Universidad Nacional Autónoma de México
Facultad de Ciencias

Genómica Computacional
Semestre 2023-2

Metodología en la identificación de subtipos de cáncer de mama

Martínez Piza Erick[1] Ramírez Muñoz Sergio Iván Alejandro[2] & Tiscareño Galvan Daniela[3]
ermarp@ciencias.unam.mx *, serchrmz01@ciencias.unam.mx[2], danielatiscareno@ciencias.unam.mx [3]

Facultad de Ciencias, Universidad Nacional Autónoma de México [1]*, Av Universidad 3000, Cd. Universitaria, Coyoacán, C.P. 04510, Ciudad de México, México.*

Resumen

Es bien sabido que los distintos tipos de cánceres, como lo es el cáncer de mama, son ecosistemas complejos de células tumorales con estructuras moleculares lo suficientemente distintas para abrir todo un área de investigación para cada instancia. A esto se suele llamar “heterogeneidad intratumoral” y es uno de los mayores retos para diagnosticar el cáncer y darle un tratamiento.

Para esto analizamos el procedimiento de una técnica llamada conCluster, la cual, de forma similar a otras técnicas de clasificación sobre scRNA-seq (secuencias de RNA de célula individual), ayudan a dar una mejor introspectiva en la clasificación de subtipos de cáncer y generando, a su vez, una mejoría en la complejidad computacional en un orden de tiempo sobre agrupamientos de datos genéticos que presenten ruido y dimensiones de tamaño elevadas.

Presentaremos nuestra pregunta de investigación y con base en la metodología seleccionada, obtendremos nuestro resultado de clusterización y, posteriormente, la identificación de subtipos, de código nucleico, sobre las células cancerosas, en particular, provenientes del cáncer de mama.

Introducción

Tan solo en 2020, 2.3 millones de mujeres fueron diagnosticadas con cáncer de mama, lo cual es equivalente a una mujer siendo diagnosticada cada 18 segundos (Sylwia Łukasiewicz et al., 2021). La creciente incidencia del cáncer de mama, la exhibe una alza anual del 3.1%, hace que la generación de nuevas estrategias de caracterizar a la enfermedad destaca la necesidad urgente de estrategias de gestión y tratamiento eficaces (Arzanova & Mayrovitz, 2022).

Las decisiones de tratamiento se encuentran fuertemente influenciadas por las características histológicas y moleculares de cada tumor. El cáncer de mama se caracteriza por su heterogeneidad molecular, la cual desempeña un papel crucial en la determinación de los tratamientos a emplear en el curso de la enfermedad. Se han desarrollado diversas clasificaciones para categorizar el cáncer de mama según las alteraciones moleculares. Actualmente, se distinguen cuatro subtipos de cáncer de mama, cuya clasificación está basada en la expresión o no expresión de receptores de estrógeno, progesterona, y HER2, o receptor 2 del factor de crecimiento epidérmico humano. Cada subtipo presenta diferentes tasas de incidencia, crecimiento, y supervivencia (Harbeck et al., 2019).

Los tumores luminal A se caracterizan por la expresión del receptor de estrógeno y/o el receptor de progesterona , así como la ausencia de HER2 y una baja expresión del marcador de proliferación celular Ki-67. A este tipo de tumores se le asocia un pronóstico más favorable en comparación a los otros subtipos de cáncer de mama, pues exhiben un crecimiento lento y bajos índices de recaída. Los tumores luminal B, a diferencia del tipo A, tienen una mayor tasa de crecimiento asociada a la expresión elevada del marcador de proliferación celular Ki-67. Este subtipo expresa el receptor de estrógeno y negativos para el receptor de progesterona. El tercer subtipo descrito, HER2 positivo, carece de receptores de estrógeno y progesterona, con una alta expresión de HER2. Este tipo de cáncer tiene un pronóstico peor en comparación a los tumores luminales, y requiere tratamientos específicos a la proteína HER2/neu. Por último el subtipo triple negativo, TNBC, carece de expresión de ER, PR y HER2, y tiene el peor pronóstico de los cuatro subtipos descritos. El TNBC es agresivo, con recaída temprana y tendencia a ser diagnosticado en etapas avanzadas. Tiene una alta tasa de proliferación, y se asocia a alteraciones en los genes de reparación del ADN e inestabilidad genómica aumentada (Orrantia-Borunda et al., 2022).

