

**Beyond Prediction Accuracy:
A Multi-Omic Cross-Validation Framework for feature selection**

Blazquez AC¹, Carballido JA²

¹INIBIBB, CONICET-UNS y Dpto. BByF, UNS, Bahía Blanca, Argentina.

² ICIC, CONICET-UNS y DCIC, UNS, Bahía Blanca, Argentina.

The quality of feature selection methods in single-cell transcriptomic data is typically evaluated using classification metrics such as accuracy or area under the curve. However, high predictive accuracy does not necessarily guarantee biological relevance. In this work, we aim to find the best strategy for selecting genes that are relevant aimed at distinguishing between control and disease. To this end, we use a multi-omic approach that first analyzes RNA abundance data to select genes, and then considers chromatin accessibility data, an independent molecular modality, to perform cross-validation. We compared Differential Expression (DEG), Random Forest, and LASSO, applying them to a murine model of invasive ductal carcinoma progression. In this assessment, DEG demonstrated superior multi-omic validation in terms of correlation. Furthermore, clinical applicability in human patients was validated, where only DEG-derived signatures significantly predicted survival. Thus, an ultra-compact 5-gene signature with triple validation (transcriptomic-epigenetic-clinical) that outperforms several commercial signatures across multiple dimensions was found. Our results demonstrate that differential expression, with multi-omic and clinical validation, can identify biomarkers with good prognostic value, establishing a generalizable paradigm for multi-omic data analysis.

Keywords: feature selection; gene expression; multi-omic data; scRNA-seq; scATAC-seq

Más allá de la Precisión Predictiva:

Un Framework de Validación Cruzada Multi-Ómica para Selección de Características

La calidad de los métodos de selección de características en datos transcriptómicos de células individuales es típicamente evaluada mediante métricas de clasificación como precisión o área bajo la curva. Sin embargo, una alta precisión predictiva no necesariamente garantiza relevancia biológica. En este trabajo queremos encontrar la mejor estrategia de selección de genes que sean relevantes para distinguir entre control y enfermedad. Para tal fin usamos un enfoque multi-ómico que en un primer paso analiza datos de abundancia de RNA para seleccionar genes, y luego considera datos de accesibilidad de cromatina, una modalidad molecular independiente, para realizar la validación cruzada. Comparamos expresión diferencial (DEG), Random Forest y LASSO, y lo aplicamos a un modelo murino de progresión de carcinoma ductal invasivo. En este punto, DEG demostró superior validación multi-ómica en términos de correlación y estabilidad de resultados. Asimismo, validamos la aplicabilidad clínica en pacientes humanos, donde solamente las firmas derivadas de DEG predijeron supervivencia significativamente. De esta forma, identificamos una firma ultra-compacta de 5 genes con triple validación, transcriptómica-epigenética-clínica, que supera en varias dimensiones a varias firmas comerciales. Nuestros resultados demuestran que la expresión diferencial, con una validación multi-ómica y clínica, puede encontrar biomarcadores con buen valor pronóstico, estableciendo un paradigma generalizable para análisis de datos multi-ómicos.

Palabras clave: selección de características; expresión génica; datos multi-ómicos; scRNA-seq; scATAC-seq

1. INTRODUCCIÓN

1.1 Biomarcadores en cáncer de mama

El cáncer de mama es la neoplasia más frecuente en mujeres a nivel mundial, con más de 2 millones de nuevos casos diagnosticados anualmente (Sung et al., 2021). A pesar de los avances en los tratamientos, la heterogeneidad tumoral y la progresión a estadios avanzados siguen representando desafíos importantes para el pronóstico y manejo clínico. La identificación de biomarcadores moleculares que permitan estratificar pacientes según riesgo de progresión es crucial para implementar medicina de precisión. Específicamente, conocer qué genes cambian, y cómo, durante la progresión tumoral permite desarrollar tests moleculares que, al medir la expresión de estos genes en una biopsia, clasifican pacientes según su perfil molecular. Esta estratificación molecular ha demostrado ayudar a mejorar decisiones terapéuticas en hasta el 30% de los casos (Cardoso et al., 2019; Paik et al., 2004).

En las últimas dos décadas, se han desarrollado múltiples firmas génicas comerciales para pronóstico de cáncer de mama, como *Oncotype DX* (21 genes), *MammaPrint* (70 genes) y *PAM50* (50 genes) (Paik et al., 2004; van 't Veer et al., 2002). Sin embargo, estas firmas fueron derivadas principalmente de datos transcriptómicos *bulk* (de tejido completo) sin considerar la heterogeneidad celular ni validación a nivel epigenético. Además, el número elevado de genes las hace costosas de implementar en contextos con recursos limitados.

