Contenido

Guía del Pipeline de survGSEA: Del Modelo de Cox a las Vías Biológicas

Proyecto: mirna_glioma **Fecha**: 2026-02-15 **Propósito**: Explicación paso a paso del análisis de enriquecimiento funcional basado en supervivencia (survival-ranked GSEA) aplicado a miRNAs en glioma.

Visión General

Este pipeline responde a una pregunta central:

¿Qué procesos biológicos y vías de señalización están regulados por los miRNAs cuya expresión se asocia con la supervivencia de los pacientes con glioma?

Para responderla, seguimos cinco pasos:

1. **Modelo de Cox** → mide la asociación de cada miRNA con supervivencia
2. **Ranking por z-score** → ordena todos los miRNAs de "peor pronóstico" a "mejor pronóstico"
3. **GSEA con miEAA** → identifica vías biológicas enriquecidas en los extremos del ranking
4. **Reducción de redundancia** → elimina términos duplicados/solapados
5. **Visualización e interpretación** → bubble plots del top 12 por base de datos

---

Paso 1: Modelo de Cox — ¿Qué miRNAs importan para la supervivencia?

¿Qué es el modelo de Cox?

El modelo de riesgos proporcionales de Cox es un modelo de regresión que evalúa la relación entre la expresión de un gen (o miRNA) y el tiempo de supervivencia. Para cada miRNA individual, ajustamos un modelo **univariado**:

$h(t) = h_0(t) \times \exp(\beta \times \text{expresión\_miRNA})$

Donde: - **h(t)**: riesgo instantáneo de muerte en el tiempo t - **$h_0(t)$**: riesgo basal (no necesitamos estimarlo — ventaja del modelo de Cox) - **β (coeficiente)**: cuánto cambia el riesgo por cada unidad de aumento en la expresión del miRNA - **exp(β) = HR (Hazard Ratio)**: la medida de efecto

¿Cómo se interpreta el coeficiente β?

β	HR = exp(β)	Interpretación
β > 0	HR > 1	Mayor expresión → **mayor riesgo** de muerte → **peor** pronóstico
β = 0	HR = 1	La expresión no tiene efecto sobre la supervivencia
β < 0	HR < 1	Mayor expresión → **menor riesgo** de muerte → **mejor** pronóstico

¿Qué es el z-score de Wald?

El z-score es simplemente el coeficiente β dividido por su error estándar:

$$z = \beta / SE(\beta)$$

Este z-score tiene dos propiedades clave: - El **signo** indica la dirección del efecto (positivo = peor pronóstico, negativo = mejor pronóstico) - La **magnitud** refleja tanto el tamaño del efecto como la precisión estadística

Es el análogo exacto del t-estadístico que se usa en expresión diferencial, pero adaptado a supervivencia.

En nuestro análisis

- Se ajustó un modelo de Cox univariado para **cada uno de los miRNAs** detectados en las muestras (n = 29 pacientes)
- La expresión fue normalizada por TMM (edgeR) y transformada a logCPM
- Se estandarizó la expresión (z-scaling) antes del modelo, de modo que el HR se interpreta como cambio por 1 desviación estándar de expresión
- Se usó el método de Efron para manejar tiempos de evento empatados
- **Resultado**: una tabla con el z-score, p-valor, HR e intervalo de confianza para cada miRNA

Paso 2: Ranking por z-score — La lista ordenada de miRNAs

Construcción del ranking

Los miRNAs se ordenan de mayor a menor z-score:

```
                ← Peor pronóstico (z positivo, alto)
  miR-XXX   z = +4.2   (HR = 3.1, expresión alta → muerte más rápida)
  miR-YYY   z = +3.8
  miR-ZZZ   z = +2.5
  ...
  miR-AAA   z = +0.1
  ─── zona neutral (z ≈ 0) ───
  miR-BBB   z = -0.3
  ...
  miR-CCC   z = -2.9
  miR-DDD   z = -3.5
miR-EEE  z = -4.7  (HR = 0.2, expresión alta → supervivencia más larga)
             ← Mejor pronóstico (z negativo, alto en valor absoluto)
```

¿Por qué un ranking continuo y no un corte binario?

