Nueva herramienta automatizada para el análisis transcriptómico integrado orientada a la interpretación biológica.

Alejandro González Sánchez[a], Cristina Martínez Toledo, Alfonso Esteban Lasso[b]

Inari Biotech SL, Dpto. Biomedicina y biotecnología UAH.

a. alejandro.go.san11@gmail.com **b.** inaribiotech@gmail.com

XI Congreso de Señalización Celular, SECUAH 2026.
XX Simposio de Dianas Terapéuticas.
16 a 18 de marzo, 2026. Universidad de Alcalá. Alcalá de Henares, Madrid. España.

Palabras clave: RNA-seq; análisis transcriptómico; pipeline automatizado; análisis integrado; biomarcadores; expresión génica diferencial

Resumen

El análisis transcriptómico por RNA-seq es una herramienta clave para caracterizar los programas de expresión génica que sustentan los mecanismos moleculares implicados en enfermedad, así como para identificar biomarcadores con potencial diagnóstico, pronóstico o terapéutico. Aunque, existen diversos métodos y herramientas para el análisis del RNA-seq, estos métodos están dispersos, requieren experiencia técnica, generan salidas heterogéneas, y rara vez producen directamente resultados integrados, orientados a interpretación biológica. En este trabajo se presenta un procedimiento automatizado de análisis transcriptómico diseñado para transformar datos crudos RNA-seq humano en resultados directamente interpretables desde el punto de vista biológico. La aproximación integra el control de calidad de las lecturas, su alineamiento al genoma de referencia, el ensamblado, la cuantificación de la expresión génica, la identificación de genes diferencialmente expresados entre condiciones y varios niveles complementarios de interpretación funcional, incluyendo enriquecimiento en términos GO y rutas KEGG, análisis de enriquecimiento de conjuntos génicos y redes de interacción proteína-proteína. El procedimiento se evaluó sobre datos de RNA-seq de 10 líneas celulares de cáncer humano, lo que permitió identificar miles de genes con cambios significativos de expresión y revelar alteraciones coordinadas en procesos relacionados con desarrollo, organización tisular, matriz extracelular, regulación de la adhesión celular, así como en múltiples procesos y vías funcionales implicados en fenotipos tumorales. La integración de análisis basados en genes individuales y en conjuntos génicos facilita detectar tanto cambios puntuales como reprogramaciones funcionales globales. A diferencia de estudios centrados en análisis aislados, esta estrategia proporciona una solución integrada, reproducible y orientada a la interpretación mecanística, que acelera la transición desde datos transcriptómicos hasta hipótesis sobre mecanismos moleculares y candidatos a biomarcador. El procedimiento constituye así una plataforma versátil para estudios de señalización celular y para el descubrimiento sistemático de firmas transcriptómicas asociadas a estados biológicos y patológicos.

Cita: González Sánchez, Alejandro; Martínez Toledo, Cristina; Esteban Lasso, Alfonso (2026) Nueva herramienta automatizada para el análisis transcriptómico integrado orientada a la interpretación biológica. Actas del XI Congreso de Señalización Celular, SECUAH 2026. XX Simposio de Dianas Terapéuticas. 16 a 18 de marzo, 2026. Universidad de Alcalá. Alcalá de Henares, Madrid. España. dianas 15 (1): e202603fp001. ISSN 1886-8746 (electronic) journal.dianas.e202603fp001 https://dianas.web.uah.es/journal/e202603fp001. URI http://hdl.handle.net/10017/15181

Introducción

La secuenciación de RNA (RNA-seq) se ha consolidado como la tecnología de referencia para el estudio de la expresión génica en sistemas biológicos complejos, permitiendo caracterizar estados celulares, procesos de diferenciación y mecanismos moleculares asociados a enfermedad. Su aplicación en oncología ha resultado especialmente relevante para la identificación de biomarcadores, la estratificación tumoral y la comprensión de alteraciones en rutas de señalización [1].

Sin embargo, la creciente complejidad de los análisis transcriptómicos no reside únicamente en las etapas iniciales de procesamiento de datos, sino en la interpretación biológica integrada de los resultados. Aunque existen numerosas herramientas para cuantificación, expresión diferencial y análisis funcional, estas suelen generar salidas fragmentadas y difícilmente integrables, lo que dificulta la transición desde matrices de expresión hasta interpretaciones mecanísticas verificables [2].

En este contexto, el presente trabajo introduce una herramienta automatizada para el análisis transcriptómico integrado, diseñada específicamente para transformar datos crudos RNA-seq humano en resultados directamente interpretables desde el punto de vista biológico. A diferencia de pipelines centrados en aspectos computacionales, este enfoque pone el acento en la coherencia biológica, la integración de

múltiples niveles de análisis funcional y la generación de tablas finales que facilitan la exploración mecanística de los datos.