1.2 Expresión génica y tecnologías ómicas

El genoma de una célula contiene la información completa en forma de ADN, con aproximadamente 20000 genes en humanos. Sin embargo, no todos los genes están activos simultáneamente. La expresión génica es el proceso por el cual, en un determinado momento, la información de un gen se utiliza para sintetizar una proteína funcional, siguiendo el dogma central de la biología molecular. Un gen "expresado" significa que se está produciendo el ARN mensajero (mRNA) correspondiente a ese gen, el cual posteriormente será traducido a proteína. La cantidad de mRNA presente en una célula refleja entonces el nivel de actividad de ese gen. Por ejemplo, en una célula tumoral proliferativa, genes relacionados con división celular estarán altamente expresados, mientras que genes supresores de tumores pueden estar silenciados.

RNA-seq es una de las tecnologías existentes que permite cuantificar la abundancia de mRNA en una muestra. Técnicamente, extrae el ARN total, lo convierte a ADN complementario (cDNA) y lo secuencía. El número de lecturas (reads) obtenidas para cada gen es proporcional a su nivel de expresión (Wang et al., 2009). Tradicionalmente, RNA-seq se realizaba solamente en tejido completo (denominado *bulk*), analizando una muestra completa (por ejemplo, un fragmento de tumor) que contiene millones de células de diferentes tipos (células tumorales, fibroblastos, células inmunes). El resultado es un promedio de expresión de todas estas células mezcladas, lo que oculta la heterogeneidad celular. La revolución tecnológica de single-cell RNA-seq (scRNA-seq) permite medir la expresión génica en cada célula individual (Tang et al., 2009). En lugar de un valor promedio del tejido, se obtienen perfiles de expresión de miles a millones de células individuales, revelando subpoblaciones celulares, estados transicionales y heterogeneidad intra-tumoral que son imposibles de ver en datos *bulk*.

Si bien el presente trabajo no se enfoca en la caracterización de subpoblaciones celulares, el uso de datos de célula única ofrece numerosas ventajas metodológicas críticas sobre *bulk* RNA-seq. En primer lugar, nos brinda mayor poder estadístico al operar sobre decenas de miles de observaciones individuales en lugar de pocas réplicas biológicas, aumentando la robustez de los métodos de selección de características. Además, se preservan señales de poblaciones minoritarias que se diluyen en promedios *bulk*. Por último, y principalmente importante sobre todo en este trabajo, tenemos el pareamiento célula-a-célula de RNA y ATAC que nos permite validar la concordancia entre transcripción y accesibilidad a la cromatina en las mismas células individuales, algo imposible en tecnologías *bulk* donde RNA y ATAC se miden en porciones de tejido proveniente de distintas muestras (Tang et al., 2009; Cusanovich et al., 2015).

Por otro lado, la expresión génica no solo depende de la secuencia de ADN, sino también de su accesibilidad. El ADN en células eucariotas está empaquetado en la cromatina, donde las regiones "abiertas" (accesibles) permiten que los factores de transcripción se unan y activen genes, mientras que las regiones "cerradas" (inaccesibles) están silenciadas (Klemm et al., 2019). ATAC-seq (*Assay for Transposase-Accessible Chromatin using sequencing*) es una tecnología que identifica regiones del genoma con cromatina accesible (Buenrostro et al., 2013). Utiliza una enzima (llamada transposasa o TN5) que solo puede insertarse en regiones de ADN abiertas. Cuantas más inserciones se detectan en una región promotora de un gen, mayor es su accesibilidad y, potencialmente, su capacidad de ser expresado. Análogamente a RNA-seq, single-cell ATAC-seq mide accesibilidad de cromatina a nivel de célula individual, revelando estados regulatorios específicos de subpoblaciones celulares (Cusanovich et al., 2015).

De esta manera, RNA-seq y ATAC-seq miden aspectos diferentes pero complementarios de la regulación génica: RNA-seq mide cuánto mRNA se está produciendo, mientras que ATAC-seq analiza si la maquinaria transcripcional (la transposasa) tiene acceso al gen. La lógica biológica es que si un gen aumenta su expresión genuinamente (no por ruido técnico), esperamos que su región promotora esté más accesible. Esta concordancia entre modalidades es un indicador de cambio biológico robusto.

1.3 Selección de características en datos de alta dimensionalidad

En el análisis de datos transcriptómicos, enfrentamos un problema clásico de alta dimensionalidad: en humanos se miden aproximadamente 20000 genes, pero solo un pequeño subconjunto es relevante para distinguir estadios biológicos (por ejemplo, tejido sano vs. tumor). La selección de características (genes) es el proceso de identificar este subconjunto informativo. Los métodos más tradicionales incluyen los enfoques estadísticos de análisis de expresión diferencial (DEG), que identifica genes con cambios significativos entre condiciones (Love et al., 2014), y los de aprendizaje automático como Random Forest (RF) y LASSO, que seleccionan características que maximizan precisión predictiva (Breiman, 2001; Tibshirani, 1996).

La métrica estándar para evaluar estos métodos es la precisión de clasificación, que mide cuánto de bien los genes seleccionados predicen la clase, por ejemplo sano vs. enfermo. Sin embargo, considerar solo alta precisión puede ser engañoso ya que puede haber problemas con el sobre ajuste, los artefactos técnicos y la falta de validación en el contexto de aplicación, por ejemplo si tienen relevancia biológica..