El diagnóstico preciso y temprano de los subtipos de cáncer de mama es crucial para implementar estrategias de tratamiento efectivas. La identificación temprana de este subtipo puede tener un impacto significativo en la elección del tratamiento, mejorando los resultados para el paciente. Como consecuencia, la implementación de nuevas tecnologías que hagan el diagnóstico más certero y preciso resulta de gran importancia. En la actualidad, la secuenciación de ARN de una sola célula ha surgido como una poderosa herramienta para caracterizar la

heterogeneidad intratumoral en el cáncer de mama. A diferencia de la secuenciación tradicional de ARN, esta técnica permite cuantificar la expresión génica a nivel de células individuales, proporcionando la capacidad de analizar las diferencias entre diversas poblaciones celulares dentro de un tumor (Gan et al., 2018).

Pregunta de Investigación

¿Se puede crear un programa eficiente en python para poder identificar subtipos de cáncer recibiendo como entrada una secuencia de RNA?

Objetivo

- Identificar subtipos de cáncer de mama a partir de una secuencia de scRNA.
- Identificar algoritmos que clasifiquen secuencias de scRNA sobre células tumorales con complejidad computacional manejable.

Método

Para el desarrollo de los resultados, se usará un método de conClustering en dónde se agrupan múltiples resultados de cluster. Este sistema toma como entrada un matriz E, que está dada por N células(filas) x G genes(columnas), la cuál pasará por 4 pasos para obtener una partición de N en K clusters.

Paso 1: Filtración de los genes

Para poder enfocarnos únicamente en las células importantes del tumor, filtraremos los genes raros y ubicuos, pues usualmente no son útiles en el clustering, e identificaremos los genes más variables del conjunto de datos que tenemos. Los genes raros casi no tienen un porcentaje de aparición con menos del 6%, en cambio los genes ubicuos tienen un porcentaje por encima del 94%, Después de la filtración, identificamos el conjunto de genes que tuvo un porcentaje más variable sobre todos los demás.

Paso 2: Reducimos la dimensión usando t-SNE

Dado que, aún después de la filtración de genes, la dimensión del conjunto puede ser muy grande usaremos una técnica llamada t-SNE para convertirlo en un

subespacio dimensional de menor tamaño. En el t-SNE el parámetro de perplejidad es muy importante para obtener el número de vecinos efectivos. La perplejidad justamente es usada para saber cuantos vecinos cercanos serán usados en el algoritmo, por lo que en esta ocasión, usaremos un número de 30 que suele ser el determinado. Todo esto para reducir el conjunto filtrado obtenido de la secuencia de scRNA en dos dimensiones.

Paso 3: Partición de células

Para esto nos basamos en las matrices obtenidas para hacer el K-means clustering con diferentes parámetros iniciales. Esto lo realizaremos T veces para obtener diferentes particiones para estas celdas individuales. Para cada uno de los resultados individuales de cluster se deriva una matriz binaria B(N x Kt) que se construyó en base a las etiquetas de clúster correspondientes de las células N donde Kt es el número de cluster en la t'ava partición. La matriz por ser binaria, sólo tendrá elementos de 1 o 0.

Paso 4: Consenso de clusters

Después de obtener las T diferentes particiones, vamos a concatenar cada una de esa matrices en una más larga, y a esta, vamos a aplicarle el K-means cluster basado en la matriz binaria recientemente obtenida, Aquí usamos el índice de Calinski-Harabaz para decidir el número de clusters. Y por último, fusionamos los resultados de cada resultado de agrupamiento individual en uno de consenso.

