PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE CIENCIAS EXACTAS Y NATURALES

MAESTRIA ONLINE EN BIOLOGIA COMPUTACIONAL

COMPARACION DE DOS MÉTODOS PARA ANALIZAR SINGLE-CELL TRANSCRIPTOMICS EN PLANTAS

Trabajo de Titulación previa a la obtención del título de Magister en Biología Computacional

VIVIANA FERNANDA QUEVEDO TUMAILLI

Quito, Octubre 2023

Derechos de autor

Expreso que soy autor del presente trabajo de titulación y consulté las referencias indicadas en el mismo. Este trabajo no fue presentado de forma previa para la obtención de ningún grado académico. La Pontificia Universidad Católica del Ecuador puede utilizar los derechos del trabajo de titulación de acuerdo con la Ley de Propiedad Intelectual y su normativa institucional.

Viviana Fernanda Quevedo Tumailli

Aprobación del director del Trabajo de Titulación

Certifico que el trabajo de Viviana Fernanda Quevedo Tumailli para la obtención del título de Máster en Biología Computacional se llevó a cabo bajo la normativa y reglamentación institucional y puede ser presentada para su calificación.

Dr. Sergio Alan Cervantes Pérez

A mi esposo y a mis hijos,

a mi padres

a mis hermanos y querida familia

Agradecimientos

En primer lugar, deseo expresar mi agradecimiento al director de este trabajo de titulación el Profesor Dr. Alan Cervantes, por compartir conmigo sus conocimientos, tiempo, experiencia y ser mi guía en el desarrollo de este estudio.

Asimismo, quiero agradecer a sus colaboradores de la Pontificia Universidad Católica del Ecuador, especialmente al Dr. Miguel Ortíz por sus valiosas recomendaciones para la edición del presente trabajo.

A la PUCE y todo su equipo profesional porque me dio la oportunidad a un nuevo mundo de conocimiento, de investigación y así me han permitido cumplir un sueño más.

Deseo expresar también todo mi agradecimiento a mis padres, a mis hermanos, a mi esposo y a mis hijos por su apoyo incondicional y su confianza a lo largo de este trayecto.

A mis demás familiares, amigos y a todas aquellas personas que han estado presentes y han dedicado parte de su tiempo al desarrollo de este trabajo, por su colaboración y paciencia.

¡Muchas gracias a todos!

Índice General

Índice de Figuras

Índice de Tablas

Resumen

La secuenciación de ARN en el campo de la investigación genética, ha permitido comprender, entre muchas cosas, la función de los genes, la regulación génica, la expresión de los genes y los mecanismos de biología molecular. La expresión génica en plantas de la especie *Arabidopsis thaliana* permite estudiar los genes que se activan o desactivan para producir proteínas y otras moléculas que son esenciales para el crecimiento, desarrollo y respuesta a su entorno. Sin embargo, a pesar de contar con una gran cantidad de datos de expresión génica en bases de datos de libre acceso, la limitante común es la falta de profundidad en la comprensión de la expresión génica. El estudio de la expresión génica se ha realizado en las últimas décadas mediante microarreglos de ARN y luego mediante secuenciación de ARN para órganos o tejidos pero no con la resolución del nivel celular. La transcriptómica unicelular permite un análisis más detallado en comparación con la transcriptómica a gran escala, permitiendo analizar las diferencias en la expresión génica entre las células individuales. Uno de los métodos más utilizados para la transcriptómica unicelular es el Método de Protoplastos, que implica la degradación enzimática de la membrana celular, pero se ha visto limitada por la rigidez de la pared celular y el uso de tratamientos enzimáticos agresivos, como alternativa se encuentra el método llamado aislamiento de núcleos que permite separar y extraer los núcleos celulares directamente. En este sentido, se desconoce el impacto de los métodos utilizados en los resultados de transcriptómica unicelular en plantas. En este trabajo, la idea fue comparar la transcriptómica unicelular en condiciones similares de estos 2 métodos mencionados para evaluar las diferencias. Se usaron 5 muestras descargadas de la base de datos GSE155304 del NCBI, 2 muestras pr1 y pr2 para el Método Protoplastos y 3 muestras mr1, mr2 y mr3 para el aislamiento de núcleos con un promedio de células de 4316 y 3198 respectivamente. El presente trabajo ha permitido estudiar diferentes Métodos entre Protoplastos y aislamiento de núcleos para transcriptómica unicelular en plantas. Los conjuntos de datos procesados dan como resultado datos más refinados y de mayor calidad. En los dos métodos se observó una fuerte correlación entre el número de genes y el número de moléculas IMU, ya que los datos son consistentes y confiables. Se observó agrupaciones celulares similares en los métodos individuales, pero diferencias notables al integrar los datos de ambos métodos. El método UMAP se utilizó para reducir la dimensionalidad de los datos y el método de agrupación de Lovaina para identificar agrupaciones de interés. Ambos métodos mostraron resultados muy similares y confiables, por lo que la recomendación sería utilizar el método que más se adecue al tejido/planta de interés.

Abstract

RNA sequencing in the field of genetic research has allowed us to understand, among many things, gene function, gene regulation, gene expression and molecular biology mechanisms. Gene expression in plants of the Arabidopsis thaliana species allows the study of genes that are turned on or off to produce proteins and other molecules that are essential for growth, development and response to their environment. However, despite having a wealth of gene expression data in freely available databases, the common limitation is the lack of depth in understanding gene expression. The study of gene expression has been performed in recent decades by RNA microarrays and then by RNA sequencing for organs or tissues but not at the resolution of the cellular level. Single-cell transcriptomics allows a more detailed analysis compared to large-scale transcriptomics, allowing analysis of differences in gene expression between individual cells. One of the most commonly used methods for single-cell transcriptomics is the protoplast method, which involves enzymatic degradation of the cell membrane, but has been limited by the rigidity of the cell wall and the use of aggressive enzymatic treatments, as an alternative is the method called nuclei isolation that allows the cell nuclei to be separated and extracted directly. In this sense, the impact of the methods used on the results of single-cell transcriptomics in plants is unknown. In this work, the idea was to compare single-cell transcriptomics under similar conditions of these 2 mentioned methods to evaluate the differences. We used 5 samples downloaded from NCBI database GSE155304, 2 samples pr1 and pr2 for the protoplast method and 3 samples mr1, mr2 and mr3 for nuclei isolation with an average cell count of 4316 and 3198 respectively. The present work has allowed us to study different methods between protoplasts and nuclei isolation for single cell transcriptomics in plants. The processed data sets result in more refined and higher quality data. A strong correlation between the number of genes and the number of IMU molecules was observed in the two methods, as the data are consistent and reliable. Similar cell clustering was observed in the individual methods, but notable differences were observed when integrating the data from both methods. The UMAP method was used to reduce the dimensionality of the data and the Louvain clustering method to identify clusters of interest. Both methods showed very similar and reliable results, so the recommendation would be to use the method that best suits the tissue/plant of interest.

1. Introducción

La secuenciación de ARN ha permitido por más de dos décadas el estudio de la expresión de los genes en un genoma. Grandes esfuerzos se han hecho desde entonces para muchos organismos, tanto que hoy en día se cuenta con una gran cantidad de datos de expresión para múltiples tejidos, órganos, condiciones de crecimiento, entre otros en los repositorios internacionales. Sin embargo, una limitante en común para todos estos datos es la falta de profundidad en el entendimiento de la expresión de los genes. Durante los últimos años, la tecnología de single-cell genomics ha emergido en modelos animales como una opción para entender la expresión génica con una resolución a nivel celular. Curiosamente, en plantas esta tecnología apenas ha sido utilizada y dirigida a órganos o tejidos específicos como la raíz en el modelo de estudio *Arabidopsis thaliana*, una planta de la familia Brassicaceae, que es de gran interés para la comunidad científica, utilizados para estudiar la biología en las plantas y la primera planta en tener su genoma completo secuenciado (The Arabidopsis genoma Initiative, 2000). Como a menudo sucede, la investigación en animales se aplica después en plantas, entonces uno de los primeros métodos para hacer single-cell que fue utilizado fue la técnica de protoplastos, que consiste en una degradación enzimática de la membrana celular. Sin embargo, al ser aplicada en plantas la mayor limitante ha sido la pared celular rígida de las células vegetales. Una alternativa a esta técnica es el aislamiento de núcleos, el cual ha dado mejores resultados. Por otro lado, la transcriptómica a gran escala es una poderosa estrategia para relacionar el genotipo con el fenotipo, pero de este proceso se obtiene el promedio de las expresiones del gen en todas las células. Sin embargo, se podría utilizar la transcriptómica unicelular, que es una potente estrategia a escala más fina con diferentes niveles de aproximación para conocer la expresión del gen en cada una de las células. Aunque a esto, poco se sabe acerca de las diferencias transcriptómicas entre las dos técnicas más utilizadas en plantas para hacer single-cell transcriptomics. En este trabajo, se pretende estudiar las diferencias entre las técnicas de protoplastos y aislamiento de núcleos en single-cell transcriptomics.¿Cuántas células es posible obtener en cada técnica?, ¿Se obtienen los mismos tipos de células?, ¿El perfil de expresión de los genes son iguales?, entre otras preguntas. Al responder estas preguntas se podría inferir y recomendar la mejor técnica para cada estudio especifico de plantas.

2. Revisión de la literatura

2.1. Tecnologías para la secuenciación de transcriptomas unicelular

Existen varias tecnologías en el campo de la secuenciación de transcriptomas unicelulares. Estas tecnologías permiten el análisis de células individuales en plantas y brindan información valiosa sobre la expresión génica a nivel de célula única. A continuación, se presentan algunas de las tecnologías:

2.1.1. Smart-Seq

Es un método robusto y reproducible para secuenciar los transcritos de células individuales (Goetz & Trimarchi, 2012). Con Smart-Seq mejoran tanto el tamaño medio de los transcritos como el número de transcritos de longitud completa. Esta técnica permite la secuenciación de transcriptomas de células individuales con alta sensibilidad y reproducibilidad. Permite un análisis más detallados de splicing alternativo y permiten identificar biomarcadores, SNP y mutaciones candidatos (Ramsköld et al., 2012). Permite un alto nivel de lecturas mapeables, mejora la cobertura de los transcritos. No es necesario conocer la secuencia del ARNm.

2.1.2. Smart-seq2

Es una técnica de transcriptómica que permite la secuenciación de transcriptomas a partir de cantidades muy bajas de ARN, incluso de células individuales (Picelli et al., 2013). Permite la amplificación completa del ARN mensajero (ARNm) y el análisis detallado de la expresión génica en cada célula. Proporciona una cobertura de lectura más uniforme a través de los transcritos que los métodos de cola de poli (A), en consonancia con el uso común de cambio de plantilla en aplicaciones diseñadas para capturar los extremos 5’ del ARN8.

2.1.3. Fluidigm C1

Permite aislar simultáneamente docenas de células individuales en función de su tamaño y forma para el análisis de RNA-seq (DeLaughter, 2018). El sistema utiliza un circuito fluídico integrado (IFC) para aislar células individuales en cámaras de reacción individuales. El sistema se utiliza para estudiar la diferenciación celular y caracterizar transcriptomas heterogéneos. (Kim & Marignani, 2022).

2.1.4. Drop-seq

Drop-seq es una técnica que involucra la encapsulación de células individuales en gotas microfluídicas junto con códigos de barras moleculares. Permite la secuenciación de transcriptomas en *Arabidopsis* y el análisis de la expresión génica a nivel de célula única. El poder de esta tecnología radica en el hecho de que, durante la secuenciación, uno puede distinguir de dónde provino la información original de célula a célula, lo que le permite hacer un mapa de expresión génica de la célula o incluso distinguir poblaciones celulares dentro de un tejido. Durante la secuenciación, se puede distinguir de dónde provino la información original de célula a célula, lo que le permite hacer un mapa de expresión génica de la célula o incluso distinguir poblaciones celulares dentro de un tejido (Macosko et al., 2015).