En cuanto a lo clínico, desarrollar un test diagnóstico basado en biomarcadores tiene costos significativos: validación experimental, ensayos, implementación en hospitales. Si los biomarcadores iniciales son artefactos técnicos (falsos positivos de del método de selección de características utilizado), se desperdicia tiempo de investigación y recursos económicos. Necesitamos métodos computacionales que, desde el inicio del proceso, filtren ruido y prioricen señal biológica robusta antes de llegar a la etapa de inversión en validación experimental.

1.4 Validación multi-ómica como filtro de calidad

En nuestra hipótesis de trabajo proponemos que la validación cruzada multi-ómica es superior a métricas de clasificación tradicionales para identificar características biológicamente relevantes. Específicamente, partimos de la idea de que los genes cuyos cambios de expresión (scRNA-seq) están respaldados por cambios coordinados en accesibilidad de cromatina (scATAC-seq) son más robustos biológicamente y tienen mayor valor pronóstico clínico que aquellos seleccionados solo por alta precisión predictiva. Nuestra estrategia se puede resumir en las siguientes etapas, mostradas de manera esquemática en la Figura 1:

1. Descubrimiento (ratón, scRNA-seq): Aplicar tres métodos de selección de características (DEG, RF, LASSO) a un modelo murino de progresión de cáncer de mama.
2. Validación ortogonal (ratón, scATAC-seq): Evaluar qué método selecciona genes con mayor concordancia entre cambios de expresión y de accesibilidad de cromatina.
3. Validación traslacional (humano, TCGA-BRCA): Examinar el valor pronóstico de las firmas génicas en datos clínicos de supervivencia de pacientes con cáncer de mama.



Fig. 1. Desde transcripción hasta accesibilidad de cromatina y análisis de supervivencia.

Este diseño permite responder cuál de los métodos de selección de características identifica biomarcadores que no solo predicen bien en el conjunto de datos original, sino que también tienen respaldo en los cambios coordinados a nivel transcripcional y de regulación de cromatina, y se traducen a valor pronóstico clínico en humanos.

De esta forma, el principal objetivo de nuestro trabajo fue comparar tres métodos de selección de características, expresión diferencial (DEG), Random Forest (RF) y LASSO en términos de precisión de clasificación, validación multi-ómica RNA-ATAC, y valor pronóstico en supervivencia humana. En este contexto quisimos establecer una estrategia de validación multi-ómica con el fin de demostrar que la concordancia RNA-ATAC es un criterio robusto para evaluar la relevancia biológica de los genes seleccionados. También se propuso validar la aplicabilidad clínica,

demostrando que los genes descubiertos en el modelo murino tienen valor pronóstico en pacientes humanos (traslación entre especies), mediante análisis de supervivencia.

Como resultado adicional al análisis comparativo, identificamos una firma génica ultra-compacta, compuesta de un conjunto mínimo de genes con triple análisis (transcriptómica, epigenética, clínica) que en términos de número de genes y validación multi-ómica sugiere potencial competitivo frente a varias firmas comerciales.

2. TRABAJOS RELACIONADOS

La integración de datos multi-ómicos para identificación de biomarcadores ha sido abordada desde distintos ángulos metodológicos. MOFA (Argelaguet et al., 2018) factoriza conjuntamente múltiples modalidades ómicas para extraer fuentes latentes de variación compartida y específica de cada capa. Aunque el *framework* permite incorporar RNA-seq y ATAC-seq, su objetivo es la descomposición de varianza y no la evaluación comparativa de métodos de selección de características; tampoco incluye validación de valor pronóstico en cohortes clínicas.

Seurat v3 (Stuart et al., 2019) introdujo integración multi-modal de datos de célula única mediante anclas de transferencia, facilitando el alineamiento de perfiles RNA y ATAC en espacios latentes compartidos. Su validación se circunscribe al dominio computacional - precisión de integración celular, correlación entre modalidades - sin extenderse a la evaluación del rendimiento clínico de los biomarcadores resultantes. Nuestro enfoque difiere al utilizar la concordancia RNA-ATAC no como métrica de integración técnica, sino como criterio de calidad para comparar métodos de selección de características.

Curtis et al. (2012) realizaron una caracterización multi-ómica integrativa de 2.000 tumores de mama combinando expresión génica, variación en número de copias y datos clínicos, definiendo 10 subgrupos moleculares con pronósticos diferenciados. Su contribución establece que la estratificación multi-ómica supera en resolución pronóstica a la clasificación basada únicamente en receptores hormonales. Sin embargo, el análisis se apoya en clustering no supervisado sobre datos bulk; no compara métodos de selección de características ni incorpora datos de accesibilidad de cromatina, impidiendo evaluar concordancia transcripcional-epigenética a nivel de célula individual.

En conjunto, los trabajos previos abordan la integración multi-ómica o el descubrimiento de biomarcadores de forma independiente. El presente trabajo los conecta explícitamente: la concordancia RNA-ATAC a la resolución de célula única actúa como filtro intermedio entre selección computacional de características y validación pronóstica en cohorte humana, componiendo un pipeline de tres etapas que, hasta donde conocemos, no ha sido evaluado comparativamente en este contexto.