Conjuntos de datos

El conjunto de datos que se utilizó fue recuperado de la página FireBrowse por parte del Board Institute en donde se descargó la información de secuencias de RNA del cáncer de mama. Este repositorio tiene en su conjunto los 1098 casos del TCGA(The Cancer Genome Atlas) con fecha de 2016.

También, para las pruebas sobre con clustering, se utilizaron muestras de secuencias scRNA de células cancerígenas recabadas de la NLM (National Library of Medicine, por sus siglas en inglés). Con códigos (NG_056086, OL451363,

KX095629, NM_001110394, KY771608, KX095591, KX095621, NM_138081, OL451372, NM_001406721).

Resultados

Como resultado, podemos ver que existen muchas bibliotecas que ya implementan los métodos utilizados para la ejecución de métodos como ConClustering, consensus clustering, que nos ayudan a obtener gráficas que nos ayuden a visualizar e identificar subtipos de cáncer de mamá, e incluso otros cánceres, dadas las secuencias de RNA obtenidas por muestras de distintas células que presenten dicha condición.

Hay que considerar que, al utilizar datos en gran cuantía, los métodos computacionales pueden encarar dificultades para procesar los datos. A su vez, los cánceres suelen mostrar una heterogeneidad de tumores considerable en la gran mayoría de características fenotípicas reconocibles, por lo que, en esos casos, es importante acumular o aglomerar correctamente el conjunto de células en distintos subtipos a partir de los datos expresados en células individuales.

Entonces, debido a estas variaciones biológicas, estos conjuntos de datos de secuencias scRNA suelen exhibir ruido y mayor complejidad espacial sobre nuestros métodos de clasificación y al tratar con ellos podemos acabar con clasificaciones con ruido y demoradas.

Para esto existen métodos como ConClustering, t-SNE, K-Means, corellation analysis, gene expression DNA micro-arrays, arreglos de metilación de ADN, secuenciaciones de mini-RNA’s, microarreglos SNP, secuenciaciones de exomas, arreglos de proteínas de fase invertida, CIDR, SNN-cliq, spectral clustering o ConK-Means, entre muchos otros. Pues con estos podemos proceder a obtener los subtipos de una variedad de cánceres, incluyendo el cáncer de mama, reduciendo el impacto de complejidad sobre los métodos de agrupación; como es nuestro caso, con métodos de consensus clustering logramos detectar subtipos de cáncer de mama de manera más eficiente y acertada. (Fig. 1).