Materiales y métodos

Conjunto de datos

La herramienta desarrollada se evaluó utilizando datos públicos de RNA-seq correspondientes a líneas celulares humanas, procedentes de la base de datos NCBI Sequence Read Archive (SRA). El conjunto de datos analizado comprende 14 muestras de RNA-seq en modo paired-end, derivadas de 10 líneas celulares distintas. En concreto, se incluyeron muestras de las siguientes líneas: PC3 (adenocarcinoma de próstata), MCF-7 (adenocarcinoma de mama), HepG2 (hepatocarcinoma), HEK293 (riñón embrionario humano), CACO-2 (adenocarcinoma colorrectal), A-549 (adenocarcinoma de pulmón), U2OS (osteosarcoma), U-251 MG (glioblastoma), A431 (carcinoma epidermoide) y SH-SY5Y (neuroblastoma). (National Center for Biotechnology Information, Sequence Read Archive.)

Estas líneas celulares abarcan orígenes tisulares diversos —epitelial, mesenquimal y neural— y presentan programas transcriptómicos claramente diferenciados, lo que las convierte en un conjunto de datos adecuado para evaluar herramientas orientadas a la interpretación funcional integrada del transcriptoma, más allá de comparaciones específicas entre dos condiciones experimentales.

Control de calidad

El control de calidad preliminar se realizó con fastp [3], aplicando filtros multiparamétricos sobre lecturas crudas para eliminar secuencias de baja calidad y adaptadores. Se generaron informes con estadísticas detalladas de calidad por muestra (puntuación Phred, contenido GC y longitud de lectura), para verificar la idoneidad de los datos antes del alineamiento.

Alineamiento

El alineamiento a GRCh38 se ejecutó con HISAT2 [4], empleando parámetros optimizados para datos paired-end. Se calcularon métricas de calidad de alineamiento y se generaron archivos BAM ordenados, listos para la cuantificación.

Ensamblaje transcriptómico

La reconstrucción de transcritos se realizó con StringTie [5] mediante un ensamblaje transcriptómico guiado por referencia, integrando anotaciones GTF existentes como guía. Se generaron tablas consolidadas que reunían los resultados de forma ordenada junto a las anotaciones génicas, facilitando el análisis downstream.

Cuantificación de la expresión génica

La cuantificación se realizó a nivel génico a partir de archivos BAM alineados al genoma humano de referencia (GRCh38). Los conteos de lecturas por gen se obtuvieron con featureCounts [6], empleando anotaciones génicas curadas y estables. Los identificadores génicos se armonizaron y se mapearon a identificadores Ensembl, descartando aquellos genes con asignación ambigua o sin correspondencia inequívoca. Adicionalmente, se calcularon métricas normalizadas (FPKM y TPM) con fines descriptivos y de visualización, manteniendo los conteos crudos como entrada para los análisis estadísticos posteriores.

Análisis de expresión diferencial

Las comparaciones entre condiciones se realizaron mediante modelos estadísticos adecuados al diseño experimental. Para comparaciones con replicación biológica se aplicaron modelos basados en distribución binomial mediante DESeq2 [7], mientras que las comparaciones exploratorias sin réplicas se abordaron mediante enfoques alternativos con edgeR [8], interpretándose con cautela [7,8].

Los genes diferencialmente expresados (DEGs) se definieron mediante umbrales estrictos de magnitud de cambio y significación estadística, generándose tablas específicas para genes sobreexpresados y subexpresados.

Interpretación funcional integrada

La interpretación biológica constituye el núcleo diferencial de la herramienta desarrollada. Para cada comparación se realizaron análisis complementarios de enriquecimiento funcional (GO y KEGG), análisis de enriquecimiento de conjuntos génicos (GSEA) y redes de interacción proteína-proteína [2,9,10].

- Enriquecimiento funcional en ontologías génicas (GO)
- Análisis de rutas (KEGG)
- Análisis de enriquecimiento de conjuntos génicos (GSEA)
- Redes de interacción proteína-proteína

Los resultados se integraron en estructuras de salida coherentes, organizadas por comparación y condición, permitiendo identificar tanto cambios puntuales como reprogramaciones funcionales globales.

Resultados

Como se resume en la figura 1, el flujo de trabajo integra el control de calidad de las lecturas, su alineamiento al genoma de referencia, el ensamblado, y la cuantificación, así como el análisis de expresión diferencial y la interpretación funcional, con el objetivo de generar resultados directamente interpretables desde el punto de vista biológico.