El GSEA clásico (Subramanian et al., 2005) trabaja con **toda la lista ordenada**, no con un subconjunto de genes “significativos”. Esto tiene ventajas:

1. **No se pierde información**: incluso miRNAs con efectos modestos contribuyen al enriquecimiento si pertenecen a la misma vía

2. **Detecta efectos coordinados**: una vía puede ser significativa aunque ningún miRNA individual lo sea, si muchos miRNAs de esa vía tienen z-scores moderados en la misma dirección
3. **No depende de un umbral arbitrario** de p-valor

Paso 3: GSEA con miEAA — ¿Qué vías regulan estos miRNAs?

¿Qué es miEAA?

miEAA (miRNA Enrichment Analysis and Annotation) es una herramienta web y API que realiza análisis de enriquecimiento específico para miRNAs. A diferencia de GSEA convencional (que usa conjuntos de genes), miEAA usa conjuntos de **miRNAs** agrupados por las vías que regulan.

¿Cómo funciona el GSEA en este contexto?

1. Se envía la **lista ordenada** de miRNAs (por z-score, de mayor a menor) a miEAA
2. Para cada vía biológica (por ejemplo, "Signaling by VEGF"), miEAA tiene una lista de miRNAs conocidos que regulan genes de esa vía (según miRPathDB)
3. El algoritmo GSEA recorre la lista ordenada de arriba a abajo y calcula un **Enrichment Score (ES)**:
   - Si los miRNAs de una vía están concentrados **arriba** del ranking (z positivos) → la vía está **enriched** (enriquecida hacia peor pronóstico)
   - Si están concentrados **abajo** (z negativos) → la vía está **depleted** (enriquecida hacia mejor pronóstico)
   - Si están distribuidos uniformemente → no hay enriquecimiento

Bases de datos consultadas

Usamos las anotaciones curadas de miRPathDB a través de miEAA en tres bases de datos:

Base de datos	Tipo de anotación	Cobertura
GO Biological Process (GO BP)	Procesos biológicos amplios (apoptosis, ciclo celular, metabolismo…)	1,724 términos
KEGG	Vías de señalización y metabólicas curadas manualmente	129 términos
Reactome	Vías moleculares con alto detalle mecanístico	647 términos

¿Qué significan "enriched" y "depleted"?

Estos términos pueden confundirse. En nuestro contexto:

Dirección	Significado biológico	Implicación clínica
Enriched	Los miRNAs que regulan esta vía tienen z-scores positivos (están arriba del ranking)	La regulación por miRNAs de esta vía se asocia con **peor** supervivencia
Depleted	Los miRNAs que regulan esta vía tienen z-scores negativos (están abajo del ranking)	La regulación por miRNAs de esta vía se asocia con **mejor** supervivencia (o su pérdida con peor pronóstico)

Ejemplo concreto: Si la vía "Signaling by VEGF" está **depleted**, significa que los miRNAs que regulan VEGF tienden a tener expresión asociada con mejor supervivencia. Esto sugiere que cuando estos miRNAs están activos (expresados), suprimen la señalización de VEGF y el paciente sobrevive más. Cuando se pierden, VEGF se desregula y el pronóstico empeora.

Resultados principales

El análisis reveló una **fuerte asimetría**:

Dirección	Total de vías (3 bases de datos)
Enriched	302 (14.2%)
Depleted	**1,830 (85.8%)**

Esta asimetría indica que la mayoría de los miRNAs asociados a supervivencia en glioma están vinculados a la **supresión de vías oncogénicas**. Cuando estos miRNAs se pierden, múltiples programas pro-tumorales se desregulan simultáneamente.