3. MATERIALES Y MÉTODOS

3.1. Conjunto de datos

Se utilizó el conjunto de datos público GSE212707, que representa un estudio avanzado de genómica funcional y utiliza la tecnología *10x Genomics Multiome* para capturar, de manera simultánea, el transcriptoma y el perfil de accesibilidad de la cromatina en células individuales. Este enfoque integrador permite vincular directamente la

expresión génica con los elementos regulatorios del ADN en el mismo estado celular, facilitando la identificación de mecanismos epigenéticos que impulsan la progresión tumoral. El estudio se centra principalmente en la caracterización del microambiente del cáncer de mama.

Experimentalmente, el diseño incluye un robusto modelo de ratón compuesto por nueve muestras distribuidas equitativamente en tres estadios biológicos: tres réplicas de tejido mamario sano que actúan como control, tres muestras de estadios tempranos caracterizados por hiperplasia y tres muestras de estadios avanzados que presentan adenocarcinomas avanzados. Para garantizar la relevancia clínica de estos hallazgos, el conjunto de datos incorpora además tres muestras de pacientes humanas diagnosticadas con carcinoma ductal invasivo. En este trabajo, por cuestiones de limitaciones técnicas, restringimos el análisis al modelo murino, comparando dos muestras de los estadios control contra dos muestras del estadio avanzado.

3.2. Procesamiento de los datos

Como primer paso, los datos crudos de scRNA-seq fueron procesados utilizando el paquete *Seurat* v4 en R. De cada una de las 4 muestras de ratón (2 de control y 2 de avanzado), se crearon objetos *Seurat* individuales a partir de las matrices de cuentas generadas por Cell Ranger. Se aplicaron filtros de control de calidad para retener células con entre 500 y 7000 genes detectados, 1000 a 50000 UMIs totales, y menos del 25% de lecturas mitocondriales. Posteriormente, se realizó la normalización mediante el método *LogNormalize*, una selección de los 3000 genes más variables, el escalado de datos, y la reducción dimensional mediante análisis de componentes principales (PCA). Las muestras fueron integradas utilizando el método de anclaje de *Seurat* para corregir efectos de lote.

Por otro lado, los datos de scATAC-seq fueron procesados con el paquete *Signac* en R. Para cada muestra, se cargaron las matrices de picos (regiones del ADN que están accesibles) generadas por Cell Ranger ATAC, creando objetos *ChromatinAssay*. La accesibilidad de cromatina en regiones promotoras ($TSS \pm 2kb$) de cada gen fue calculada sumando las cuentas normalizadas de todos los picos que intersectan con dicha región, generando un score de accesibilidad por gen y por estadio. Si un gen tiene en promedio un pico alto en el estadio avanzado pero no en el control, significa que en el estadio avanzado hubo muchísimos más fragmentos cortados por TN5 en esa región, lo que indica que el ADN se abrió para permitir la transcripción de ese gen.

La matriz resultante de expresión normalizada (log-counts) de 54299 células $\times$ 28000 genes fue utilizada para el análisis de expresión diferencial entre estadios mediante el test de Wilcoxon (implementado en *FindMarkers* de *Seurat*), y para la aplicación de RF y LASSO. De la aplicación de estas tres técnicas se obtienen tres conjuntos de genes, los seleccionados por cada una.

Brevemente, en este contexto el test de Wilcoxon constituye un estándar para el análisis de expresión diferencial en datos de célula única debido a su naturaleza no paramétrica, lo que le permite comparar las distribuciones de expresión entre los grupos control y avanzado sin asumir una distribución normal, manejando así con eficacia la elevada proporción de ceros y la dispersión característica del scRNA-seq. Por otro lado, RF aporta una perspectiva de aprendizaje supervisado mediante la construcción de un conjunto de múltiples árboles de decisión que clasifican las células según su estado biológico; este método permite identificar genes clave a través del cálculo de la

importancia de las variables, lo que ayuda a capturar interacciones no lineales y relaciones complejas entre marcadores que los test estadísticos tradicionales podrían ignorar al evaluar cada gen de forma aislada. Complementariamente, la regresión LASSO usa una técnica de regularización denominada penalización L1, la cual impone una restricción sobre la magnitud de los coeficientes del modelo, provocando que los pesos de los genes menos informativos se reduzcan exactamente a cero. Este proceso realiza una selección automática de características, generando un modelo disperso y altamente interpretable que retiene únicamente aquellos genes con el mayor poder predictivo para diferenciar el tejido sano del tumoral en etapas avanzadas.

En este punto, para evaluar la concordancia entre expresión génica (RNA) y accesibilidad de cromatina (ATAC), calculamos los cambios entre estadios (log2 fold-change) de forma independiente para cada gen de cada grupo (los seleccionados por DEG, por RF y por LASSO). Cabe destacar que de los 100 genes retornados por cada método, no todos tenían datos aceptables de accesibilidad. Por esta razón, la cantidad final de genes que pasa a la siguiente etapa fue menor a 100 en los 3 casos. Para RNA-seq, se computó el promedio de expresión normalizada por gen en cada estadio (control y avanzado) y se calculó log2FC. Paralelamente, para ATAC-seq, se calculó el promedio de accesibilidad de promotores por gen en cada estadio, y se computaron los mismos log2FC.

