WHITE PAPER

CARACTERÍSTICAS DE UN ANÁLISIS SINGLE CELL RNA-SEQ

Enrique de la Rosa
Bioinformatics Scientist

info@dreamgenics.com
985 088 180 I 613 038 948
www.dreamgenics.com

Resumen

La secuenciación masiva ha revolucionado nuestra comprensión de los sistemas biológicos en los últimos años. En particular, la tecnología Single cell RNA-seq se ha posicionado en diversas disciplinas biomédicas como una herramienta esencial que permite analizar el transcriptoma completo de cada célula de manera individual. De esta manera, el análisis Single cell RNA-seq puede revelar procesos biológicos y mecanismos moleculares específicos que podrían no ser detectados mediante técnicas de secuenciación de ARN a nivel poblacional o de tejido.

Introducción

En los últimos años, los avances en el ámbito de la Secuenciación de Próxima Generación (NGS, por sus siglas en inglés), o tecnologías de alto rendimiento, han propiciado un profundo avance en nuestra comprensión de los sistemas biológicos, la diversidad humana y las enfermedades.

Estas tecnologías genómicas, transcriptómicas y otras tecnologías multi-ómicas están siendo cada vez más empleadas en el estudio de células individuales. Este enfoque de secuenciación unicelular posibilita la exploración de poblaciones celulares complejas y desconocidas sin que se pierda la heterogeneidad celular, permitiendo así la revelación de relaciones reguladoras entre genes y la comprensión de diferencias y relaciones evolutivas entre distintas células.

Esta práctica tiene amplias aplicaciones en campos como la oncología, microbiología, neurología, reproducción, inmunología y otras, convirtiéndose así en una herramienta crucial para la investigación biomédica.

Una de las tecnologías más destacadas y relevantes en este ámbito es el RNA-seq de célula única (scRNA-seq), que analiza el transcriptoma de cada célula de manera individual. Esta técnica proporciona una cantidad extraordinaria de datos, lo que representa un gran desafío en su análisis e interpretación.

El objeto central de estudio de esta técnica es la

matriz de conteos, en la que las células se organizan en columnas y los genes en filas.

Posibles ventajas de scRNA-seq frente a otras técnicas de secuenciación incluyen su capacidad para capturar la heterogeneidad celular sin necesidad de agrupación previa, lo que permite un análisis más preciso y detallado de la diversidad celular en un tejido o muestra. Además, al analizar el transcriptoma de células individuales, scRNA-seq puede revelar procesos biológicos y mecanismos moleculares específicos que podrían no ser detectados mediante técnicas de secuenciación de ARN a nivel poblacional o de tejido. Esto hace que scRNA-seq sea una herramienta poderosa para comprender la biología celular en un nivel más profundo y detallado.

Pipeline e interpretación de outputs

Preparación de muestras y secuenciación

De forma previa a la obtención del objeto central de análisis o matriz de conteos se hace necesario secuenciar el transcriptoma de cada célula individual. Como ejemplo, la tecnología 10X Chromium emplea una sofisticada técnica de microfluidos para capturar células individualmente y preparar bibliotecas de cDNA.

Específicamente, esta tecnología utiliza Microflujos de Gel en Emulsión (GEM) que contienen oligonucleótidos con códigos de barras, lo que permite la identificación única de cada célula capturada. Un aspecto notable es que estos oligonucleótidos con códigos de barras incorporan una secuencia de Identificador Molecular Único (UMI), compuesta por varios pares de bases., que facilita la distinción de las moléculas específicas y únicas capturadas en cada célula.

Además de capturar las células, los GEM tienen la capacidad de incorporar los reactivos necesarios para la transcripción inversa, lo que permite la preparación de las bibliotecas de cDNA dentro de ellos de manera simultánea. Posteriormente, las bibliotecas obtenidas son amplificadas y secuenciadas.

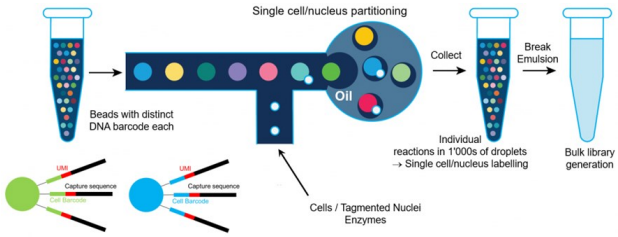

Figura 1. Protocolo de secuenciación single-cell mediante GEMs.

Análisis de la matriz de *counts*

El análisis de la matriz de datos de single-cell requiere de diversos pasos que emplean algoritmos muy sofisticados, entre los que se incluyen algunos de *machine learning* no supervisado, como la reducción de la dimensionalidad o el *clustering.*

Una vez tenemos la matriz de *counts* generada, el primer paso será realizar un **control de calidad**, eliminando los genes menos expresados y las células con valores de expresión considerados como *outliers*. Un punto realmente importante aquí es el filtrado de células que superan el umbral de los recuentos que mapean con el ADN mitocondrial.