Vías depleted más significativas

Las vías con mayor significancia estadística (P-ajustado más bajo) incluyen:

KEGG (6 vías con FDR < 0.05): - *MicroRNAs in cancer* ($P_{adj} = 2.39\times10^{-7}$) — 205 miRNAs - *Adherens junction* ($P_{adj} = 2.25\times10^{-5}$) — adhesión célula-célula - *Hippo signaling pathway* ($P_{adj} = 4.16\times10^{-4}$) — control de proliferación y tamaño celular - *TGF-β signaling pathway* ($P_{adj} = 0.011$) — transición epitelio-mesenquimal (EMT)

Reactome (top 15 con $P_{adj} < 0.002$): - *Generic Transcription Pathway* ($P_{adj} = 1.06\times10^{-7}$) — regulación transcripcional global - *Signaling by VEGF* ($P_{adj} = 9.49\times10^{-6}$) — angiogénesis tumoral - *Cellular Senescence* ($P_{adj} = 4.33\times10^{-5}$) — barrera anti-tumoral - *Transcriptional regulation by RUNX3* ($P_{adj} = 8.84\times10^{-5}$) — supresor tumoral en glioma - *PIP3 activates AKT signaling* ($P_{adj} = 2.52\times10^{-4}$) — eje PI3K/AKT oncogénico - *Loss of Function of SMAD4 in Cancer* ($P_{adj} = 2.86\times10^{-4}$) - *Interleukin-4 and Interleukin-13 signaling* ($P_{adj} = 3.24\times10^{-4}$) — modulación inmune - *Oncogenic MAPK signaling* (P_{adj}

= 4.91×10^{-4}) - *GRB2 events in EGFR signaling* ($P_{adj} = 8.75\times10^{-4}$) - *Oncogene-Induced Senescence* ($P_{adj} = 8.93\times10^{-4}$)

GO Biological Process ($P_{adj} \sim 10^{-8}$ a 10^{-6}): - *Cellular macromolecule biosynthetic process* ($P_{adj} = 7.67\times10^{-8}$) - *Macromolecule metabolic process* ($P_{adj} = 2.39\times10^{-7}$) - *Positive regulation of gene expression* ($P_{adj} = 1.89\times10^{-6}$)

Estos resultados apuntan a una **depleción convergente** de la regulación por miRNAs en los principales ejes oncogénicos: PI3K/AKT, MAPK/ERK, VEGF (angiogénesis), TGF-β/SMAD, y Wnt/β-catenina.

Vías enriched más significativas

Son muchas menos y con menor significancia: - *Circadian entrainment* (KEGG; $P_{adj} = 0.033$) — única vía enriched con FDR < 0.05 - Varios términos de apoptosis mediada por p53 (TP53 Regulates Cell Cycle Genes, FOXO-mediated transcription, PUMA activation) - Regulación del ciclo celular (cell cycle phase transition)

Esto sugiere que un subconjunto de miRNAs asociados a mal pronóstico podría actuar reforzando mecanismos apoptóticos y de arresto del ciclo celular, posiblemente como respuesta compensatoria.

Paso 4: Reducción de redundancia — Eliminar ruido sin perder señal

¿Por qué es necesaria?

Las bases de datos de vías tienen mucha redundancia. Por ejemplo, en KEGG: - “Pathways in cancer” comparte muchos genes con “MicroRNAs in cancer” - “Non-small cell lung cancer” comparte genes con “Pancreatic cancer”, “Bladder cancer”, etc.

Si 50 vías de cáncer salen significativas pero comparten el 80% de sus miRNAs, eso no son 50 hallazgos independientes — es básicamente el mismo hallazgo repetido 50 veces.

Método 1: Similitud de Jaccard + clustering jerárquico (KEGG y Reactome)

Este método compara directamente los conjuntos de miRNAs entre vías:

```
                    miRNAs de la vía A: {miR-17, miR-20a, miR-93, miR-
106a, miR-21}
                    miRNAs de la vía B: {miR-17, miR-20a, miR-93, miR-
155, miR-34a}

    Intersección (A ∩ B) = {miR-17, miR-20a, miR-93} = 3 miRNAs
     Unión (A ∪ B) = {miR-17, miR-20a, miR-93, miR-106a, miR-21, miR-
155, miR-34a} = 7 miRNAs

    Jaccard = |A ∩ B| / |A ∪ B| = 3/7 = 0.43
```

- **J = 0**: las vías no comparten ningún miRNA (completamente diferentes)
- **J = 1**: comparten exactamente los mismos miRNAs (totalmente redundantes)