2.1.5. 10x Genomics Chromium

Utiliza una tecnología de microgotas para la secuenciación de células individuales. Permite el análisis de transcriptomas en alta resolución en *Arabidopsis*, generando información sobre la expresión génica a nivel de célula única. El sistema Chromium es un método basado en microfluidos de secuenciación de ARN de una sola célula que permite la secuenciación de una sola célula con su tecnología Next GEM. El sistema 10x Genomics permite apuntar a miles de células por muestra, lo que resulta en un bajo costo por célula en el caso de proyectos de alto rendimiento. El sistema Chromium es un dispositivo que permite la partición de cientos a cientos de miles de celdas individuales en minutos (Gao et al., 2020).

2.1.6. MATQ-seq

Es un método altamente sensible y cuantitativo para la secuenciación unicelular de ARN total incluido el ARN no codificante y no poliadenilado. Captura la variación biológica genuina entre transcriptomas completos de células individuales. Puede detectar la variación transcripcional entre las células de la misma población. Además, elimina el sesgo de la PCR utiizando una estrategia de código de barras molecular (Sheng et al., 2017).

2.1.7. Seq-Well

Seq-Well es una tecnología que utiliza una matriz de pocillos microfluídicos para la secuenciación de células individuales. Permite el análisis de transcriptomas de

Arabidopsis a nivel de célula única y el estudio de la expresión génica. Es un scRNA-seq portátil y de bajo costo diseñado para muestras clínicas de bajo volumen y entornos independientes de los recursos (Gierahn et al., 2017).

2.1.8. CEL-seq

Proporciona resultados más reproducibles, lineales y snsibles que un método de amplificación basado en PCR. Un protocolo que satisface la demanda de amplificación lineal por transcripción in vitro para material suficiente agrupando muestras con código de barras, permitiendo así la amplificación lineal eficiente de ARN de células individuales y su análisis por secuenciación (Hashimshony et al., 2012)**.**

2.1.9. MARS-seq

Diseñado para el muestreo in vitro de miles de células mediante RNA-seq multiplexado manteniendo un estricto control sobre los sesgos de amplificación y los errores de etiqutado (Jaitin et al., 2014). Perfila la dinámica transcripcional de células individuales en un flujo de trabajo paralelo masivo y automatizado con alta resolución. Este método ofrece alta sensibilidad, bajo ruido técnico y bajo costo por célula.

2.1.10. inDrop-seq

Es un método que utiliza microesferas de hidrogel para introducir los oligonucleótidos con código de barras en células individuales. Las células se encapsulan en microesferas de hidrogel y los oligonucleótidos con código de barras se introducen en las microesferas. Luego, las microesferas se abren y el ARN se transcribe y amplifica de manera inversa. El método se utiliza para el etiquetado de una sola célula de alto rendimiento. es un método poderoso para la secuenciación de ARN de una sola célula que ofrece un alto rendimiento y un bajo costo por célula (Klein et al., 2015) (M. Klein & Macosko, 2017)**.**

2.1.11. DNBelab C4

Es un método que permite la creación de perfiles transcripcionales de una sola célula de alto rendimiento. Tiene enormes beneficios en costo y portabilidad (Liu et al., 2019).

2.2. Transcriptómica de célula única vs. la secuenciación de ARN en masa

La principal diferencia entre la técnica transcriptómica de célula única (single-cell transcriptomics) y la secuenciación de ARN en masa (bulk RNA-seq) radica en el nivel de resolución con el que se analizan las expresiones génicas en una muestra biológica. A continuación se mencionan algunas diferencias clave y las ventajas de la técnica de single-cell transcriptomics:

Resolución a nivel de célula única: En secuenciación de ARN en masa se analiza el promedio de expresiones génicas de todas las células presentes en una muestra. Esto proporciona información sobre la expresión promedio de los genes en la población celular, pero no revela la variabilidad entre células individuales. En cambio, con la técnica transcriptómica de célula única, se analiza el perfil de expresión génica de cada célula individualmente en una muestra. Esto permite capturar heterogeneidades celulares, identificar subtipos celulares, detectar células raras y comprender las relaciones entre ellas de manera mucho más detallada.

Identificación de subtipos celulares y estados transicionales: En secuenciación de ARN en masa se puede proporcionar información sobre los tipos celulares predominantes presentes en una muestra, pero no se puede distinguir entre subtipos celulares o estados de diferenciación en el mismo nivel de detalle que la transcriptómica de célula única. En cambio la técnica transcriptómica de célula única permite la identificación y caracterización de subtipos celulares y estados transicionales, lo que es fundamental para comprender mejor la diversidad celular en un tejido o sistema.

Células raras y eventos biológicos sutiles: En secuenciación de ARN en masa las células raras o minoritarias pueden quedar enmascaradas por las células más abundantes en un análisis bulk, mientras que en transcriptómica de célula única, facilita la detección y el estudio de células raras, que pueden ser cruciales en el contexto de enfermedades o procesos biológicos específicos.

Detección de heterogeneidad intracelular: En la secuenciación de ARN en masa no permite discernir diferencias en la expresión génica dentro de una población celular homogénea, mientras que en transcriptómica de célula única, revela la heterogeneidad de expresión génica dentro de una población de células aparentemente homogénea, lo que es esencial para comprender la variabilidad celular y los mecanismos reguladores.

Mejora en la comprensión de redes de regulación génica: En cuanto a la técnica transcriptómica de célula única, facilita el análisis de interacciones entre genes y la inferencia de redes de regulación génica a nivel de célula única, lo que puede proporcionar información sobre las vías biológicas activas en diferentes tipos celulares y condiciones.

2.3. Métodos actuales para aislar las células

2.3.1. Protoplastos

Un protoplasto es una célula vegetal a la cual se le ha removido completamente la pared celular, mediante métodos mecánicos o enzimáticos. Observados al microscopio, los protoplastos presentan una forma esférica, con el citoplasma rodeado por una membrana plasmática; en el interior de la célula se puede visualizar el núcleo. Es posible obtener protoplastos a partir de cualquier órgano, tejido y tipo de planta.

En el mercado existen preparaciones comerciales de enzimas para aislar protoplastos (pectinasas, celulasas o hemicelulasas), cuyas fuentes pueden ser de origen fúngico o bacteriano. Usando una mezcla de estas enzimas se puede hidrolizar las paredes celulares y obtener altos rendimientos de protoplastos viables, adecuados para diferentes fines en la investigación.

2.3.2. Aislamiento de núcleos

El aislamiento de ARN es la extracción de ácido ribonucleico de las células, que luego se utiliza en una serie de experimentos y procedimientos. Los investigadores y científicos suelen utilizar el proceso de aislamiento de ARN para estudiar una variedad de enfermedades y funciones celulares. Este es un proceso difícil que requiere atención a los detalles y precaución para proteger el ARN.

Numerosos laboratorios llevan a cabo este proceso y se ofrecen a la venta numerosos kits como kits de aislamiento de ARN.

Si bien el proceso de aislamiento del ARN no es sencillo, con el paso de los años se ha simplificado. Para evitar daños, el ARN debe maniobrar adecuadamente durante el procedimiento. Antes de aislar el ARN del resto de la célula, los científicos lo mantienen intacto utilizando productos químicos especializados y técnicas de trituración celular. Luego, para romper las paredes celulares y eliminar la protea que rodea el ARN, se utiliza una solución especializada hecha de enzimas. Si el científico no quiere estudiar ambos al mismo tiempo, el ADN puede eliminarse utilizando una solución diferente.

2.4. Herramientas tecnológicas usadas en este estudio

2.4.1. Tecnología 10x Gemomics

10x Genomics es una empresa estadounidense de biotecnología que diseña y fabrica tecnología de secuenciación de genes utilizada en la investigación científica. Ofrece herramientas poderosas y confiables que impulsan los descubrimientos científicos e impulsan el progreso exponencial para dominar la biología y mejorar la salud humana.

La tecnología 10x Genomics se puede utilizar para el análisis de células individuales en una amplia variedad de organismos, incluyendo la planta modelo *Arabidopsis thaliana*. Algunos enfoques comunes para el análisis de células individuales en *Arabidopsis* utilizando la tecnología 10x Genomics incluyen:

1.- Análisis de expresión génica a escala de célula única: La tecnología 10x Genomics permite la secuenciación de ARNm a partir de células individuales, lo que permite la identificación de genes que se expresan en cada célula y la cuantificación de sus niveles de expresión. Este enfoque se ha utilizado para identificar subpoblaciones de células en diferentes tejidos de *Arabidopsis* y para caracterizar los cambios en la expresión génica durante el desarrollo y la respuesta al estrés (Denyer et al., 2019; Petricka et al., 2019).

2.- Análisis de la variación genética: La tecnología 10x Genomics también se puede utilizar para identificar la variación genética a escala de célula única. Esto se ha utilizado en estudios para identificar mutaciones somáticas en células individuales de *Arabidopsis* y para estudiar la variabilidad genética en poblaciones de células (Klionsky et al., 2019).

3.- Análisis de la interacción celular: La tecnología 10x Genomics también se ha utilizado para estudiar la interacción celular en *Arabidopsis*. Esto se ha utilizado para identificar interacciones entre células de diferentes tipos y para estudiar la formación de estructuras tridimensionales en diferentes tejidos (Sukumar et al., 2019).

2.4.2. Biomage (Breathing life into biological data)

Biomage se podría traducir al español como “Dando vida a los datos biológicos”. Es una plataforma de acceso gratuito que se ha desarrollado en la Escuela de Medicina de Harvard. Permite el análisis de datos biológicos celulares y moleculares. Es una plataforma que apoya a la investigación científica y el nuevo descubrimiento de nuevos conocimientos a través de datos biológicos complejos.

Los investigadores pueden visualizar, analizar y compartir datos de células y tejidos a gran escala. A través de algoritmos de Machine Learning y el análisis estadísticos avanzados pueden ayudar a identificar patrones, relaciones y características en los datos y generar nuevas hipótesis para la investigación.

2.4.3. Cellenics

Es una herramienta bioinformática de Biomage para el análisis de datos de la secuenciación de ARN de una sola célula (scRNA-seq). Está basada en la nube, trabaja en el espacio virtual desde el navegador en el que se puede almacenar por ejemplo, las muestras de RNA y trabajar directamente en los archivos sin descargar software alguno. La herramienta es de código abierto para datos scRNA-seq, permite explorar y analizar un conjunto de datos sin tener que ser experto en programación. Principalmente tiene módulos como el procesamiento de datos, exploración de datos, y gráficos y tablas.

2.4.4. CellRanger

Es una herramienta de análisis de datos de células individuales de 10x Genomics que se utiliza para procesar los datos de secuenciación de ARNm a partir de células individuales. La herramienta Cell Ranger se ha utilizado para analizar datos de células individuales en una variedad de organismos, incluyendo *Arabidopsis* thaliana.

El proceso de análisis de datos con Cell Ranger para *Arabidopsis* sigue el mismo procedimiento general que para otros organismos, pero con algunas consideraciones específicas para esta planta. Para realizar el análisis de células individuales de *Arabidopsis* con Cell Ranger, se deben seguir los siguientes pasos:

- Preprocesamiento de los datos de secuenciación: los datos de secuenciación de ARNm a partir de células individuales de *Arabidopsis* se deben preprocesar para eliminar las lecturas de baja calidad y adaptadores.

- Mapeo de lecturas: Las lecturas se alinean al genoma de *Arabidopsis* utilizando un algoritmo de alineación de lecturas.

- Identificación de células individuales: Las células individuales se identifican y se agrupan utilizando un análisis de agrupamiento de células.

- Análisis de expresión génica: Los datos de expresión génica se cuantifican utilizando un análisis de expresión génica de célula única.

- Análisis de enriquecimiento de GO: Se pueden utilizar herramientas como clusterProfiler para identificar los genes y funciones biológicas enriquecidas en diferentes subpoblaciones de células.

Al procesar los datos de secuenciación de múltiples muestras con Cell Ranger, es posible detectar y corregir errores y variaciones técnicas, lo que garantiza la precisión y reproducibilidad de los resultados obtenidos.