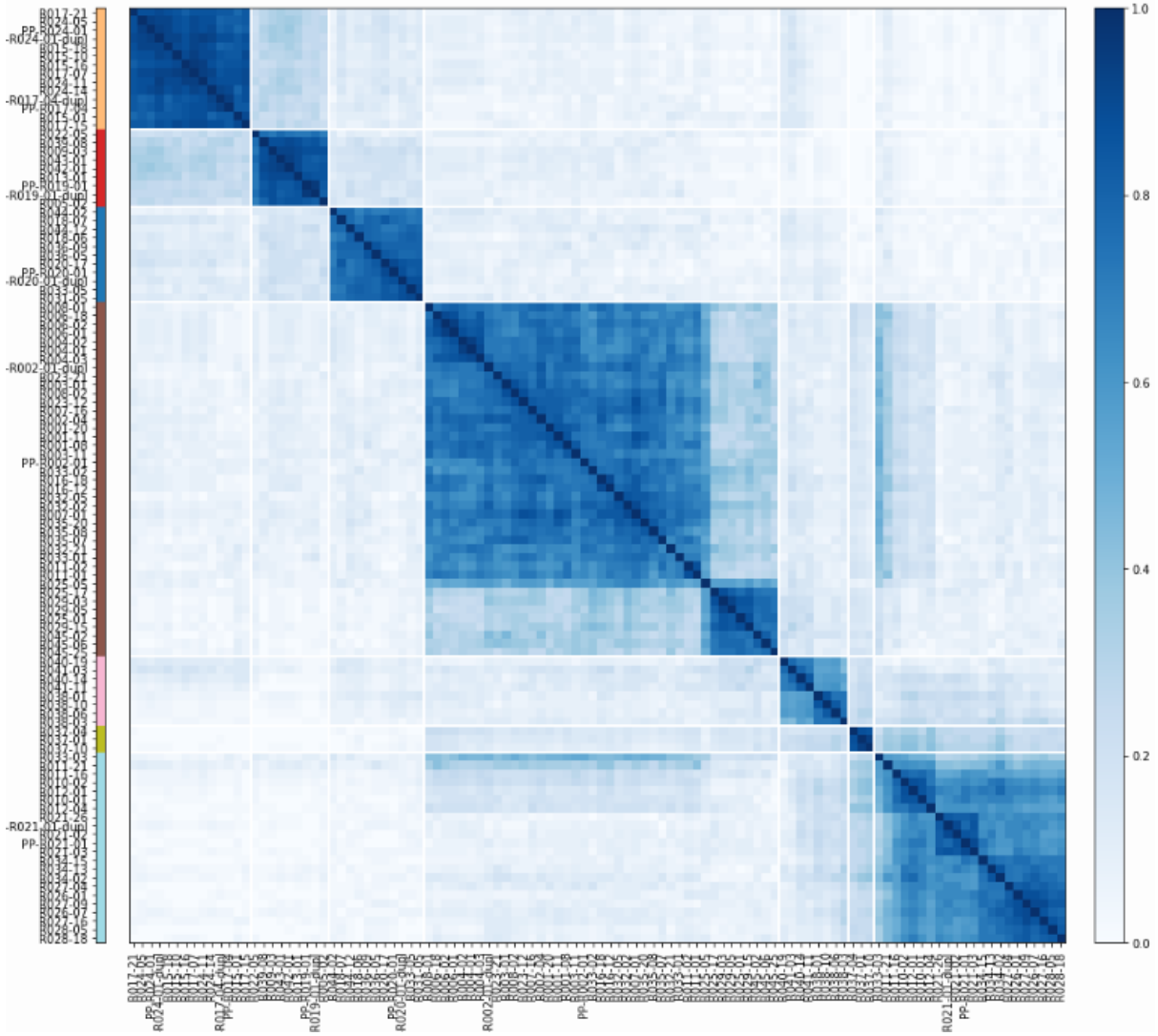

Fig 1. Consenso de clustering por K-Means generado exitosamente por medio de la matriz binaria obtenida con los métodos de la paquetería *pyckmeans* y las métricas de clustering (índice de Calinski Harabasz, e.o.) asociadas a esta.

Para tomar otro punto de vista sobre el método de consensus clustering, podemos observar la categorización proveída por (Gan, Y., Li, N., Zou, G. et al.). (Fig 2).