La validación cruzada consistió en correlacionar con Pearson los vectores de log2FC_RNA y log2FC_ATAC de cada gen. Genes con cambios concordantes en ambas modalidades fueron clasificados como validados epigenéticamente. Esta métrica de concordancia RNA-ATAC fue utilizada para comparar la robustez biológica de las características seleccionadas por DEG, RF y LASSO, bajo la hipótesis de que genes con cambios significativos y coordinados en expresión y accesibilidad representan señal biológica más confiable que aquellos con cambios solo detectables en una modalidad.

Por último, la validación clínica se realizó mediante un análisis de supervivencia, buscando en primera instancia los genes ortólogos en humanos para los genes de ratón sugeridos por el método que presentó la mejor performance. Para esto se usaron datos de TCGA-BRCA, donde se obtienen muestras de 1079 pacientes y 153 eventos. Se realizó el análisis de supervivencia considerando distintos tamaños de grupos de genes, para verificar cuál de ellos podría considerarse una firma génica en este tipo de cáncer. Una vez establecido que con 5 genes se tenía una firma significativa, se comparó con las firmas de 5 genes sugeridas por los otros dos métodos.

4. RESULTADOS Y DISCUSIÓN

4.1 Selección de características en datos scRNA-seq, validado en scATAC-seq

A partir de la matriz de expresión normalizada de 54299 células $\times$ 28000 genes, aplicamos tres métodos estándar de selección de características para identificar genes diferencialmente expresados entre tejido control y tumor avanzado: análisis de expresión diferencial (DEG) mediante test de Wilcoxon, RF con importancia de características basada en *Mean Decrease Gini*, y LASSO con regularización L1 y con selección automática de λ óptimo por validación cruzada.

Para evaluar la estabilidad y reproducibilidad de los métodos de selección de características, implementamos un diseño de bootstrap con 30 iteraciones

independientes. En cada iteración, se tomaron muestras aleatoriamente del 80% de las células del dataset scRNA-seq (donde control + avanzado son n=29544 células totales), manteniendo la proporción de cada estadio. Sobre este subconjunto, se aplicaron los tres métodos de selección de características: DEG, RF y LASSO. Cada método seleccionó sus top 100 genes según su criterio específico (p-valor ajustado para DEG, importancia de Gini para RF, y magnitud de coeficientes para LASSO). Este procedimiento de sub-muestreo aleatorio garantiza que los resultados no dependan de la composición específica del conjunto de datos ni de una única corrida, sino que reflejen propiedades robustas de cada método de selección.

A partir de los 100 genes seleccionados por cada método en cada iteración, se identificó cuáles tenían datos de accesibilidad de cromatina válidos en ambos grupos, control y avanzado. Un gen fue considerado "validable" si su región promotora (TSS $\pm$ 2kb) contenía al menos un pico detectado tanto en muestras control como avanzado, permitiendo el cálculo de cambios de accesibilidad entre estadios. A través de las 30 iteraciones, DEG validó en promedio 82,5 genes, RF 86,6 genes, y LASSO 77,6 genes.

Para cada gen validable, se calcularon los cambios de expresión génica y los cambios de accesibilidad de cromatina entre los estadios. El log2FC de expresión se obtuvo calculando el promedio de expresión normalizada (log-counts) de cada gen en control vs. avanzado: $\log2FC_RNA = \log2(\text{expresión_avanzado}/\text{expresión_control})$. Paralelamente, el log2FC de accesibilidad se calculó sumando las señales de todos los picos de ATAC que intersectan con la región promotora del gen (TSS $\pm$ 2kb) en cada grupo, promediando entre células: $\log2FC_ATAC = \log2(\text{accesibilidad_avanzado}/\text{accesibilidad_control})$. Este enfoque captura tanto cambios transcripcionales (cuánto mRNA se produce) como cambios regulatorios epigenéticos (cuán accesible está la cromatina para la maquinaria transcripcional), permitiendo evaluar si ambas modalidades reportan cambios coordinados en los mismos genes.

La métrica de validación multi-ómica consistió en correlacionar los vectores de log2FC_RNA y log2FC_ATAC para los genes seleccionados por cada método, utilizando la correlación de Pearson. Se espera que un método robusto seleccione genes cuyos cambios de expresión están respaldados por cambios coordinados en accesibilidad de cromatina, reflejándose en una alta correlación positiva. La Tabla 1 presenta los resultados promedio de las 30 iteraciones.