Cell Ranger se ha utilizado para analizar datos de células individuales de *Arabidopsis* en varios estudios, como el análisis de células del meristema radical de *Arabidopsis* (Denyer et al., 2019) y la identificación de subpoblaciones de células en hojas y raíces (Klionsky et al., 2019).

Al procesar los datos de expresión génica se generan 3 archivos por muestra:

- barcodes.tsv: contiene la lista de códigos de barras (UMI, Identificadores Únicos de Moléculas, por sus siglas en inglés) que están disponibles en la muestra.
- características.tsv o genes.tsv: contiene la lista de características/genes que se reconocen.
- matrix.mtx: representa una matriz que contiene el número de transcritos detectados, con genes a lo largo de la fila y códigos de barras/celdas a lo largo de la columna. Esto se llama una matriz de conteo .

2.4.5. R

Es un entorno y lenguaje de programación con un enfoque al análisis estadístico. Es uno de los más utilizados en investigación científica, siendo además muy popular en los campos de aprendizaje automático, minería de datos, investigación biomédica, bioinformática y matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes con funcionalidades de cálculo y graficación.

2.5. Especie *Arabidopsis Thaliana*

La *Arabidopsis* es un género de plantas herbáceas de la familia de las brasicáceas, que han sido objeto de intenso estudio en época reciente como modelos para la investigación fitobiológica. *Arabidopsis thaliana* fue la primera planta cuyo genoma se secuenció por completo, una tarea completada en diciembre del 2000 por el proyecto AGI (Iniciativa para el Genoma de la *Arabidopsis,* por sus siglas en ingles).

Arabidopsis thaliana es una especie de planta con flor importante para la investigación en biología vegetal debido a su tamaño y ciclo de vida, su genoma secuenciado, la disponibilidad de herramientas genéticas y su importancia biológica.

Cuenta con cinco cromosomas y 125 millones de pares de bases por lo que es relativamente pequeño comparado con otras especies vegetales facilitando su estudio. Estas características hacen que *Arabidopsis* sea un modelo útil para el estudio de la biología molecular y la genética de las plantas.

A continuación, se detallan las razones mencionadas anteriormente:

Tamaño y ciclo de vida: *Arabidopsis* es una planta pequeña que tiene un ciclo de vida corto, lo que significa que se puede cultivar en grandes cantidades en un espacio relativamente pequeño y en un tiempo relativamente corto. Esto hace que sea fácil para los investigadores hacer estudios a gran escala.

Genoma secuenciado: El genoma de *Arabidopsis* thaliana fue secuenciado a principios de los años 2000, lo que significa que se puede estudiar la genética y la biología molecular de esta planta con un nivel de detalle sin precedentes.

Disponibilidad de herramientas genéticas: Los investigadores han desarrollado una amplia gama de herramientas genéticas para trabajar con *Arabidopsis*, incluyendo mutantes, líneas de expresión génica y técnicas de edición genómica.

Importancia biológica: Aunque *Arabidopsis* es una planta pequeña y relativamente simple, comparte muchas características biológicas con otras plantas más grandes y complejas. Por lo tanto, los resultados obtenidos a partir de estudios en *Arabidopsis* pueden aplicarse a una amplia variedad de plantas.

2.5.1. Gráficos estadísticos

2.5.1.1. t-SNE (T-distributed Stochastic Neighbor Embedding) y/O UMAP (Uniform Manifols Approximation and Projection).

Es un algoritmo de aprendizaje automático para la visualización de datos. Los gráficos t-SNE (Incrustación de vecinos estocásticos distribuidos en T, por sus siglas en inglés) modelan cada objeto de alta dimensionalidad por un punto de 2 o 3 dimensiones de manera que los objetos similares se modelan por puntos cercanos

y los objetos dispares se modelan por puntos distantes con alta probabilidad. Las visualizaciones producidas por t-SNE son significativamente mejores que las producidas por otras técnicas en casi todos los conjuntos de datos (Com & Hinton, 2008).

Por otro lado, UMAP (Aproximación y proyección de variedad uniforme, por sus siglas en inglés) es un algoritmo similar al t-SNE para reducciones dimensionales no lineal (McInnes et al., 2018) para clusterizar.

UMAP se construye a partir de un marco teórico basado en la geometría riemanniana y la topología algebraica. El resultado es un algoritmo escalable práctico que se aplica a datos del mundo real. El algoritmo UMAP es competitivo con t-SNE en cuanto a calidad de visualización y podría decirse que conserva más de la estructura global con un rendimiento de tiempo de ejecución superior. Además, UMAP no tiene restricciones computacionales en la dimensión incrustada, lo que lo hace viable como una técnica de reducción de dimensión de propósito general para el aprendizaje automático (McInnes et al., 2018).

2.5.1.2. Diagrama de puntos

Es un gráfico estadístico que consta de puntos de datos trazados en una escala bastante simple, generalmente usando círculos rellenos. Se utiliza para visualizar la distribución de datos en un conjunto de muestras. Es una forma gráfica de representar los valores individuales de una variable en un eje y los valores de otra variable en el eje x. Cada punto representa un valor individual en el conjunto de datos.

El diagrama de puntos es particularmente útil para mostrar la variabilidad en los datos y detectar posibles valores atípicos. También puede ser utilizado para comparar la distribución de datos en diferentes grupos.

Una forma común de utilizar el diagrama de puntos es en el análisis de secuenciación de ARN de una sola célula en plantas (scRNA-seq) para visualizar la distribución de la expresión génica en diferentes células individuales. Al igual que

en otros organismos, el diagrama de puntos se utiliza para representar los valores de expresión de un solo gen en diferentes células individuales (Zhang, H., et al., 2021).

En el diagrama de puntos para scRNA-seq en plantas, cada punto representa una célula y su posición en el eje vertical indica el nivel de expresión del gen en esa célula en particular. Los puntos pueden ser coloreados para indicar diferentes grupos o subtipos de células.

2.5.1.3. Mapa de calor

Es una técnica de visualización de datos que muestra la magnitud de un fenómeno en forma de color en dos dimensiones. La variación de color puede ser por matiz o intensidad, dando pistas visuales obvias al lector sobre cómo el fenómeno se agrupa o varía en el espacio.

Los mapas de calor tienen una amplia gama de posibilidades entre las aplicaciones debido a su capacidad para simplificar los datos y hacer que el análisis de datos sea visualmente atractivo para leer. En el campo biológico, los mapas de calor se utilizan para representar visualmente conjuntos de datos grandes y pequeños. La atención se centra en patrones y similitudes en el ADN, el ARN, la expresión génica, etc. Al trabajar con estos conjuntos de datos, los científicos de datos en bioinformática, se centran en diferentes conceptos, algunos de los cuales son la detección comunitaria, la asociación y la correlación, y el concepto de centralidad, donde los mapas de calor son una forma convincente de resumir visualmente los resultados y compartirlos con otras profesiones que no pertenecen al campo de la biología o la bioinformática. Particularmente, se utiliza en el análisis de datos de secuenciación de ARN de una sola célula en plantas que consiste en una representación visual de los niveles de expresión génica de una gran cantidad de genes en diferentes células individuales, donde cada célula se representa como una fila y cada gen se representa como una columna en la matriz de datos.

El mapa de calor muestra la expresión de genes específicos en diferentes células, lo que permite a los investigadores identificar patrones y correlaciones entre las células y los genes. La intensidad del color en cada celda del mapa de calor refleja el nivel de expresión del gen en esa célula. Los genes que se expresan a niveles más altos aparecen en tonos más oscuros, mientras que los genes que se expresan a niveles más bajos aparecen en tonos más claros.

3. Metodología

Se descargaron las muestras de secuenciación de ARN de la especie *Arabidopsis thaliana* de la base de datos GSE155304 del National Center of Biotechnology information (GEO Accession viewer (nih.gov)), dos muestras renombradas como pr1 y pr2 que corresponden al método protoplastos y tres muestras mr1, mr2 y mr3 que corresponden al método de aislamiento de núcleos. De estos datos de secuenciación en formato FASTQ se generaron tres archivos para cada muestra: barcodes.tsv, features.tsv y countmatrix.mtx a traves de Cell Ranger de la tecnología 10x Genomics Chromium. Finalmente estos procesos son cargados en la herramienta Cellenics, procesados, analizados y sus resultados son visualizados para compararlas entre los dos métodos protoplastos y el aislamiento de núcleos (Figura 1)

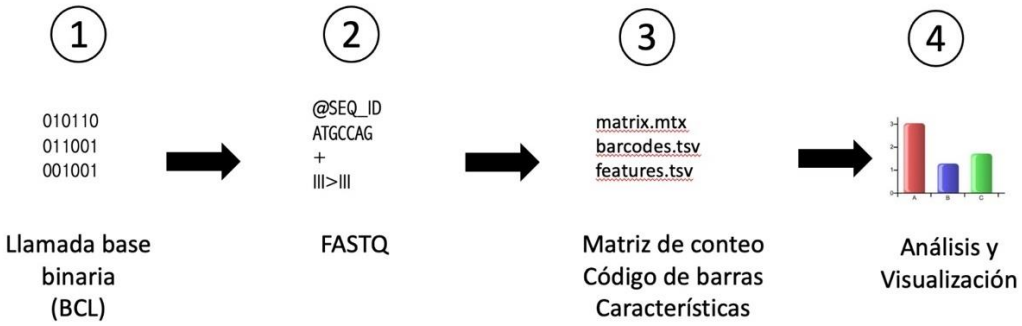

Figura 1. Flujo de trabajo del análisis de datos de secuenciación unicelular

Se procedió a crear tres enfoques en la herramienta Cellenics para cargar los datos. Cada enfoque corresponde a cada método, el primero corresponde a la exploración de expresión a nivel celular utilizando datos (pr1, pr2) con protoplastos, el segundo método corresponde a la exploración de la misma muestra pero utilizando el enfoque de núcleos (mr1, mr2, mr3). El tercer enfoque se ha creado combinando las muestras de ambos métodos protoplastos y aislamiento de núcleos (pr1 y pr2, mr1, mr2, mr3) con el fin de

compararlas para ver las diferencias. Las muestras en forma de carpetas se cargaron en la plataforma Cellenics para procesar, analizar, visualizar y dar sentido a los datos de la secuenciación de ARN de la especie *Arabidopsis thaliana*. Cada carpeta de datos proporciona tres archivos: barcodes.tsv, genes.tsv y matrix.mtx. Se añadieron metadatos para cada enfoque, principalmente en el tercer enfoque que contiene muestras múltiples, es decir, “protoplasto” y “aislamiento de núcleos”.

Una vez cargado los datos, se procedió a procesar los datos, este proceso es importante para limpiar los datos eliminando las gotas vacías, las células muertas, los dobletes y las células de mala calidad con la finalidad de asegurar que los datos procesados sean de alta calidad y arrojen resultados precisos durante el análisis posterior. El procesamiento de datos constó de siete pasos secuenciales, filtro clasificador, filtro de distribución del tamaño de celda, filtro de contenido mitocondrial, filtro de número de genes frente a UMI, filtro doblete, integración de datos y configuración de la incrustación. La salida de cada paso en este módulo se convierte en la entrada para el siguiente paso. Los pasos 1 hasta 5 consisten en filtros para eliminar datos no deseados y de mala calidad de cada muestra individual. En el paso 6, se integran varios conjuntos de datos de muestra para eliminar los efectos por lotes y se realiza la reducción de la dimensionalidad. Finalmente, en el paso 7, se configura la incrustación (por ejemplo, UMAP o t-SNE) y se aplica la agrupación.

Después del procesamiento de datos, se procedió a explorar los datos usando diferentes resultados, tales como, la incrustación de UMAP, la lista de genes presentes en el conjunto de datos ordenados por dispersión, reporte de los genes más variables del conjunto de datos y el mapa de calor mostrando los genes marcadores.