La **normalización** será el siguiente paso, haciendo los datos más comparables entre sí. La normalización redistribuye los valores de recuento en un intervalo pequeño sin afectar las diferencias relativas entre ellos.

El siguiente paso será la **reducción de dimensionalidad**. En términos simples, es la transformación de datos desde un espacio de alta dimensionalidad a uno de baja dimensionalidad, de manera que la representación de baja dimensionalidad retenga algunas propiedades significativas de los datos originales. Para esta aproximación se realiza una reducción dimensional lineal llamada Análisis de Componentes Principales (PCA, por sus siglas en inglés).

En este sentido, tratamos de buscar el número de estas componentes principales que cubran la mayor variabilidad de la muestra. Un punto destacado es que, a pesar de que ya están los valores normalizados en la matriz de conteos, es obligatorio realizar **escalado y centrado** de los mismos.

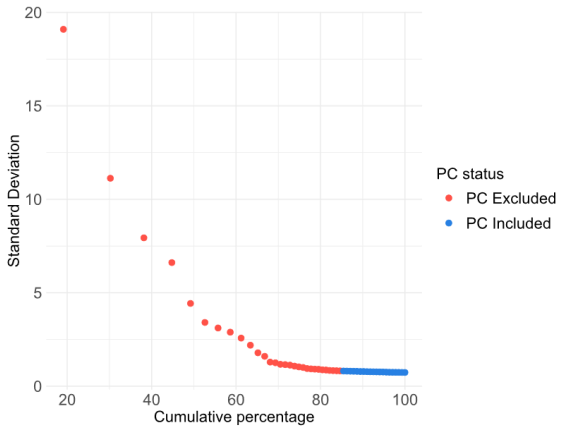

Figura 2. Análisis de Componentes Principales (PCA). Reducción de la dimensionalidad.

En el siguiente paso del análisis empezaremos a encontrar los primeros resultados del proceso analítico de single-cell: el ***clustering***. Se realiza un grafo de k-vecinos más cercanos (KNN, por sus siglas en inglés) con distancias euclídeas entre las células. Luego, este grafo KNN se utiliza para construir el grafo de Vecinos Compartidos más Cercanos (SNN, por sus siglas en inglés). Estos algoritmos clasifican fácilmente cada célula buscando similitud en un número determinado de vecinos celulares (k) y clasificándolos en el grupo que contiene células más relacionadas.

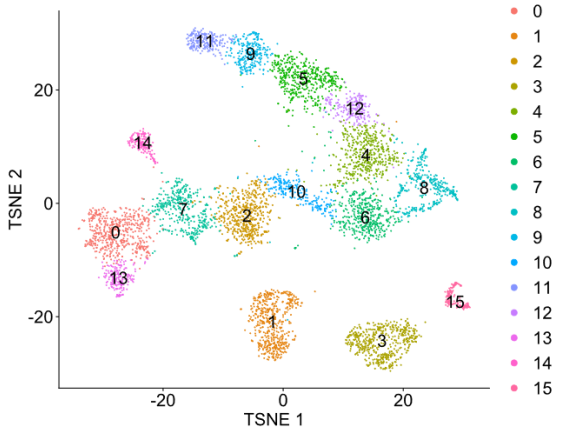

Figura 3. Resultado del clustering con los clusters o poblaciones celulares encontrados por el algoritmo.

Para este resultado de *clustering*, podemos realizar un etiquetado en función al tipo celular dado por citometría de flujo en caso de que los datos de single-cell estén supervisados, o podemos realizar una anotación en función a una base de datos conocida que recoja firmas génicas para los distintos tipos celulares.

Se trata, por tanto, del siguiente paso en el análisis: anotación de *clusters* con el tipo celular correspondiente.

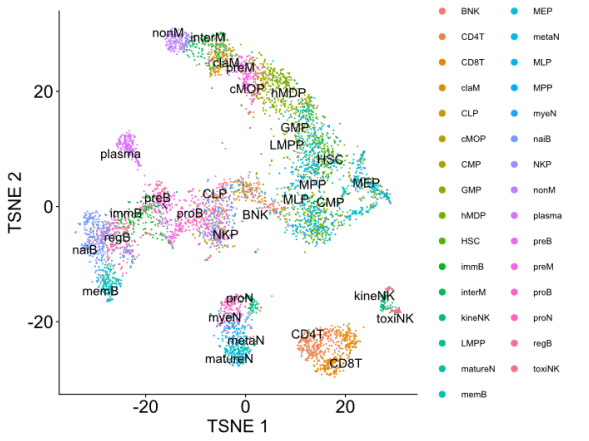

Figura 4. Resultado del *clustering* con los clusters o poblaciones celulares anotados por el tipo celular. En este caso los datos venían supervisados por citometría de flujo.

Finalmente, nos centraremos y pondremos el foco en los genes que nos han permitido distinguir los distintos *clusters* o poblaciones celulares, es decir, buscaremos los **genes diferencialmente expresados** entre ellos.