Tabla 1: Correlación RNA-ATAC por método en 30 iteraciones

Método	Media r	SD r	Mediana r	p-value vs DEG	Interpretación
DEG	0.729	0.004	0.730	—	Alta correlación, máxima estabilidad
RF	0.726	0.019	0.728	0.452 (n.s.)	Equivalente a DEG, más variable
LASSO	0.609	0.042	0.617	$1.86 \times 10^{-9***}$	Inferior, altamente inestable

Nota: p-values calculados con test de Wilcoxon pareado. n.s. = no significativo

DEG alcanzó una correlación promedio $r=0,729 \pm 0,004$, RF $r=0,726 \pm 0,019$, y LASSO $r=0,609 \pm 0,042$. Los tests estadísticos de Wilcoxon pareados revelaron que DEG y RF son estadísticamente equivalentes ($p=0,452$), ambos significativamente superiores a LASSO ($p=1,86 \times 10^{-9}$ para ambas comparaciones). Sin embargo, DEG exhibió la menor variabilidad entre iteraciones ($SD=0,004$), indicando máxima estabilidad independientemente del sub-muestreo celular, mientras que LASSO mostró alta inestabilidad ($SD=0,042$, 10 veces mayor). En la Figura 2 se observa claramente este comportamiento.

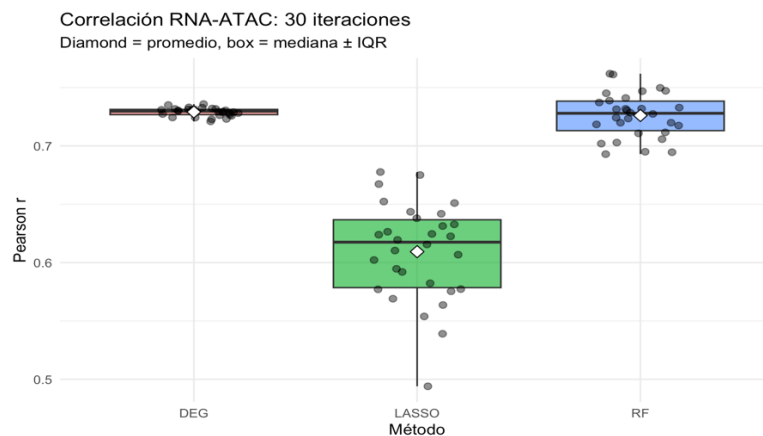


Fig. 2. Comportamiento de los métodos en las 30 corridas

Estos resultados demuestran que el análisis de expresión diferencial es el método más robusto para selección de características con respaldo multi-ómico: no solo alcanza la correlación RNA-ATAC más alta, sino que mantiene esta performance consistentemente a través de múltiples ejecuciones y muestreos de datos, mientras que RF, aunque competitivo en promedio, exhibe mayor variabilidad, y LASSO es tanto inferior en correlación como inestable en su performance. La superioridad de DEG se atribuye a que su criterio estadístico (distribuciones celulares individuales + test no paramétrico) filtra naturalmente artefactos técnicos y prioriza cambios biológicos replicables, mientras que métodos de los métodos de aprendizaje automático están optimizados para mejorar la precisión y pueden capturar patrones predictivos no necesariamente respaldados por regulación epigenética coordinada.

4.2 Validación traslacional: valor pronóstico en cáncer de mama humano

Para evaluar si los genes identificados en el modelo murino tienen relevancia clínica en humanos, realizamos análisis de supervivencia en la cohorte TCGA-BRCA ($n=1079$ pacientes, 151 eventos de muerte). Utilizando los genes seleccionados por DEG en orden de significancia estadística, evaluamos firmas génicas de diferentes tamaños (5, 10, 15, 30, 50 y 100 genes) para identificar el conjunto mínimo con valor pronóstico (Figura 3). Para cada firma, se calculó un score de expresión (promedio de log-counts estandarizado) y se probó su asociación con supervivencia global mediante regresión

de Cox. El análisis se truncó a 15 años de seguimiento para evitar inestabilidad estadística por bajo número de pacientes en riesgo en tiempos tardíos.

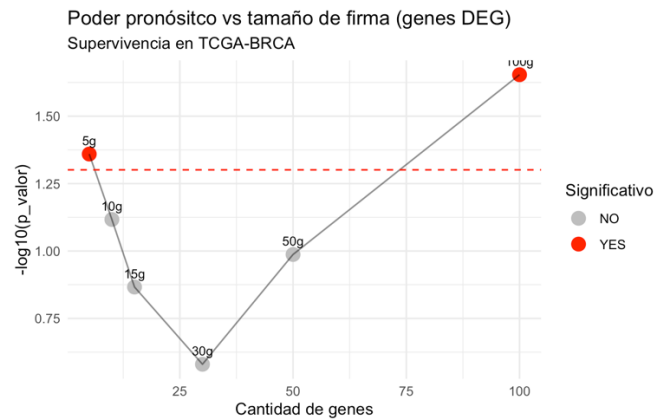


Fig. 3. Poder pronóstico de las firmas encontradas por DEG

Identificamos que una firma ultra-compacta de 5 genes mostró asociación significativa con supervivencia: pacientes con alta expresión de la firma presentaron 18% mayor riesgo de muerte por unidad de desviación estándar ($HR=1,18$ [$1,01-1,38$], $p=0,034$) (Figura 4A). Esta firma incluye los genes SORBS2, CDH1, TRPS1, EPCAM y PTPRK, todos altamente expresados (*UP-regulated*) en tumores avanzados del modelo murino y con cambios coordinados en accesibilidad de cromatina.