Finalmente, se procedió a visualizar y analizar las tablas y los gráficos para obtener información de sus datos de forma rápida y sencilla. Gráficos tales como, incrustación categórica, gráfico de frecuencia y gráfico de trayectoria. También se puede observar gráficos que representan la expresión de genes individuales en conjunto de células, como diagrama de violín, diagrama de puntos, entre otros.

4. Resultados

Se crearon tres enfoques diferentes en la plataforma Cellenics (Figura 2, Figura 3, Figura 4). En cada enfoque se usaron metadatos, por ejemplo, las muestras dentro del conjunto de datos se han asignado como “protoplasto” y/o “aislamiento” en la pista de metadato llamado “método” (Figura 4). La asignación de metadatos permitió la comparación de grupos para determinar genes expresados diferencialmente y la visualización de grupos en la plataforma.

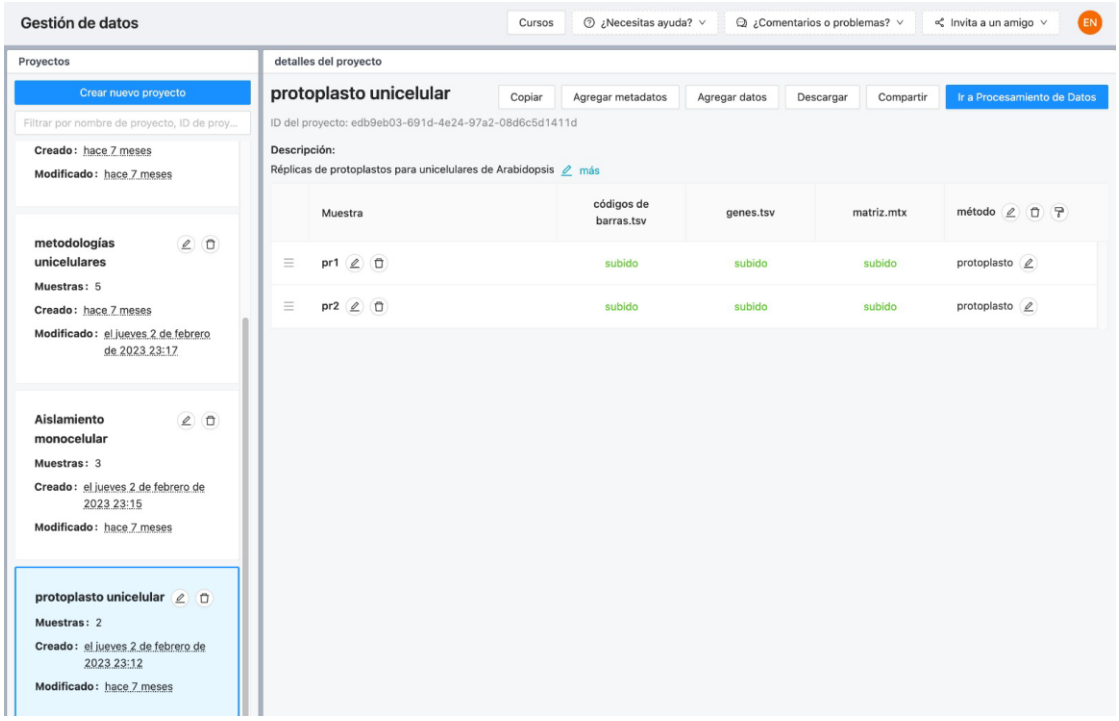

Figura 2. Enfoque 1: Método de protoplastos con 2 muestras pr1 y pr2

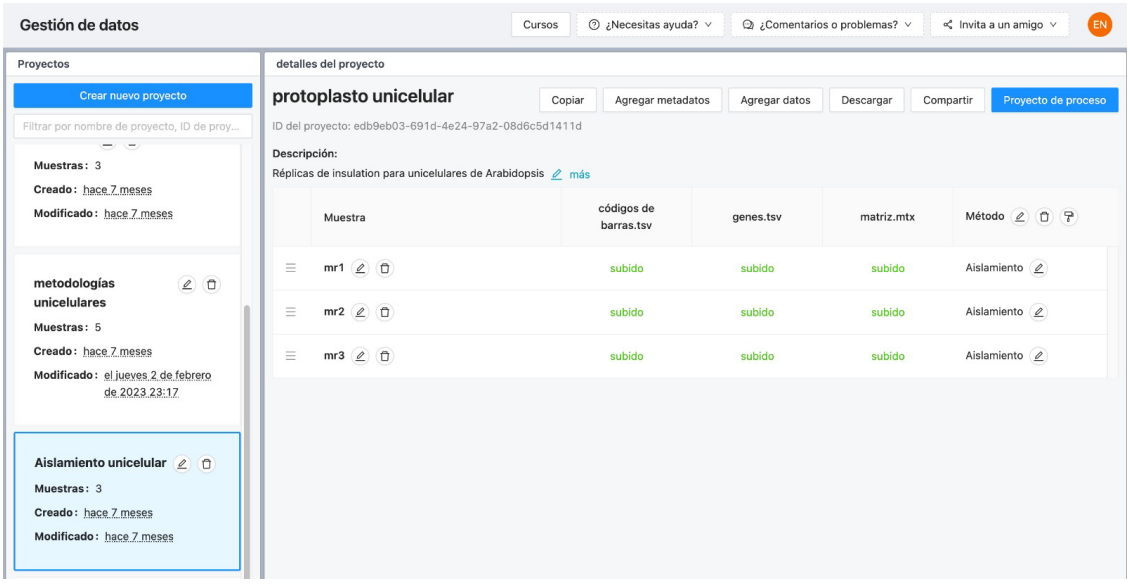

Figura 3. Método de aislamiento de núcleos con 3 muestras mr1, mr2 y mr3

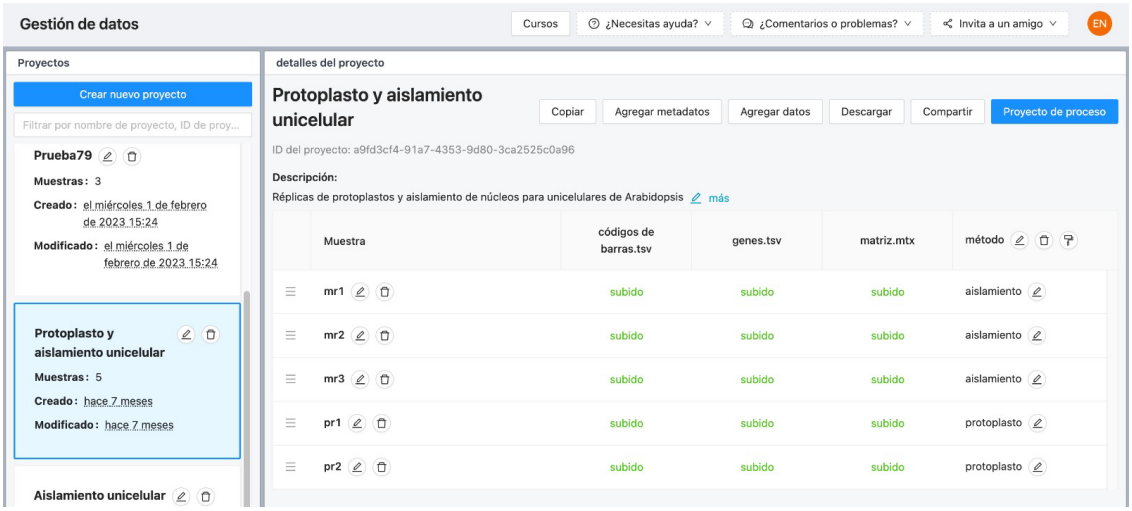

Figura 4. Combinación entre Protoplastos y aislamiento de núcleos con 5 muestras mr1, mr2, mr3, pr1 y pr2

A continuación, se detalla el procesamiento de datos en la plataforma Cellenics en tres enfoques diferentes.

4.1. Procesamiento de datos en el Enfoque 1: Método de Protoplastos

Antes de empezar el procesamiento de datos del enfoque 1, los números estimados de células de *Arabidopsis thaliana* de las dos muestras del Método de Protoplastos pr1 y pr2 cuentan con 4196 (Tabla 1) y 4437 (Tabla 2) células respectivamente.
En el paso 1 "**filtro clasificador**" se mantienen los números estimados de células descritos en el párrafo anterior después del paso de filtrado, al igual que el número total de genes iguales a 23815/23750 para pr1/pr2. Así como también, no hay cambios de proporción para los valores de la mediana de genes por célula y la mediana de recuentos UMI por celda, es decir se mantienen en 0.000% para el porcentaje de cambios (Tabla 1 y Tabla 2). Esto significa que todas las células se retienen niguna se filtra, es decir, no hay gotas vacías en el conjunto de datos, todas son válidas para el análisis.

Tabla 1. Estadística de filtrado para la muestra pr1 del método protoplastos

Estadísticas	# antes	# después	% cambió
Número estimado de células	4196	4196	0.000
Número total de genes	23815	23815	0.000
Número medio de genes por célula	1466.5	1466.5	0.000
Recuentos medios de UMI por celda	3908.5	3908.5	0.000

Tabla 2. Estadística de filtrado para la muestra pr2 del método protoplastos

Estadísticas	# antes	# después	% cambió
Número estimado de células	4437	4437	0.000
Número total de genes	23750	23750	0.000
Número medio de genes por célula	1650	1650	0.000
Recuentos medios de UMI por celda	4757	4757	0.000

Al igual que en el enfoque 1, algunos pasos del procesamiento de datos no fueron tomados en cuenta por varias razones, en el paso 2 "**filtro de distribución del tamaño de celda"** muestran los mismos resultados de manera gráfica donde se

detectan nuevamente gotas vacías y se afina el paso 1 y en el paso 3 **"filtro de contenido mitocondrial"** no se visualizó ningún gráfico porque no hay datos de secuencias mitocondriales.

Luego de aplicarse el paso 4 **"filtro de número de genes frente a UMI"** se visualiza un decremento del 0.143%, de 4196 se obtuvo 4190 células para pr1. Esto quiere decir que 6 células se eliminaron y no pasaron este filtro (Figura 5). De igual manera, existe un decremento del 0.158% para pr2, es decir, de 4437 se modifica a 4430 células, 7 células no pasaron el filtro (Figura 6). En las Tablas 3 y 4 se puede observar el porcentaje de pr1 para el número total de genes igual a 0.004%, para pr2 no hubo cambios. En cuanto, a los valores de la mediana de genes por célula así como también la mediana de recuentos de UMI por celda ha habido un incremento del 0.170% y 0.102% respectivamente para pr1. Para pr2 el incremento fue del 0.273% para la mediana de genes por celular y 0.116% para la mediana de recuentos de UMI por celda.

Las Figuras 5 y 6 muestran la correlación entre el número de genes y el número de moléculas UMI. La correlación es cercana a 1 lo que significa que los datos están bien representados.