Procedimiento: 1. Se calcula la similitud de Jaccard entre todos los pares de vías 2. Se convierte a distancia: D = 1 - J 3. Se aplica clustering jerárquico (UPGMA/enlace promedio) 4. Se corta el dendrograma a un umbral definido 5. Dentro de cada cluster, se selecciona como **representante** la vía con menor P-ajustado (la más significativa)

Umbrales utilizados:

Base de datos	Umbral Jaccard	Justificación
KEGG	0.25	Vías más pequeñas con menos solapamiento, umbral más permisivo
Reactome	0.50	Vías más grandes y jerárquicas con alto solapamiento, umbral más estricto

Ejemplo real: En KEGG depleted, el cluster más grande agrupó 51 vías bajo el representante *"MicroRNAs in cancer"*, absorbiendo vías como *Pathways in cancer*, *Proteoglycans in cancer*, *PI3K-Akt signaling*, *MAPK signaling*, *p53 signaling*, y múltiples cánceres específicos (glioma, melanoma, páncreas, etc.), todas con >25% de solapamiento en miRNAs.

Nota sobre los singletons: Las vías que no se agruparon con ninguna otra (similitud Jaccard por debajo del umbral con todas las demás) se mantienen en la tabla final como términos **únicos y no redundantes**. Esto es importante porque un término significativo que no se parece a ningún otro no debe eliminarse — no es redundante, es único.

Método 2: Similitud semántica con rrvgo (GO BP)

Para GO Biological Process, la similitud de Jaccard no es óptima porque los términos GO tienen una estructura jerárquica (ontología). Por ejemplo: - "Regulation of apoptotic process" es padre de "Positive regulation of apoptotic process" - Estos términos son semánticamente redundantes aunque no compartan exactamente los mismos miRNAs

El paquete **rrvgo** (reduce + visualize Gene Ontology) calcula la **similitud semántica** entre términos GO usando el método Rel (Relevance), que combina: - La posición en la jerarquía GO (terms más específicos son más informativos) - El contenido de información (IC) basado en la frecuencia de anotación

Umbral utilizado: 0.90 (muy estricto, solo agrupa términos altamente similares semánticamente)

Resultado de la reducción

Base de datos	Dirección	Antes	Después	Reducción
KEGG	enriched	24	14	41.7%
KEGG	depleted	105	35	65.7%
Reactome	enriched	129	69	46.5%
Reactome	depleted	518	290	44.0%
GO BP	enriched	219	219	~0% (singletons preservados)
GO BP	depleted	1,505	1,505	~0% (singletons preservados)

Total: 2,719 → 2,132 términos (587 redundantes eliminados, 21.6% de reducción)

La reducción es más agresiva para KEGG (vías con alto solapamiento entre sí) que para GO BP (donde la mayoría de los términos son semánticamente distintos a umbral 0.90).

Referencia metodológica

Esta aproximación de reducción está basada en el **Enrichment Map** de Merico et al. (2010), PLoS ONE 5(11):e13984, que propuso usar el coeficiente de Jaccard para construir redes de similitud entre gene sets y agrupar los redundantes.

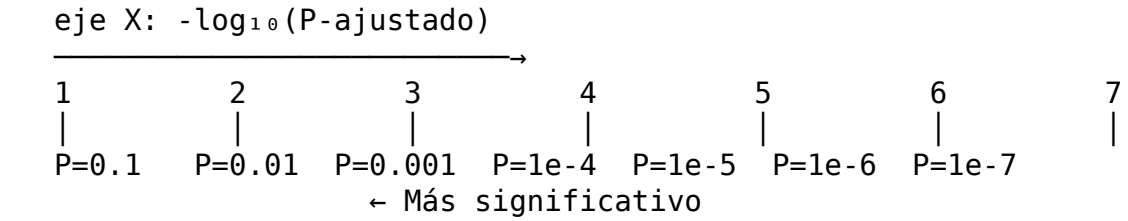
Paso 5: Interpretación de los Bubble Plots (Top 12)

¿Qué muestran las gráficas?

Los bubble plots muestran las **12 vías más significativas** (menor P-ajustado) de cada base de datos, después de la reducción de redundancia.