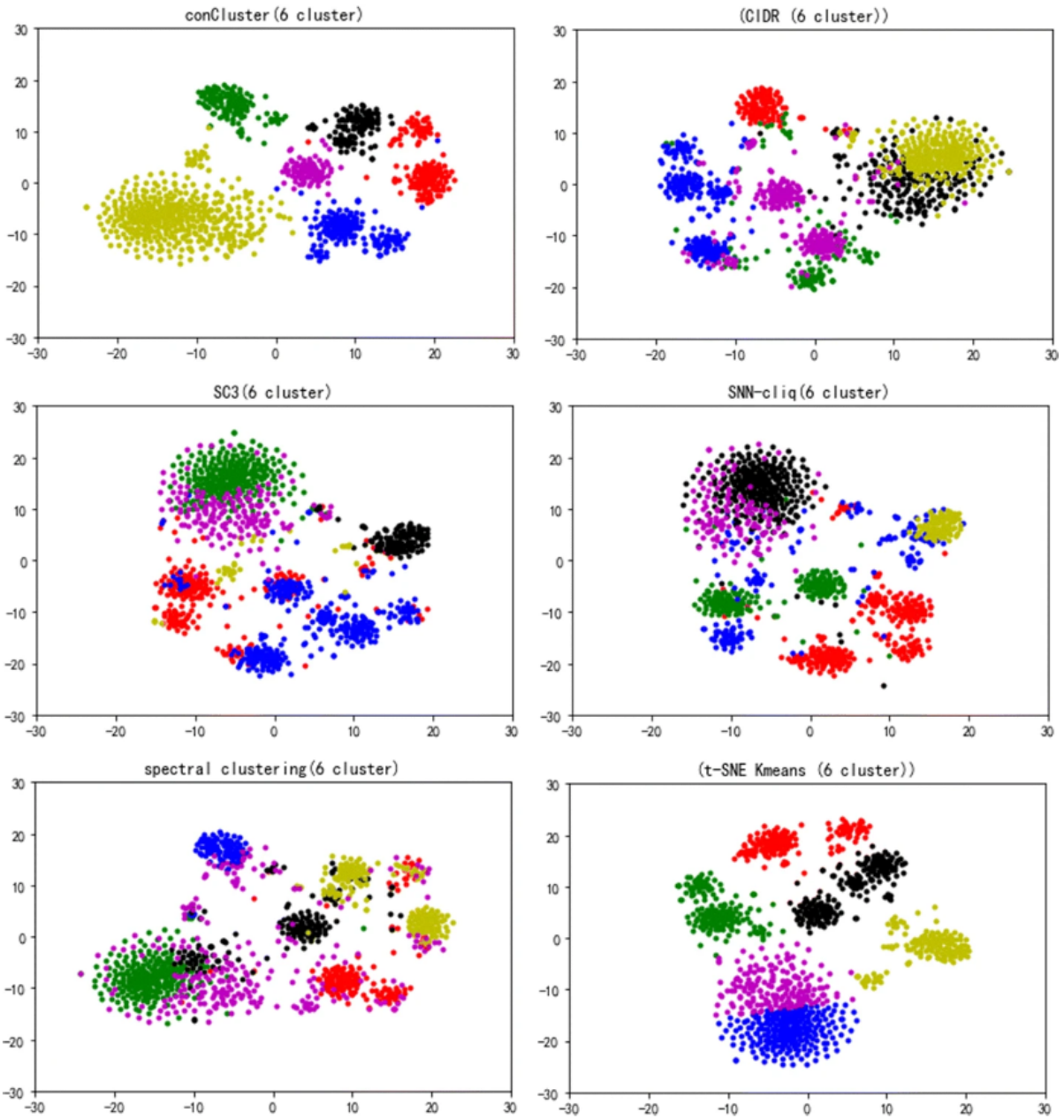

Fig 2. Mismo método de consenso que en la Figura 1, en conjunto con algunos de los métodos mencionados anteriormente. Presentados con una vista de puntos sobre el plano.

Discusión

Manejar este tipo de herramientas y métodos requiere de una gran capacidad de procesamiento computacional pues mantenemos algoritmos que transitan grandes cantidades de datos; hay que hacer uso de métodos y conceptos provenientes de Big Data y Machine Learning (entre otros/otras) para obtener resultados relevantes y utilizables a la hora de procesar secuencias de células tumorales -aún más, de cualquier secuencia genómica que se desee estudiar o analizar-, esto pues no siempre existen algoritmos polinomiales (en tiempo eficiente)

que nos permitan manejar datos en masa y con una cantidad exponencial de posibles variaciones (en nuestro caso, biológicas y tecnológicas).

Sin embargo, podemos hacer uso de estos métodos sobre datos más reducidos para mantener una observación sobre su funcionamiento y funcionalidades, lo que nos permite generar conclusiones acerca de estos y afirmar la existencia de algoritmos heurísticos y estadísticos que ayuden a clasificar nuestros datos en matrices de secuencias individuales.