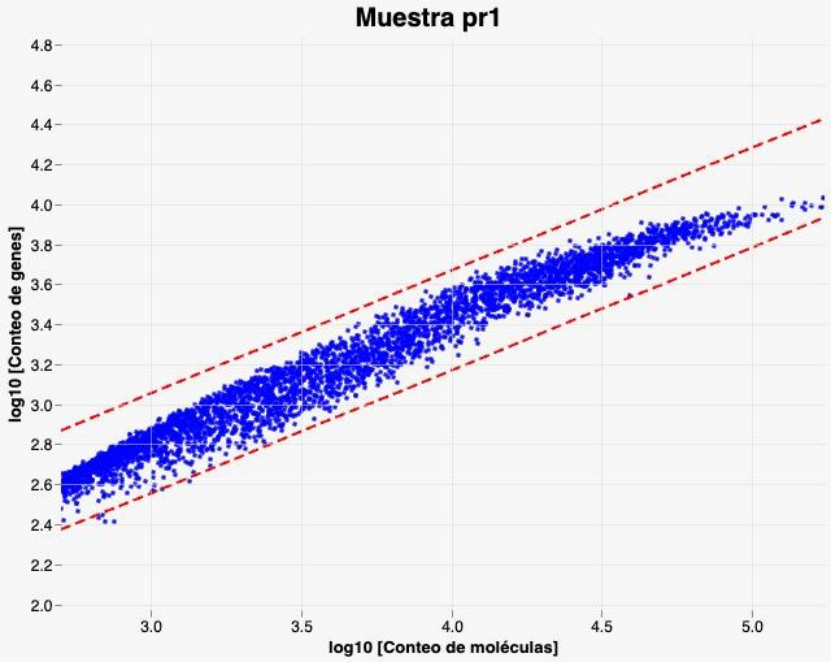

Figura 5. Fuera de la línea roja se muestran las 6 células que se eliminaron en la muestra pr1

Tabla 3. Valores de la primera muestra (pr1) del método protoplastos después de aplicar el filtro de número de genes frente a UMI

Estadísticas	# antes	# después	% cambió
Número estimado de células	4196	4190	-0.143
Número total de genes	23815	23814	-0.004
Número medio de genes por célula	1466.5	1469	+0.170
Recuentos medios de UMI por celda	3908.5	3912.5	+0.102

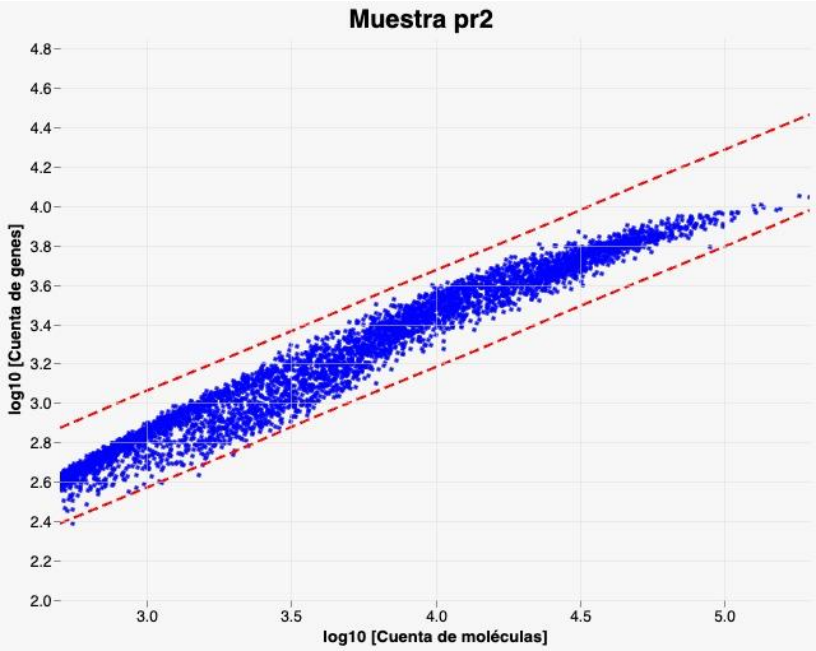

Figura 6. Fuera de la línea roja se muestran las 7 células que se eliminaron en la muestra pr2

Tabla 4. Valores de la segunda muestra (pr2) del método protoplastos después de aplicar el filtro de número de genes frente a UMI

Estadísticas	# antes	# después	% cambió
Número estimado de células	4437	4430	-0.158
Número total de genes	23750	23750	0.000
Número medio de genes por célula	1650	1654.5	+0.273
Recuentos medios de UMI por celda	4757	4762.5	+0.116

Luego de aplicarse el paso 5 **“filtro doblete”** se visualiza un decremento del 7.61%, de 4190 se obtuvo 3871 células para pr1. Esto quiere decir que 319 células se eliminaron y que no pasaron este filtro. De igual manera, existe un decremento del 7.81% para pr2, es decir, de 4430 se modifica a 4084 células, 346 células no pasaron el filtro. Esto puede suceder, porque a veces dos o más células pueden acabar en la misma gota causando problemas en el análisis posterior. En las Tablas 5 y 6 se puede observar el porcentaje de pr1 y pr2 para el número total de genes igual a -0.91% y -0.72% y con una pérdida de genes de 217 y 173 respectivamente. Los valores de la mediana de genes por célula así como también la mediana de recuentos de UMI por celda ha habido un decremento del 8.85% y 13.78% respectivamente para pr1. Para pr2 el decremento fue del 7.76% para la mediana de genes por célula y 9.29% para la mediana de recuentos de UMI por celda.
La Figura 7 y Figura 8 se muestran semejantes donde el umbral de probabilidad para pr1 es 0.55240 y pr2 es 0.50592.

Tabla 5. Valores de la primera muestra (pr1) del método protoplastos después de aplicar el filtro doblete

Statistics	# before	# after	% changed
Estimated number of cells	4190	3871	-7.613
Total number of genes	23814	23597	-0.911
Median number of genes per cell	1469	1339	-8.850
Median UMI counts per cell	3912.5	3373	-13.789

Tabla 6. Valores de la segunda muestra (pr2) del método protoplastos después de aplicar el filtro doblete

Estadísticas	# antes	# después	% cambió
Número estimado de células	4430	4084	-7.810
Número total de genes	23750	23577	-0.728
Número medio de genes por célula	1654.5	1526	-7.767
Recuentos medios de UMI por celda	4762.5	4320	-9.291

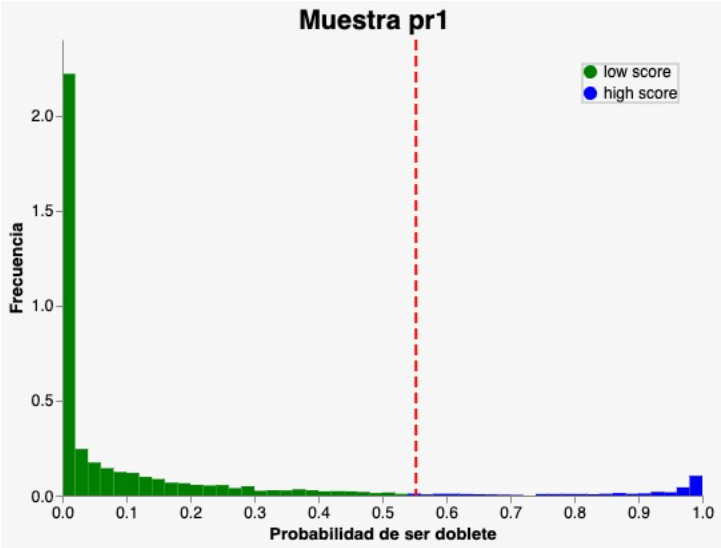

Figura 7. Umbral de probabilidad para pr1 es 0.55240

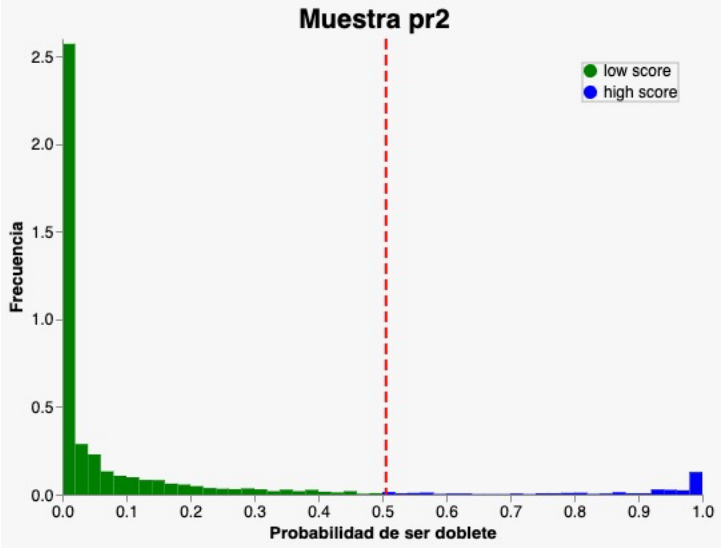

Figura 8. Umbral de probabilidad para pr2 es 0.50592

En el paso 6 **"integración de datos"** se aplica el método Seurat v4, se eliminan los efectos por lotes y se reducen la dimensionalidad de los datos a traves del método de Análisis de Componentes Principales (PCA). Este paso es muy importante porque se puede visualizar que ambas muestras están agrupadas de manera similar, es decir pr1 de color rojo y pr2 de color azul se parecen (Figura 9). Esta figura muestra como se ven los datos etiquetados y combinados según la tecnología de la secuenciación de la que provienen los datos. Las muestras pr1 y pr2 están bien integrados mostrando una buena distribución de cada una.

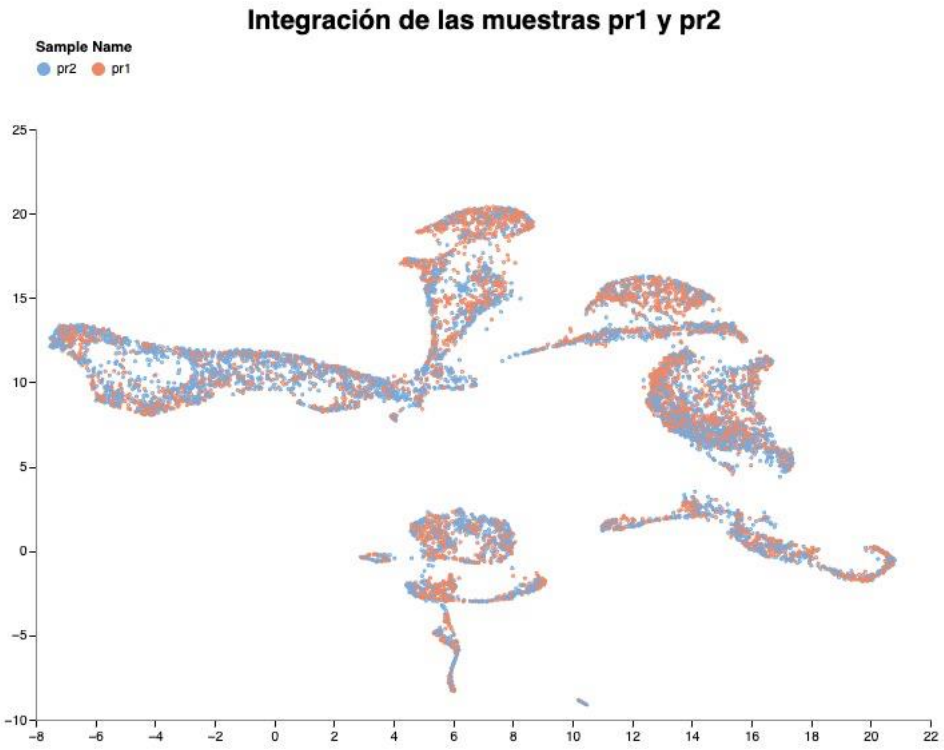

Figura 9. Integración de datos de las dos muestras del método de protoplastos

Finalmente, en el paso 7 **"configuración de la incrustación"** se usa el método estocástico llamado UMAP para la reducción de dimensionalidad siendo este método efectivo para visualizar los grupos de puntos de datos y sus proximidades relativas con una distancia mínima de 0.3, esta distancia puede estar dentro del rango de 0.001 a 1. En este caso el valor 0.3 controla con que fuerza se permite que la incrustación comprima los puntos, garantizando que los puntos incrustados se distribuyan de manera más uniforme. En este paso también se usa el método de agrupación Lovaina basado en el algoritmo de agrupación más popular y más eficiente en el análisis de datos de scRNA-seq con una resolución de 0.8, este parámetro permite obtener menos agrupaciones, en este caso se ha creado 21 agrupaciones (Figura 10). En resumen, utilizando las muestras que provienen de protoplastos muestran buena calidad y al recrear los análisis podemos separar las células en 22 distintos clusters, donde es importante resaltar que de cierta forma pareciera haber tipos celulares desconectados entre todas las células de la raíz. Lo cual podría asemejar los principales tipos celulares que contrastan.