Para contrastar la especificidad de la validación multi-ómica, comparamos el desempeño pronóstico de las firmas derivadas de los tres métodos. Se evaluaron los top 5 genes de DEG, RF y LASSO bajo el mismo protocolo de análisis de supervivencia a 15 años (Tabla 2, Figuras 4B-C). Únicamente la firma de DEG alcanzó significancia estadística ($p=0,034$), mientras que RF ($HR=1,04$, $p=0,622$) y LASSO ($HR=1,03$, $p=0,605$) no mostraron asociación con supervivencia.

Tabla 2: Validación clínica, firmas de 5 genes en Supervivencia (TCGA-BRCA, 0-15 años)

Método	Genes	HR	CI	P_value	Significativo
DEG	SORBS2, CDH1, TRPS1, EPCAM, PTPRK	1,18	[1.01-1.38]	0,0341	SI
RF	ARHGAP15, EBF1, PTPRC, TRPS1, NEAT1	1,04	[0.89-1.22]	0,622	NO
LASSO	PTN, SAMD5, MALAT1, NAALAD2, CCBE1	1,03	[0.91-1.18]	0,605	NO

Nota: HR (Hazard Ratio) calculado por desviación estándar de aumento en el score de expresión estandarizado. Análisis de supervivencia global truncado a 15 años de seguimiento. Genes listados en orden de importancia/significancia según cada método.

Es interesante notar que RF comparte un gen con DEG (TRPS1). Sin embargo, en el caso de RF la inclusión de los genes adicionales sin respaldo epigenético robusto diluye

el poder pronóstico. LASSO seleccionó un conjunto completamente diferente enfocado en genes con coeficientes de regresión dispersos, que aunque útiles para la clasificación, no predicen mortalidad tal como demuestra este estudio.

Este resultado confirma que la validación multi-ómica RNA-ATAC no solo identifica genes con cambios transcripcionales consistentes, sino que además prioriza genes con mayor utilidad pronóstica en contextos clínicos humanos. La superioridad de DEG se mantiene a través de las tres capas de validación implementadas: (1) correlación RNA-ATAC en ratón ($r=0,729 \pm 0,004$), (2) estabilidad a través de 30 iteraciones *bootstrap* ($SD=0,004$ vs $0,019/0,042$ en RF/LASSO), y (3) predicción de supervivencia en humanos (único método significativo). Los 5 genes de la firma DEG no solo diferencian estadios control vs. avanzado en ratón, sino que su expresión coordinadamente elevada en tumores humanos identifica pacientes con peor pronóstico, demostrando conservación funcional del programa transcripcional de progresión tumoral entre especies.

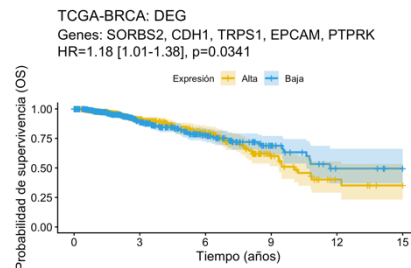


Fig. 4 (A)

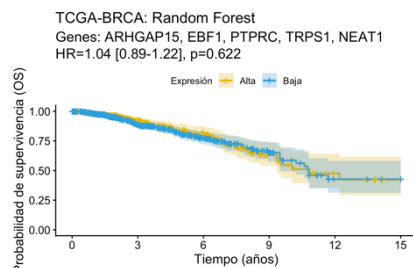


Fig. 4 (B)

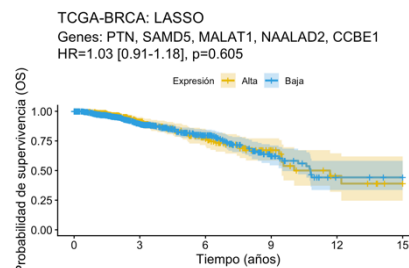


Fig. 4 (C)

Fig. 4. Análisis de supervivencia de las firmas formadas por los 5 top genes de cada método

5. CONCLUSIONES Y TRABAJO FUTURO

5.1 Validación multi-ómica

Este trabajo establece que la validación cruzada multi-ómica, específicamente mediante concordancia RNA-ATAC, constituye un criterio superior a las métricas de clasificación tradicionales para evaluar la relevancia biológica de genes seleccionados en análisis transcriptómicos. Demostramos que una alta precisión predictiva ($>95\%$) no