También, incluso podríamos comparar los métodos para que, dependiendo del caso que se esté estudiando, utilicemos el algoritmo más eficiente para nuestra clasificación como se muestra en la siguiente figura (Fig. 3) obtenida de (Gan, Y., Li, N., Zou, G. et al.).

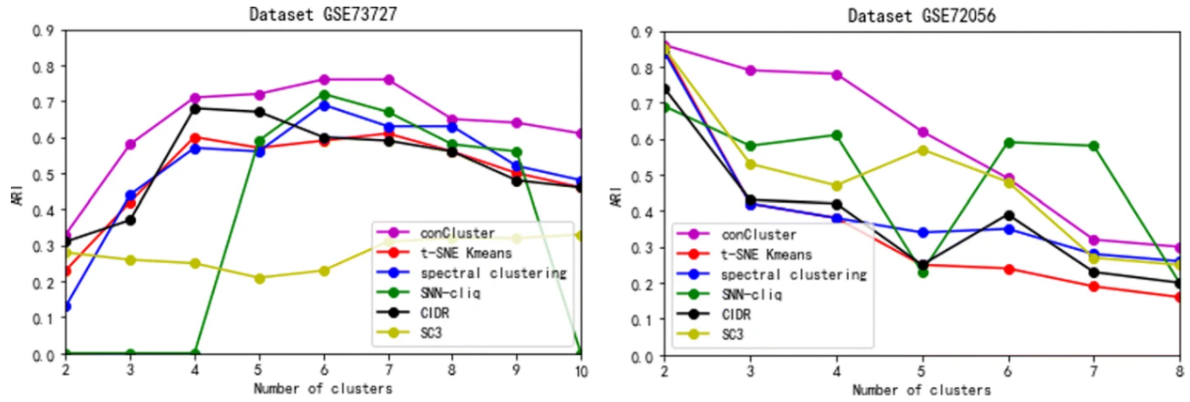

Fig 3. Comparación de rendimiento entre consensus clustering y otros métodos de clasificación de scRNA-seq mencionados anteriormente.

Conclusión

Finalmente, podemos concluir que, dados los métodos mencionados en este documento, sí existen algoritmos eficientes y utilizables que puedan ayudarnos a clasificar subtipos de cáncer de mama y otros tipos de células tumorales. Con esto es claro que nuestra pregunta de investigación es confirmada y existe la posibilidad de implementar estos métodos, en especial el de conClustering (consensus clustering), en un lenguaje de programación de preferencia, en particular -para nuestro caso- en python.

Referencias

- Arzanova, E., & Mayrovitz, H. N. (2022). *The Epidemiology of Breast Cancer*. 1–20. https://doi.org/10.36255/exon-publications-breast-cancer-epidemiology

- Gan, Y., Li, N., Zou, G., Xin, Y., & Guan, J. (2018). *Identification of cancer subtypes from single-cell RNA-seq data using a consensus clustering method*. *11*(S6). https://doi.org/10.1186/s12920-018-0433-z
- Harbeck, N., Frédérique Penault-Llorca, Cortes, J., Gnant, M., Nehmat Houssami, Poortmans, P., Ruddy, K. J., Tsang, J., & Cardoso, F. (2019). *Breast cancer*. *5*(1). https://doi.org/10.1038/s41572-019-0111-2
- Orrantia-Borunda, E., Anchondo-Nuñez, P., Lucero Evelia Acuña-Aguilar, Francisco Octavio Gómez-Valles, & Ramírez-Valdespino, C. A. (2022). *Subtypes of Breast Cancer*. 31–42. https://doi.org/10.36255/exon-publications-breast-cancer-subtypes
- Sylwia Łukasiewicz, Marcin Czeczelewski, Forma, A., Baj, J., Sitarz, R., & Andrzej Stanisławek. (2021). *Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review*. *13*(17), 4287–4287. https://doi.org/10.3390/cancers13174287
- Broad Institute. (2019) *TCGA data version 2016_01_28 for BRCA FireBrowse*