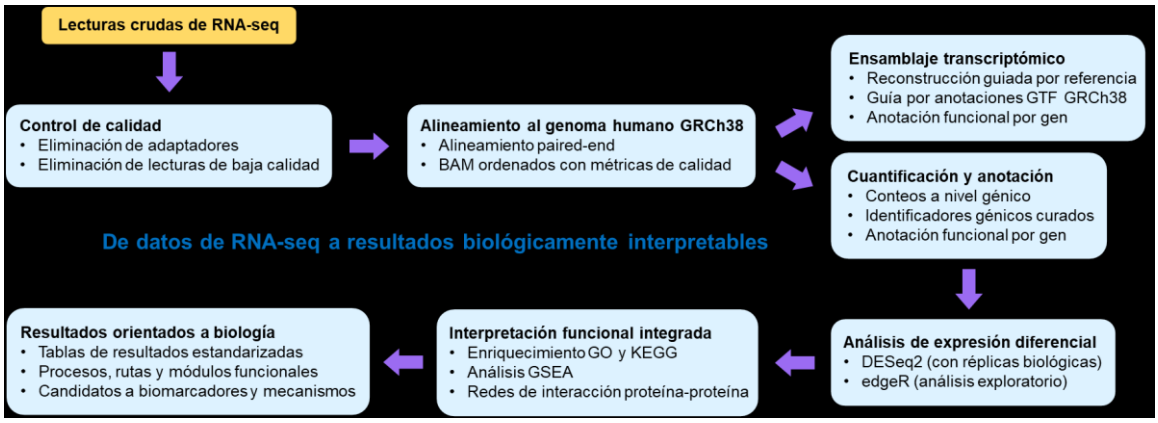

Figura 1. Esquema simplificado del flujo de análisis transcriptómico automatizado desarrollado en este estudio. El procedimiento transforma datos crudos de RNA-seq humano en resultados biológicamente interpretables mediante el control de calidad de las lecturas, el alineamiento al genoma humano de referencia (GRCh38), el ensamblaje transcriptómico guiado por referencia, la cuantificación de la expresión génica, el análisis de expresión diferencial y una interpretación funcional integrada, permitiendo identificar procesos biológicos alterados, rutas de señalización y posibles biomarcadores.

El control de calidad identificó y eliminó adaptadores, lecturas de baja calidad y contaminantes, mejorando significativamente los Phred scores promedio (>Q30) y asegurando homogeneidad en el contenido GC y en la longitud de las lecturas. Se consiguió alinear eficientemente >90% de los fragmentos paired-end al genoma GRCh38, generando archivos BAM con altas tasas de mapeo único, bajos duplicados y cobertura uniforme de regiones génicas transcritas.

En relación con el panorama transcriptómico general, la cuantificación permitió detectar expresión génica robusta en decenas de miles de genes, con patrones claramente diferenciados entre líneas celulares. El filtrado por estabilidad y anotación redujo la complejidad inicial, aumentando la fiabilidad biológica del conjunto final de genes analizados.

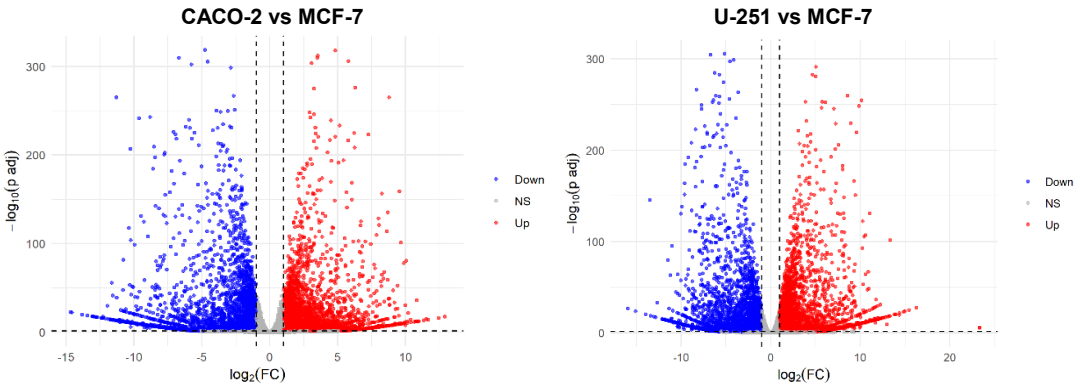

Figura 2. Gráficos de volcán de dos comparaciones representativas con alto número de genes diferencialmente expresados (CACO-2 vs MCF-7 y U-251 MG vs MCF-7). Se muestran en rojo los genes sobreexpresados en CACO-2 y U-251 MG, en azul los genes sobreexpresados en MCF-7 y en gris los genes no significativos (log2FC ≥ 1; p ajustado ≤ 0.05).