Se generan dos tipos: 1. **Multipanel** (3 facetas): GO BP | KEGG | Reactome en una sola figura, para comparación directa 2. **Por base de datos**: Una gráfica individual por cada combinación base de datos × dirección

¿Cómo leer cada gráfica?



- **Eje X**: -log₁₀(P-ajustado). Valores más altos = más significativo estadísticamente
- **Eje Y**: Nombres de las vías (ordenadas por significancia, la más significativa arriba)
- **Tamaño de la burbuja**: Número de miRNAs observados en esa vía (conjuntos más grandes = burbujas más grandes)
- **Color**: Verde (#009E73) para enriched, naranja (#D55E00) para depleted (paleta colorblind-safe)

Ejemplo de lectura

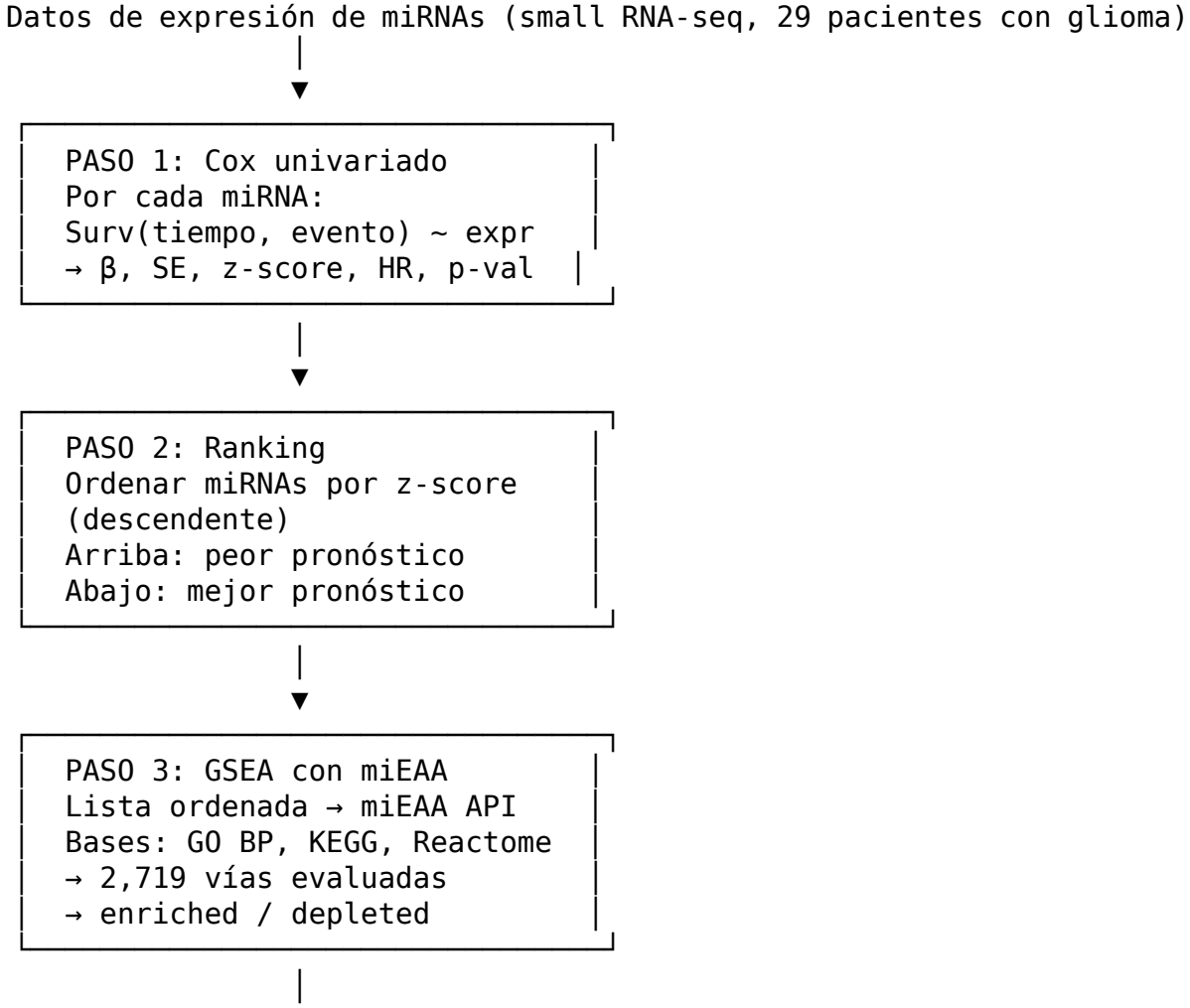
Si en un bubble plot de **Reactome depleted** vemos:

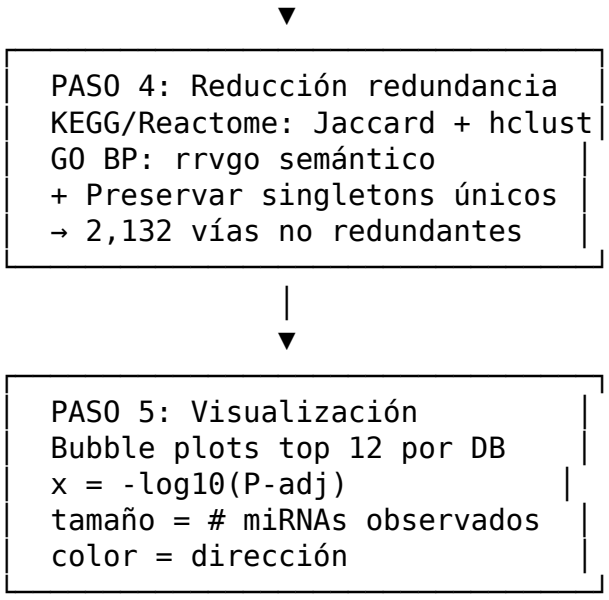
```
Signaling by VEGF          ●●●          (burbuja mediana, x ≈ 5)
Cellular Senescence         ●●●●●●       (burbuja grande, x ≈ 4.4)
PIP3 activates AKT         ●●●●●●●      (burbuja grande, x ≈ 3.6)
```

Esto nos dice: - **Signaling by VEGF** es la vía más significativa ($P_{adj} \approx 10^{-5}$) con ~39 miRNAs - **Cellular Senescence** involucra más miRNAs (~126) aunque es ligeramente menos significativa - **PIP3 activates AKT** tiene muchos miRNAs (~108) apuntando al eje PI3K/AKT

Las tres son **depleted**: los miRNAs que regulan estas vías tienden a tener z-scores negativos (su expresión alta se asocia con mejor supervivencia). Esto implica que la **pérdida** de estos miRNAs reguladores podría estar liberando estas vías oncogénicas, contribuyendo a peor pronóstico.

Resumen del Pipeline Completo





Referencias

1. **Cox PH model**: Cox DR (1972) Regression models and life-tables. *J R Stat Soc B* 34(2):187-220
2. **GSEA original**: Subramanian A et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* 102(43):15545-50
3. **Survival GSEA validation**: Lee S, Kim J, Lee S (2011) A comparative study on gene-set analysis methods for assessing differential expression associated with the survival phenotype. *BMC Bioinformatics* 12:377
4. **SGSEA package**: Deng X, Thompson JA (2023) An R package for Survival-based Gene Set Enrichment Analysis. *Research Square* (preprint)
5. **miEAA**: Kern F et al. (2020) miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic Acids Res* 48(W1):W521-W528
6. **miRPathDB**: Backes C et al. (2017) miRPathDB: a new dictionary on microRNAs and target pathways. *Nucleic Acids Res* 45(D1):D90-D96
7. **Enrichment Map (Jaccard)**: Merico D et al. (2010) Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* 5(11):e13984
8. **rrvgo**: Sayols S (2023) rrvgo: a Bioconductor package for interpreting lists of Gene Ontology terms. *MicroPublication Biology*

Scripts del Pipeline

Paso	Script	Descripción
1-3	`scripts/13_survGSEA_miEAA.R`	Cox univariado + ranking + GSEA con miEAA
4-5	`scripts/18_reduce_redundancy_jaccard.R`	Reducción Jaccard/rrvgo + bubble plots
Auxiliar	`scripts/query_pathway_clusters.R`	Consulta de clusters (qué vías se agruparon)