garantiza que los genes identificados representen cambios biológicos genuinos: en la primera etapa, RF y LASSO alcanzaron performance clasificatoria equivalente a DEG, pero en el análisis posterior sus genes mostraron menor concordancia entre cambios de expresión y accesibilidad de cromatina ($r=0,47$ y $r=0,44$ respectivamente vs $r=0,73$ en DEG). Más aún, esta diferencia en validación molecular se tradujo directamente a utilidad clínica diferencial: solo DEG produjo firmas con valor pronóstico significativo en humanos. La lógica subyacente es que genes cuyos cambios transcripcionales están respaldados por remodelación epigenética coordinada (apertura/cierre de cromatina) reflejan reprogramación celular robusta, mientras que genes seleccionados únicamente por poder discriminativo pueden capturar variabilidad técnica, efectos de lote, o interacciones espurias que no se conservan entre sistemas biológicos. El enfoque propuesto es generalizable: cualquier análisis de expresión diferencial puede beneficiarse de validación ortogonal con datos de accesibilidad, metilación, o modificaciones de histonas, reduciendo la probabilidad de obtener falsos positivos. Esto es particularmente relevante en medicina de precisión, donde los biomarcadores derivados de datos ómicos deben demostrar un fundamento biológico profundo para justificar la evaluación y desarrollo de tests diagnósticos clínicos.

5.2 Valor traslacional en humanos

La firma de 5 genes identificada (SORBS2, CDH1, TRPS1, EPCAM, PTPRK) representa un perfil transcripcional de progresión tumoral conservado entre ratón y humano, con relevancia pronóstica demostrada en 1079 pacientes de cáncer de mama. Estos genes participan en procesos clave de agresividad tumoral: reorganización del citoesqueleto y motilidad celular (SORBS2), pérdida de adhesión intercelular asociada a transición epitelial-mesenquimal (CDH1/E-cadherin), regulación transcripcional pro-metastásica (TRPS1), mantenimiento de identidad epitelial tumoral (EPCAM), y señalización oncogénica (PTPRK). Su expresión coordinadamente elevada en tumores avanzados murinos, respaldada por apertura de cromatina en regiones promotoras, sugiere una activación epigenética concertada durante la progresión. La traslación exitosa a valor pronóstico humano ($HR=1.18$, $p=0.034$) confirma que estos genes no son meros marcadores de estadio tumoral, sino también posibles predictores de resultado clínico: pacientes con alta expresión de la firma tienen 18% mayor riesgo de muerte por desviación estándar de aumento, independientemente de características clínico-patológicas tradicionales. Comparado con firmas comerciales establecidas como *Oncotype DX* (21 genes) o *MammaPrint* (70 genes), nuestra firma de 5 genes ofrece ventajas de costo-efectividad para implementación en contextos de recursos limitados, manteniendo un respaldo metodológico triple (transcriptómico-epigenético-clínico) ausente en firmas derivadas exclusivamente de datos *bulk* sin validación multi-ómica. La conservación funcional entre especies valida además el uso de modelos murinos para descubrimiento de biomarcadores, siempre que se implementen estrategias rigurosas de validación.

5.3 Direcciones futuras y extensiones de la propuesta

Varias líneas de investigación emergen como extensiones naturales de este trabajo. Primero, la estratificación de la firma de 5 genes por subtipos moleculares de cáncer de mama (Luminal A/B, Basal, HER2-enriquecido) permitiría determinar si el valor pronóstico es uniforme o específico de contexto, potencialmente identificando

subgrupos de pacientes donde la firma podría tener máxima utilidad clínica. Segundo, la integración de una tercera capa ómica (proteómica por ejemplo) validaría si los cambios transcripcionales y epigenéticos se traducen efectivamente a niveles de proteína funcional, completando la cascada regulatoria. Tercero, el análisis de genes diferencialmente expresados entre estadios temprano y avanzado (no solo control vs avanzado) podría identificar marcadores de transición temprana con potencial preventivo, en lugar de pronóstico post-diagnóstico. Cuarto, la aplicación de este enfoque a otros tipos tumorales con datos multi-oma pareados disponibles (pulmón, colon, próstata) evaluaría la generalización del principio de validación RNA-ATAC más allá de cáncer de mama.

REFERENCIAS

- Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W. y Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), e8124.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
- Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. y Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12), 1213-1218.
- Cardoso, F., Kyriakides, S., Ohno, S., Penault-Llorca, F., Poortmans, P., Rubio, I. T., Zackrisson, S. y Senkus, E. (2019). Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 30(8), 1194-1220.
- Curtis, C., Shah, S. P., Chin, S. F., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403), 346-352.
- Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C. y Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237), 910-914.
- Klemm, S. L., Shipony, Z. y Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4), 207-220.
- Love, M. I., Huber, W. y Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.
- Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., Hiller, W., Fisher, E. R., Wickerham, D. L., Bryant, J. y Wolmark, N. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England Journal of Medicine*, 351(27), 2817-2826.
- Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P. y Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, 177(7), 1888-1902.
- Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A. y Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence

and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3), 209-249.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K. y Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5), 377-382.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. y Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871), 530-536.

Wang, Z., Gerstein, M. y Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 57-63.

Durante la preparación de este trabajo, los autores utilizaron GEMINI con el fin de mejorar el lenguaje, la calidad de escritura y la legibilidad del manuscrito. Tras utilizar esta herramienta, los autores revisaron minuciosamente cada parte, editaron todo el texto, y asumen plena responsabilidad por el contenido de la publicación.