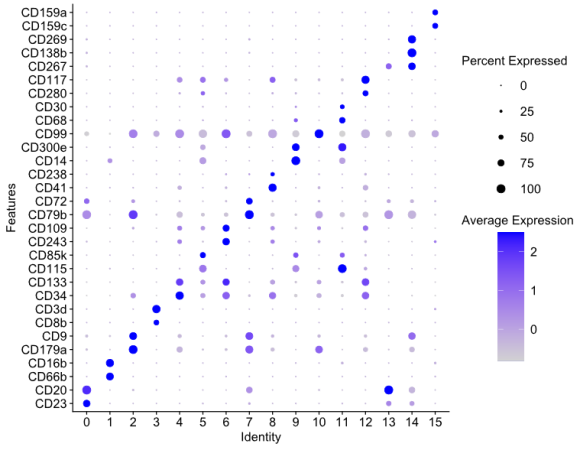

Figura 5. DotPlot con el TOP2 genes diferencialmente expresados por cada cluster.

Sobre estos genes podemos llevar a cabo un análisis de enriquecimiento de términos de ontología de genes (GO, por sus siglas en inglés) o vías metabólicas (WikiPathways) para terminar así de completar toda la información y contextualizar dónde están las principales diferencias entre las distintas poblaciones/subpoblaciones celulares encontradas.

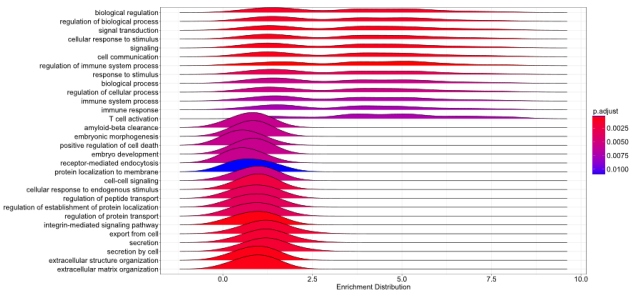

Figura 6. RidgePlot con los términos GO enriquecidos para una firma dada.

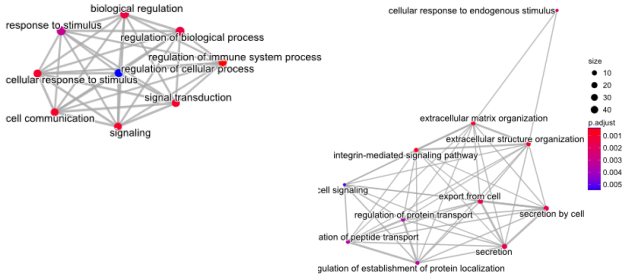

Figura 7. Network Plot con los términos GO enriquecidos para una firma dada.

Llegados a este punto tenemos completado un análisis genérico de scRNA-seq, pero aún quedan herramientas y posibles algoritmos aplicables a este tipo de estudios, como el **análisis de trayectorias, RNA velocity y el análisis de contactos celulares.**

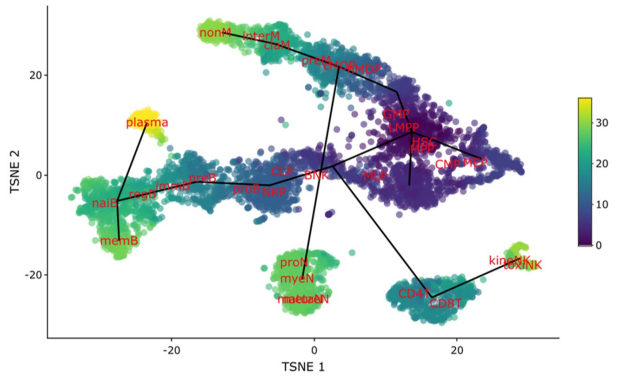

Figura 8. Análisis de trayectorias con las células coloreadas acorde al hiperparámetro pseudotime.

Como se puede observar, es posible aislar una ingente cantidad de datos o información en un análisis scRNA-seq, dado que poseemos información de cada una de las células y con ello el poder de resolución de la técnica es abrumador.

Si sumamos este hecho a su potencial aplicación a todo tipo de estudios biomédicos, podremos entender por qué se trata una de las herramientas más prometedoras que existen en la actualidad y por qué ha despertado el interés de la comunidad científica haciendo que su uso esté cada vez más demandado.

Bibliografía

1. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun. 2020 May 8;11(1):2285.
2. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018 Aug 7;50(8):96.
3. Wu Y, Zhang K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. at Rev Nephrol. 2020 Jul;16 (7):408-421.
4. Tang X, Huang Y, Lei J, Luo H, Zhu X. The single-cell sequencing: new developments and medical applications. Cell Biosci. 2019 Jun 26;9:53.
5. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018 Jun;36(5):411-420.
6. Xie X, Liu M, Zhang Y, et al. Single-cell transcriptomic landscape of human blood cells. Natl Sci Rev. 2020 Aug 24;8 (3):nwaa180.

Ampliar información

Encuentra más información sobre nuestro servicio de análisis Single cell RNA-seq en nuestra página web www.dreamgenics.com

Contacto comercial

Puedes contactar con nuestro equipo comercial a través de las siguientes vías:

info@dreamgenics.com

+34 985 088 180 I 613 038 948