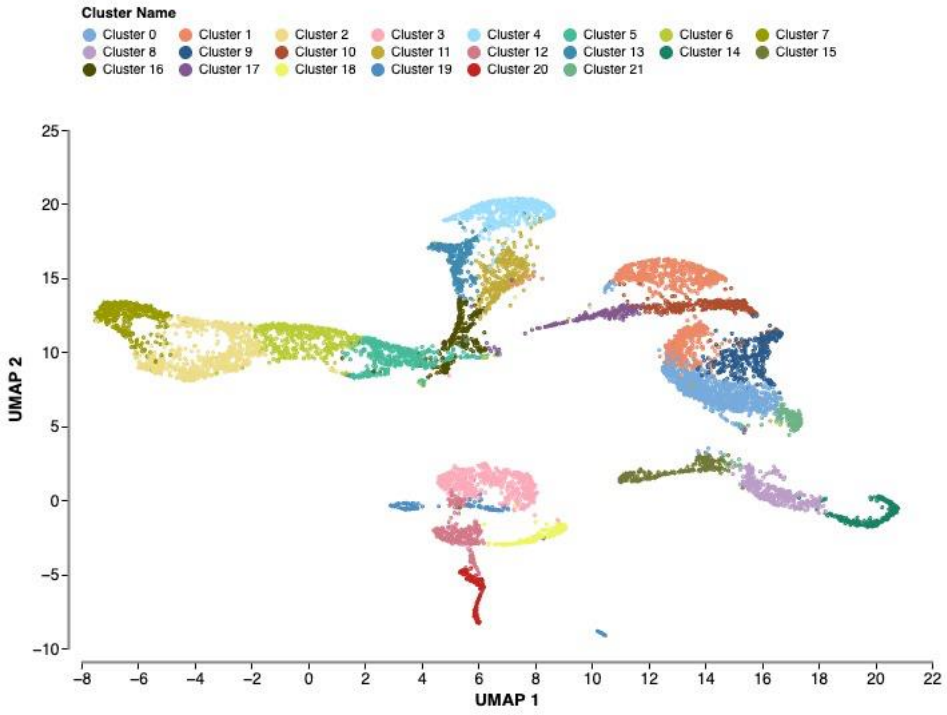

Figura 10. Configura la incrustación de las dos muestras del método de protoplastos

4.2. Procesamiento de datos en el Enfoque 2: Método de Aislamiento de Núcleos

Antes de empezar el procesamiento de datos del enfoque 2, los números estimados de células de *Arabidopsis thaliana* de las tres muestras del método de Aislamiento de Núcleos mr1, mr2 y mr3 cuentan con 1744 (Tabla 7), 2003 (Tabla 8) y 5849 (Tabla 9) respectivamente.

En el paso 1 **"filtro clasificador"** se mantienen los números estimados de células descritos en el párrafo anterior después del paso de filtrado, al igual que el número total de genes iguales a 23050/20672/24772 para mr1/mr2/mr3. Así como también, no hay cambios de proporción para los valores de la mediana de genes por célula y la mediana de recuentos UMI por celda, es decir se mantienen en 0.000% para el porcentaje de cambios (Tabla 7, Tabla 8, Tabla 9). Esto significa que todas las células se retienen niguna se filtra, es decir, no hay gotas vacías en el conjunto de datos, todas son válidas para el análisis.

Tabla 7. Estadística de filtrado para la muestra mr1 del método aislamiento de núcleos

Estadísticas	# antes	# después	% cambió
Número estimado de células	1744	1744	0.000
Número total de genes	23050	23050	0.000
Número medio de genes por célula	970	970	0.000
Recuentos medios de UMI por celda	1577	1577	0.000

Tabla 8. Estadística de filtrado para la muestra mr2 del método aislamiento de núcleos

Estadísticas	# antes	# después	% cambió
Número estimado de células	2003	2003	0.000
Número total de genes	20672	20672	0.000
Número medio de genes por célula	527	527	0.000
Recuentos medios de UMI por celda	768	768	0.000

Tabla 9. Estadística de filtrado para la muestra mr3 del método aislamiento de núcleos

Estadísticas	# antes	# después	% cambió
Número estimado de células	5849	5849	0.000
Número total de genes	24772	24772	0.000
Número medio de genes por célula	1707	1707	0.000
Recuentos medios de UMI por celda	2684	2684	0.000

Algunos pasos del procesamiento de datos no fueron tomados en cuenta por varias razones, en el paso 2 “**filtro de distribución del tamaño de celda**” muestran los mismos resultados pero de manera gráfica gráfica donde se detectan nuevamente gotas vacías y se afina el paso 1 y en el paso 3 “**filtro de contenido mitocondrial**” no se visualizó ningún gráfico porque no hay datos de secuencias mitocondriales.

Luego de aplicarse el paso 4 **"filtro de número de genes frente a UMI"** se visualiza un decremento del 1.72%, de 1744 se obtuvo 1714 células para mr1. Esto quiere decir que 30 células se eliminaron y no pasaron este filtro (Figura 11). También, existe un decremento del 0.15% para mr2, es decir, de 2003 se modifica a 2000 células, 3 células no pasaron el filtro (Figura 12). De igual manera, existe un decremento del 1.197% para mr3, es decir, de 5849 se modifica a 5779 células, 70 células no pasaron el filtro (Figura 13). En las Tablas 10, 11 y 12 se puede observar el porcentaje de mr1 para el número total de genes igual a 0.143%, para mr2 0.005% y para mr3 0.178% observándose un decremento de 33, 1 y 44 genes respectivamente. En cuanto, a los valores de la mediana de genes por célula se observa un incremento de 1.031%/0.095%/0.644% para cada muestra. Finalmente, la mediana de recuentos de UMI por celda ha habido un incremento del 0.335% para mr3, por otro lado, para mr1 y mr2 hay un decremento del 0.444% y 0.065 respectivamente.

Las Figuras 11, 12 y 13 muestran la correlación entre el número de genes y el número de moléculas UMI. La correlación es cercana a 1 lo que significa que los datos están bien representados.

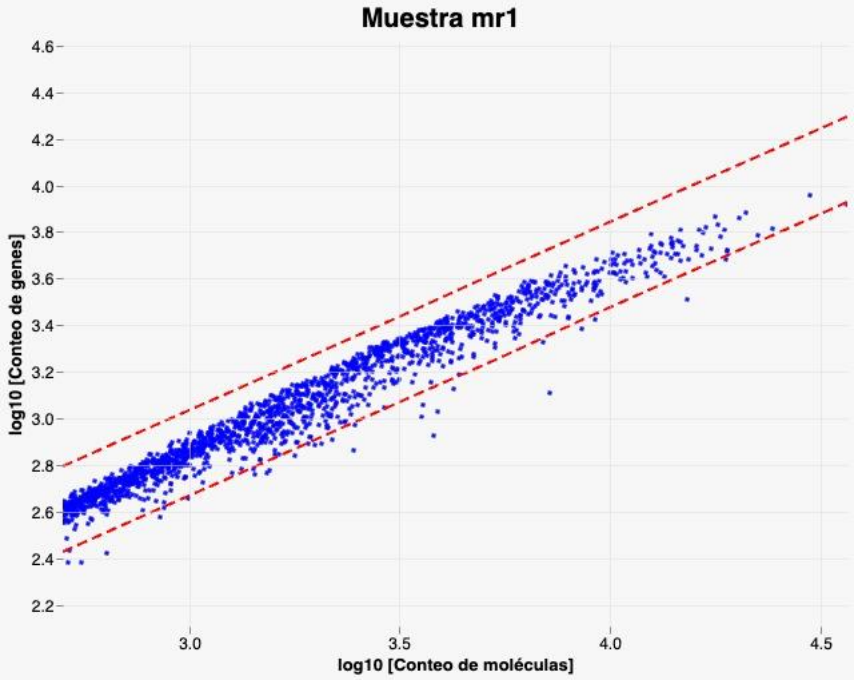

Figura 11. Fuera de la línea roja se muestran las 30 células que se eliminaron en la muestra mr1

Tabla 10. Valores de la primera muestra (mr1) del método aislamiento de núcleos después de aplicar el filtro de número de genes frente a UMI

Estadísticas	# antes	# después	% cambió
Número estimado de células	1744	1714	-1.720
Número total de genes	23050	23017	-0.143
Número medio de genes por célula	970	980	+1.031
Recuentos medios de UMI por celda	1577	1570	-0.444

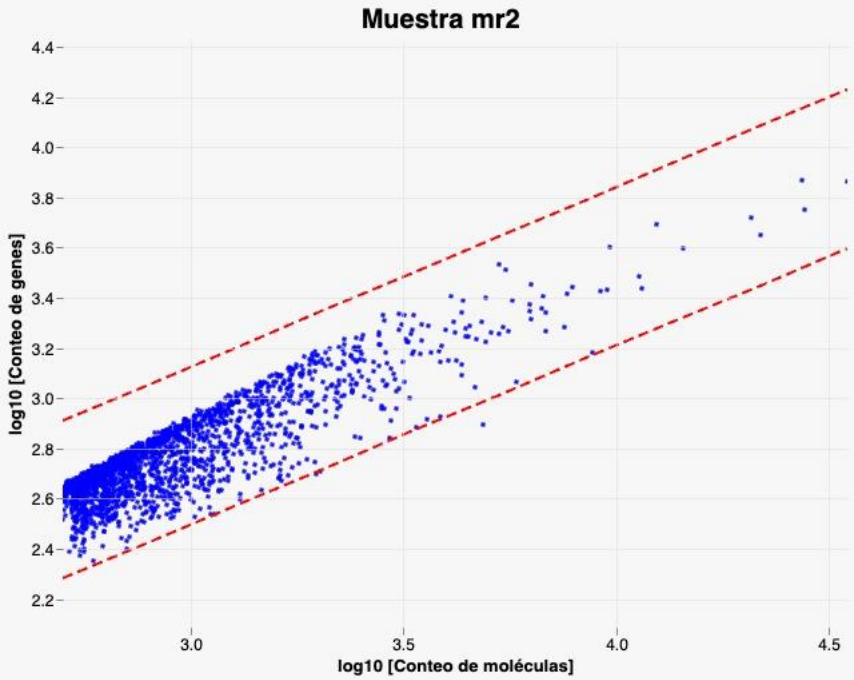

Figura 12. Fuera de la línea roja se muestran las 3 células que se eliminaron en la muestra mr2

Tabla 11. Valores de la primera muestra (mr2) del método aislamiento de núcleos después de aplicar el filtro de número de genes frente a UMI

Estadísticas	# antes	# después	% cambió
Número estimado de células	2003	2000	-0.150
Número total de genes	20672	20671	-0.005
Número medio de genes por célula	527	527.5	+0.095
Recuentos medios de UMI por celda	768	767.5	-0.065

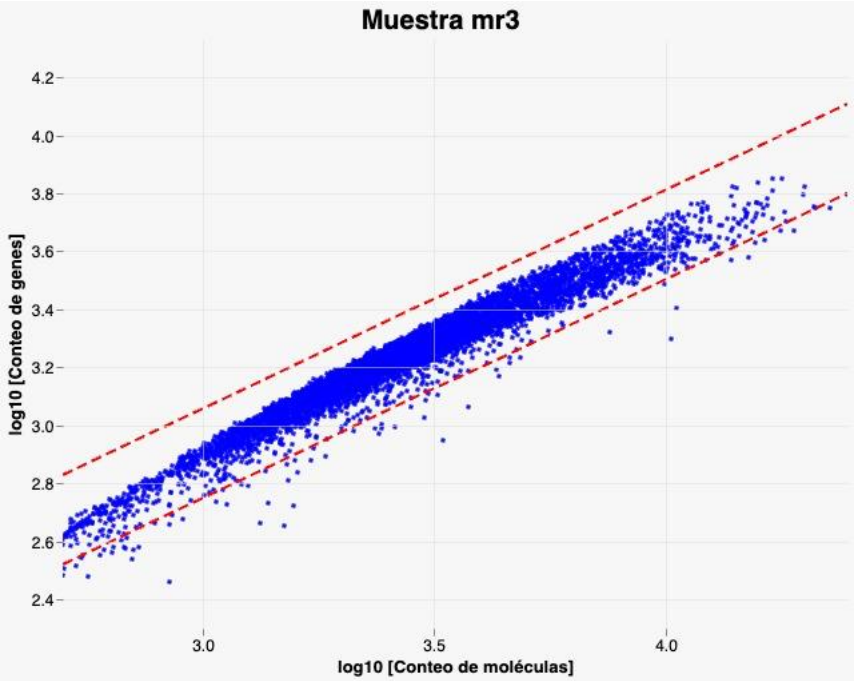

Figura 13. Fuera de la línea roja se muestran las 70 células que se eliminaron en la muestra mr3