En cuanto a los genes diferencialmente expresados, el análisis identificó miles de genes con cambios significativos entre condiciones, incluyendo reguladores de desarrollo, componentes de la matriz extracelular y genes asociados a adhesión celular y migración. Como se muestra en la figura 2, los gráficos de volcán resumen la magnitud (log2FC) y la significación (p ajustado) de los genes diferencialmente expresados en dos comparaciones representativas (CACO-2 vs MCF-7 y U-251 MG vs MCF-7).

Respecto a las alteraciones funcionales coordinadas, los análisis de enriquecimiento funcional revelaron una regulación convergente de procesos biológicos relacionados con organización tisular, señalización celular, remodelado de la matriz extracelular y programas de diferenciación, coherentes con los fenotipos de las líneas analizadas.

Además, el análisis GSEA permitió detectar firmas transcriptómicas consistentes incluso en casos donde los cambios individuales no alcanzaban significación estadística, reforzando la interpretación a nivel de sistemas biológicos [9].

Finalmente, la integración de interacciones proteína-proteína destacó módulos funcionales densamente conectados, sugiriendo la activación conjunta de rutas moleculares relevantes y facilitando la identificación de genes candidatos con potencial interés como biomarcadores [10].

Discusión

Este trabajo presenta una herramienta de análisis transcriptómico que trasciende la ejecución secuencial de métodos estándar, priorizando la interpretación biológica integrada de los resultados. La principal aportación del TFM radica en la estructuración sistemática de salidas orientadas al biólogo experimental, reduciendo la barrera entre análisis bioinformático y generación de hipótesis biológicas.

A diferencia de pipelines centrados en optimización computacional, esta estrategia enfatiza la coherencia funcional, la trazabilidad de los resultados y la posibilidad de explorar los datos desde múltiples niveles de organización biológica. La integración de análisis basados en genes individuales y conjuntos génicos permite capturar tanto eventos discretos como reprogramaciones transcriptómicas globales.

Si bien el procedimiento se evaluó en líneas celulares tumorales, su diseño modular lo hace aplicable a otros contextos biológicos y patológicos. Futuras extensiones podrían incorporar otros tipos de datos ómicos y aumentar la flexibilidad frente a diseños experimentales complejos [11]. La herramienta desarrollada constituye una plataforma reproducible y orientada a la interpretación biológica para el análisis de datos de RNA-seq humano. Su aplicación facilita la transición desde datos transcriptómicos hasta hipótesis mecanísticas, acelerando el descubrimiento de firmas transcriptómicas y biomarcadores.

Agradecimientos

Este trabajo ha sido financiado por una ayuda de colaboración público-privada del Plan Estatal de Investigación Científica, Técnica y de Innovación-Next Generation EU, concedida por el Ministerio de Ciencia e Innovación (CPP2023-010929). Los autores agradecen el apoyo recibido para el desarrollo de esta investigación.

Bibliografía

1. Ozsolak, F., and Milos, P. M. 2011. RNA sequencing: Advances, challenges and opportunities. Nature Reviews Genetics, 12 (2), 87–98.
2. Xu, S., Hu, E., Cai, Y., Xie, Z., Luo, X., Zhan, L., Tang, W., Wang, Q., Liu, B., Wang, R., Xie, W., Wu, T., Xie, L., and Yu, G. 2024. Using clusterProfiler to characterize multiomics data. Nature Protocols, 19 (11), 3292–3320.
3. Chen, S., Zhou, Y., Chen, Y., & Gu, J. 2018. Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560.
4. Kim, D., Langmead, B., & Salzberg, S. L. 2015. HISAT: a fast spliced aligner with low memory requirements. Nature Methods, 12(4), 357–360. https://doi.org/10.1038/nmeth.3317.
5. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology, 33(3), 290–295. https://doi.org/10.1038/nbt.3122
6. Liao, Y., Smyth, G. K., & Shi, W. 2014. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656
7. Love, M. I., Huber, W., & Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550. https://doi.org/10.1186/s13059-014-0550-8

8. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

9. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102 (43), 15545–15550.

10. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., and von Mering, C. 2023. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research, 51 (D1), D638–D646.

11. Yu, K., Chen, B., Aran, D., Charalel, J., Yau, C., Wolf, D. M., van ’t Veer, L. J., Butte, A. J., Goldstein, T., and Sirota, M. 2019. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. Nature Communications, 10 (1), 3574.