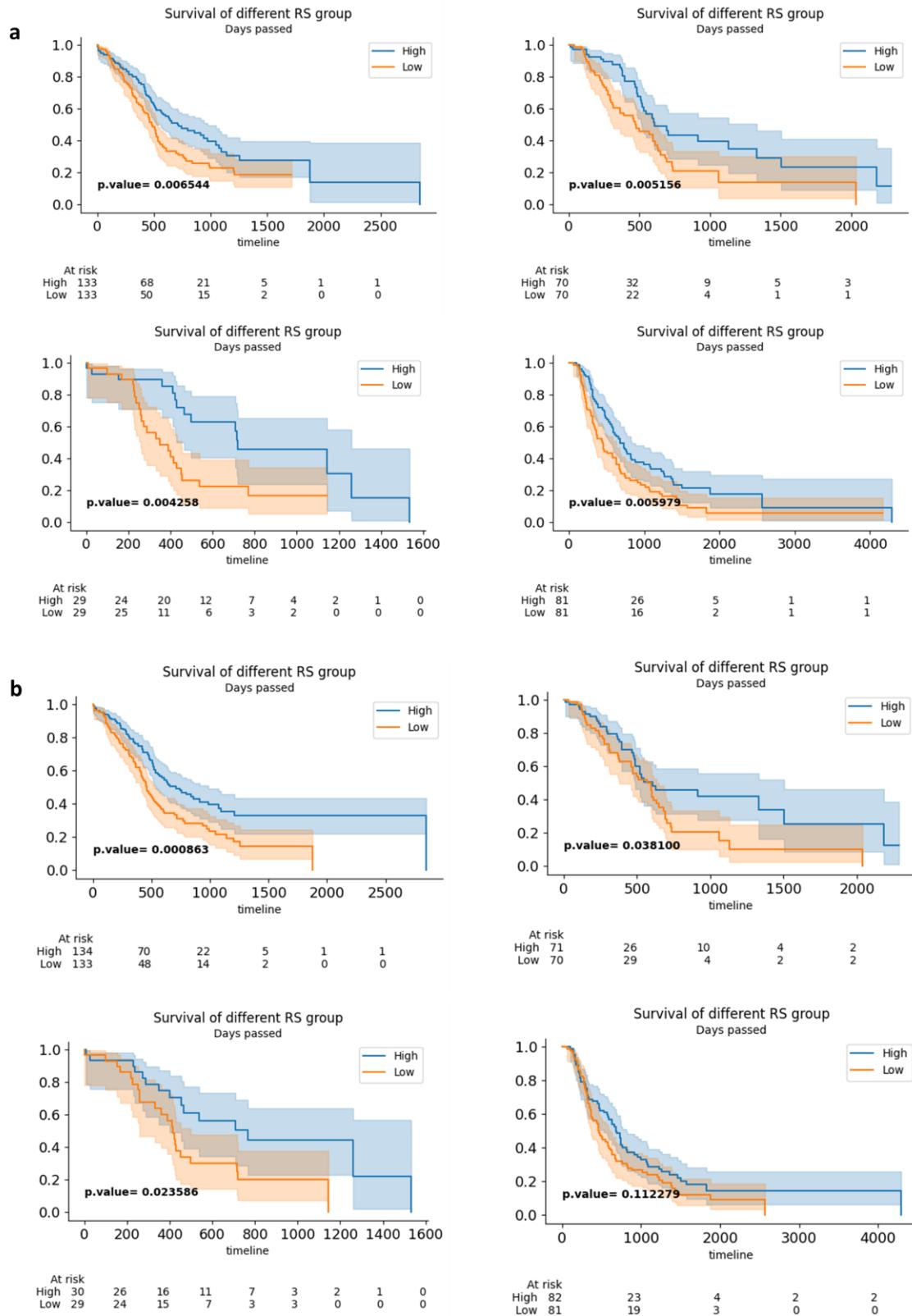


**Figura 7: Análisis de expresión diferencial y funcional del experimento de RNA-Seq.** Boxplot de expresión de BPTF en muestras tumorales frente a normales de pacientes de ADP (GEPIA) (**7.a**). Heatmap puntuación de enriquecimiento de cada muestra para firmas de regulación del sistema inmune (**7.b**). Volcano plots mostrando genes diferencialmente expresados ( $| \text{Log2 Fold Change} | > 1$  &  $\text{FDR} < 0.05$ ) en los contrastes: shSCR TNF $\alpha$ + vs shSCR TNF $\alpha$ - (**7.c**), shBPTF(sh1) TNF $\alpha$ + vs shSCR TNF $\alpha$ + (**7.e**) y sh1 TNF $\alpha$ - vs shSCR TNF $\alpha$ - (**7.d**), heatmap mostrando el valor de expresión normalizada por pyDeseq2 de los DEGs (**7.f**) y un gráfico de barras mostrando el NES para cada firma (coloreados por  $-\log_{10}(\text{FDR})$ ) (**7.g**) de este último contraste.



**Figura 8: Elección de parámetro de regularización y filtrado de genes mediante regresión Cox Lasso.**

Coeficientes asociados a los genes obtenidos por regresión Cox Lasso en el dataset de entrenamiento (ICGC-AU2) para todos los valores testados del parámetro de regularización alfa (**8.a**). Elección del parámetro alfa (línea vertical naranja) que maximiza el índice de concordancia obtenido por el modelo Cox Lasso (**8.b**). Representación de genes y coeficientes seleccionados por el modelo Cox Lasso (**8.c**).



**Figura 9: Curvas de Kaplan-Meier de los pacientes de las cuatro cohortes estratificadas por la expresión de LYZ y FGF13.** Curvas de Kaplan-Meier de grupos de alto (azul) y bajo (naranja) nivel de expresión de LYZ (**9.a**) y FGF13 (**9.b**) en los sets de datos: ICGC-AU2, (arriba izquierda, n=265), TCGA (arriba derecha, n=141), ICGC-AU (abajo izquierda, n=59) e ICGC-CA (abajo derecha, n=163).