Tabla 12. Valores de la primera muestra (mr3) del método aislamiento de núcleos después de aplicar el filtro de número de genes frente a UMI

Estadísticas	# antes	# después	% cambió
Número estimado de células	5849	5779	-1.197
Número total de genes	24772	24728	-0.178
Número medio de genes por célula	1707	1718	+0.644
Recuentos medios de UMI por celda	2684	2693	+0.335

Luego de aplicarse el paso 5 **"filtro doblete"** se visualiza un decremento del 6.184%, de 1714 se obtuvo 1608 células para mr1. Esto quiere decir que 106 células se eliminaron y que no pasaron este filtro. Existe un decremento del 5.400% para mr2, es decir, de 2000 se modifica a 1892 células, 108 células no pasaron el filtro. De igual manera, existe un decremento del 10.071% para mr3, es decir, de 5779 se modifica a 5197 células, 582 células no pasaron el filtro. Esto puede suceder, porque

a veces dos o más células pueden acabar en la misma gota causando problemas en el análisis posterior. En las Tablas 13, 14 y 15 se puede observar el porcentaje de mr1, mr2 y mr3 para el número total de genes igual a -1.269%, -1.079% y -1.278% y con una pérdida de genes de 292, 223 y 316 respectivamente. Se observa que ha habido decrementos en los valores de las tres muestras mr1, mr2 y mr3 iguales a 6.276%, 1.801% y 5.763% para la mediana de genes por célula y 6.242%, 3.127% y 7.204% para la mediana de recuentos de UMI por celda.

La Figura 14, Figura 15 y Figura 16 no son semejantes porque el umbral de probabilidad para cada muestra es diferente, así tenemos los siguientes resultados: mr1 = 0.86447, para mr2 = 0.68752 y para mr3 = 0.55810.

Tabla 13. Valores de la primera muestra (mr1) del método aislamiento de núcleos después de aplicar el filtro doblete

Estadísticas	# antes	# después	% cambió
Número estimado de células	1714	1608	-6.184
Número total de genes	23017	22725	-1.269
Número medio de genes por célula	980	918.5	-6.276
Recuentos medios de UMI por celda	1570	1472	-6.242

Tabla 14. Valores de la primera muestra (mr2) del método aislamiento de núcleos después de aplicar el filtro doblete

Estadísticas	# antes	# después	% cambió
Número estimado de células	2000	1892	-5.400
Número total de genes	20671	20448	-1.079
Número medio de genes por célula	527.5	518	-1.801
Recuentos medios de UMI por celda	767.5	743.5	-3.127

Tabla 15. Valores de la primera muestra (mr3) del método aislamiento de núcleos después de aplicar el filtro doblete

Estadísticas	# antes	# después	% cambió
Número estimado de células	5779	5197	-10.071
Número total de genes	24728	24412	-1.278
Número medio de genes por célula	1718	1619	-5.763
Recuentos medios de UMI por celda	2693	2499	-7.204

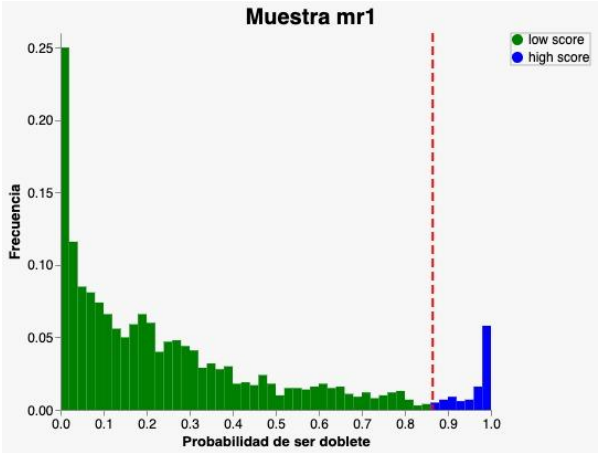

Figura 14. Umbral de probabilidad para mr1 es 0.86447

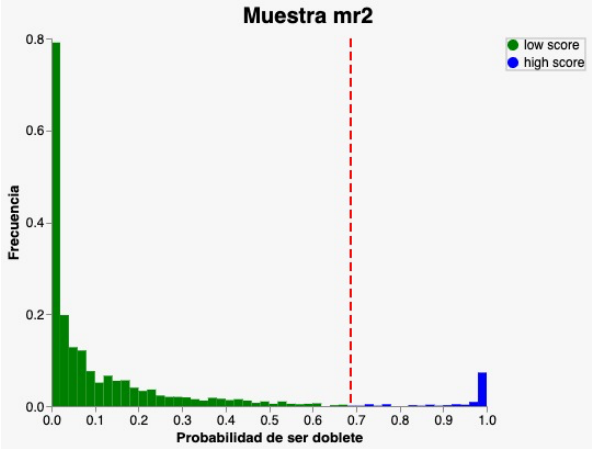

Figura 15. Umbral de probabilidad para mr2 es 0. 68752

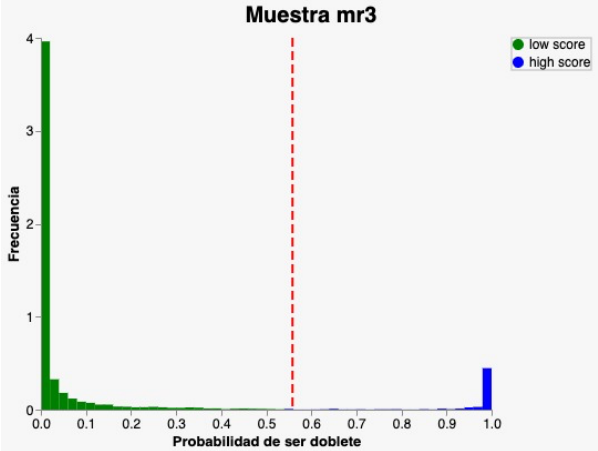

Figura 16. Umbral de probabilidad para mr2 es 0.55810

El paso 6 **“integración de datos”** se aplica la misma configuración que el enfoque 1. Este paso es muy importante porque se puede visualizar que las tres muestras están agrupadas, mr1 de color azul, mr2 de color amarillo y mr3 de color rojo (Figura 17). Esta integración de datos no es totalmente similar, se puede obsevar mayor concentración de células de la muestra mr2 en una sola área. Esto se debe a que el número total de genes de las muestras mr1 y mr3 se encuentran dentro del rango en función a la desviación estandar, mientras que la muestra mr2 está fuera de este rango.

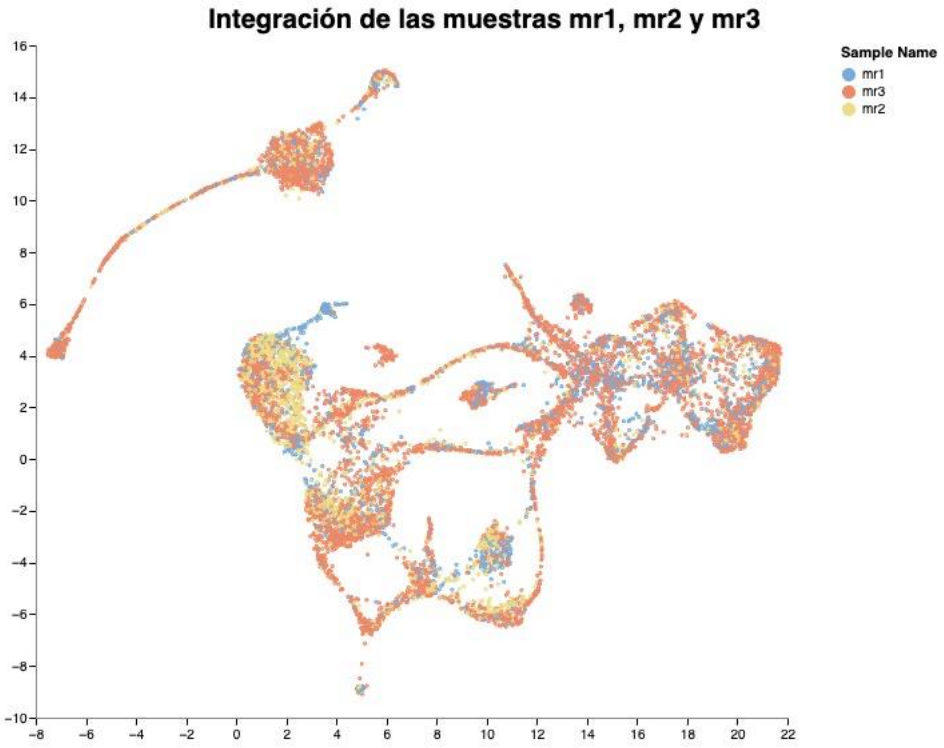

Figura 17. Integración de datos de las tres muestras del método de aislamiento de núcleos

Finalmente, en el paso 7 **“configuración de la incrustación”** al igual que en el método protoplastos se usa el método estocástico UMAP con una distancia mínima de 0.3 que garantice que los puntos incrustados se distribuyan de manera más uniforme. Se usa el método de agrupación Lovaina con una resolución de 0.8 permitiéndo obtener menos agrupaciones, en este caso se han creado 22 clusters (Figura 18). Basados en los perfiles transcriptómicos, las células se agrupan en 22 diferentes clusters (desde 0 a 21). En resumen, los datos provenientes del método de aislamiento de núcleos proveen datos de buena calidad de forma similar al

método de protoplastos. Interesantemente, ambos conjuntos de datos resolvieron el mismo número de tipos celulares (22 clusters), lo cual confirma que ambos métodos pueden ser utilizados para identificar la diversidad de células de la raíz. Sin embargo, es importante resaltar que a diferencia del agrupamiento de protoplastos, el agrupamiento de aislamiento de núcleos muestra una mayor conexión entre los clusters, lo que sugiere una mayor resolución de estados celulares que dinámicamente conectan entre tipos celulares.

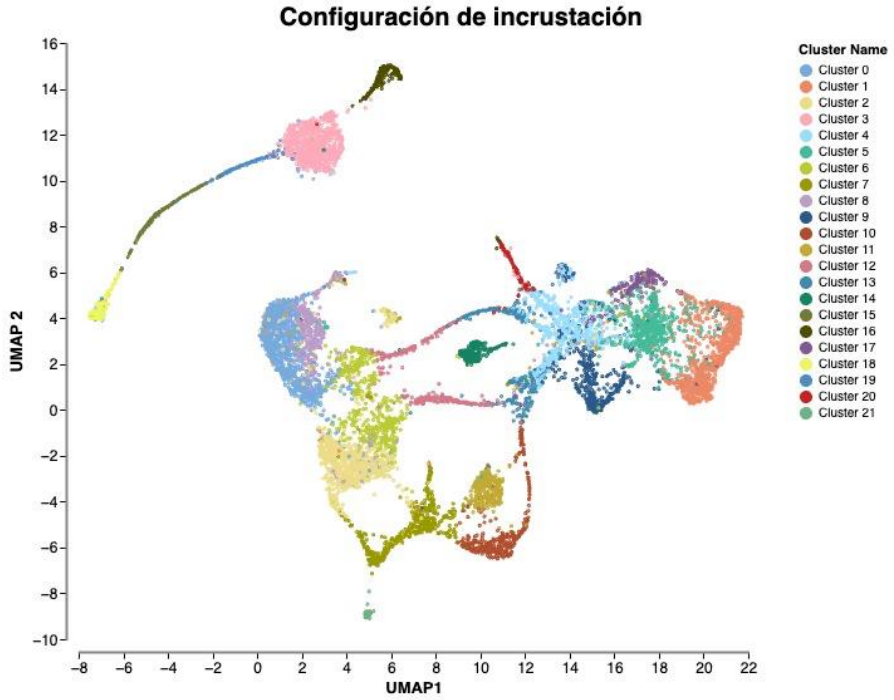

Figura 18. Configura la incrustación de las tres muestras del método de Aislamiento de núcleos

4.3. Procesamiento de datos en el Enfoque 3: Combinación de Métodos (Protoplastos + Aislamiento de Núcleos)

En este enfoque se combinó las cinco muestras de *Arabidopsis thaliana*, dos que corresponde al método de protoplastos y tres al método de aislamiento de núcleos. Los cinco primeros pasos desde el filtro clasificador hasta el filtro doblete se registran los mismos valores expuestos en los dos enfoques anteriores.

El paso 6 **"integración de datos"** es muy importante porque se puede visualizar que las muestras están agrupadas de manera similar (Figura 19). En esta integración se utilizó el mismo método Seurat V4 que en los enfoques anteriores.

Existe una reducción de dimensionalidades comprimiendo los datos para permitir la visualización en dos dimensiones, el método utilizado para este caso fue el Análisis de Componentes Principales (PCA). Se ha usado el método de normalización predeterminado llamado LogNormalize que permite una visualización y comparación de las diferencias de expresión entre células. En la Figura 19 se observa la incrustación coloreada por muestra generada después de la reducción dimensional. Se visualizan las muestras mr2, mr1, pr2 y pr1 muy similares, esto significa que tanto las muestras de protoplastos como las muestras de aislamiento de núcleos son muy parecidas a pesar de las diferencias entre los métodos, sin embargo, hay pequeña diferencia con la muestra mr3. La Figura 20 muestra un gráfico de frecuencia que visualiza la contribución de cada muestra en cada grupo. La Figura 21 es un gráfico de codo que mapea la contribución porcentual de cada Componente Principal (PC) a la variación total en el conjunto de datos. En este caso, se ha definido la configuración predeterminada para la cantidad de PC igual a 30 y un porcentaje de variación explicada del 89.34%.

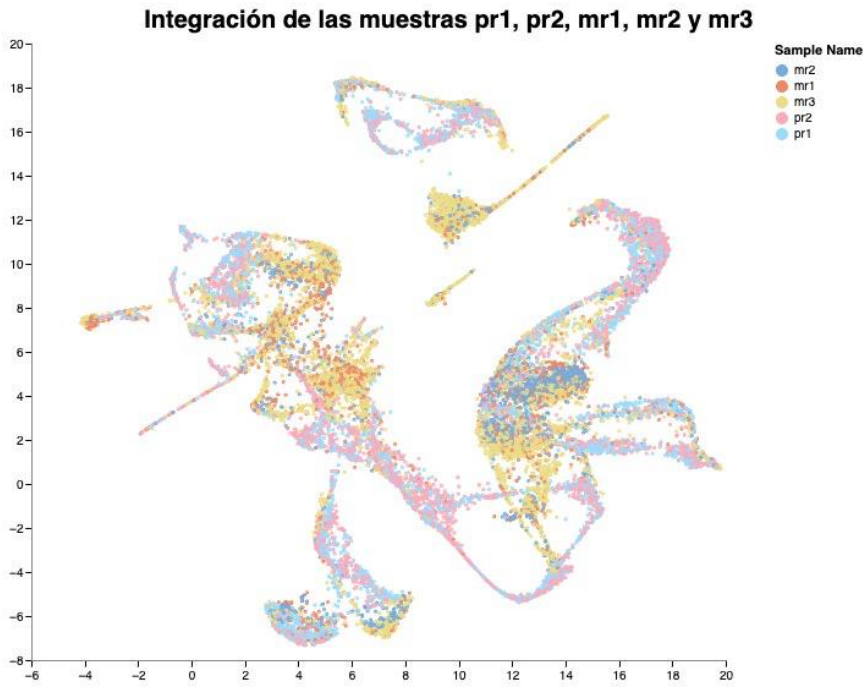

Figura 19. Integración de datos de las cinco muestras combinando los dos métodos

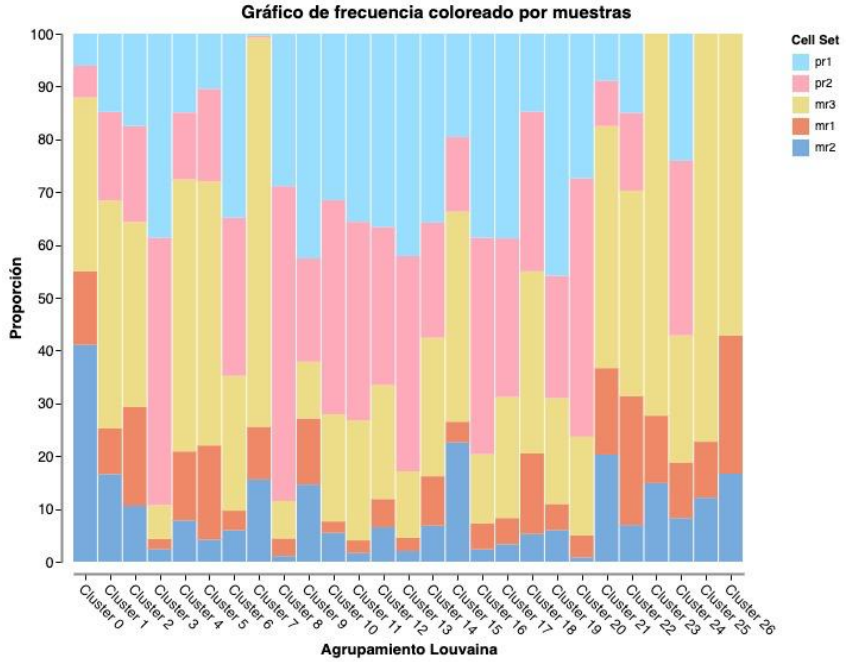

Figura 20. Gráfico de frecuencia coloreado por muestras combinando los dos métodos

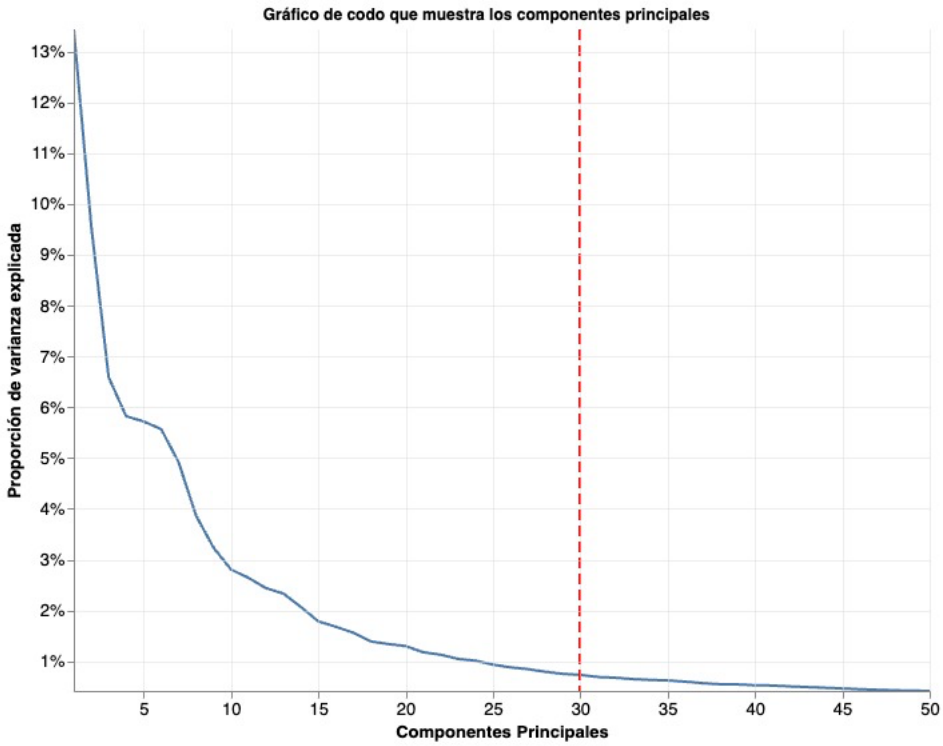

Figura 21. Gráfico de codo que muestra los componentes principales

Finalmente, en el paso 7 **“configuración de la incrustación”** se usa el método predeterminado UMAP que permite visualizar los datos. UMAP es una técnica con un algoritmo que se ajusta más fácilmente a la paralelización y funciona más rápido que el método tSNE (Figura 22). En este gráfico se pueden visualizar los puntos de datos incrustados agrupados y coloreados de acuerdo con las anotaciones del grupo. Se agrupan las células de alta similitud usando el método de agrupamiento de Louvain utilizado por Cellenics por defecto. Los grupos están codificados por colores y numerados desde el clúster 0 hasta el clúster 26 para identificarlos y ser explorados más adelante.

La combinación de datos de protoplastos y aislamiento de núcleos muestran que se complementan bastante bien para mostrar una mejor resolución de la definición de los tipos celulares y por ende, de los transcriptomas.

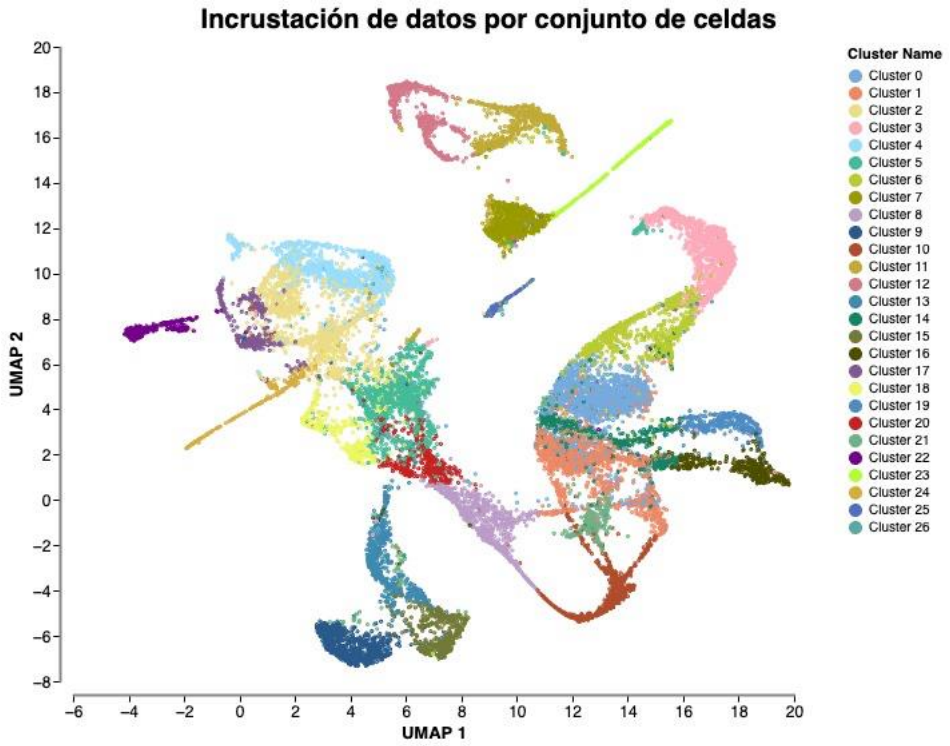

Figura 22. Incrustación de todas las muestras con trama de color por conjunto de celdas

En la figura 23 se muestra una incrustación de UMAP coloreada por la puntuación de probabilidad del doblete. Se visualiza una distribución uniforme de color, no se muestra una coloración alta, lo que indica que hay escaces de población de dobletes presentes

en los datos. En la figura 24 muestra una incrustación de UMAP coloreada según el número de genes. Finalmente la figura 25 muestra una incrustación de UMAP coloreada según la cantidad de UMI.

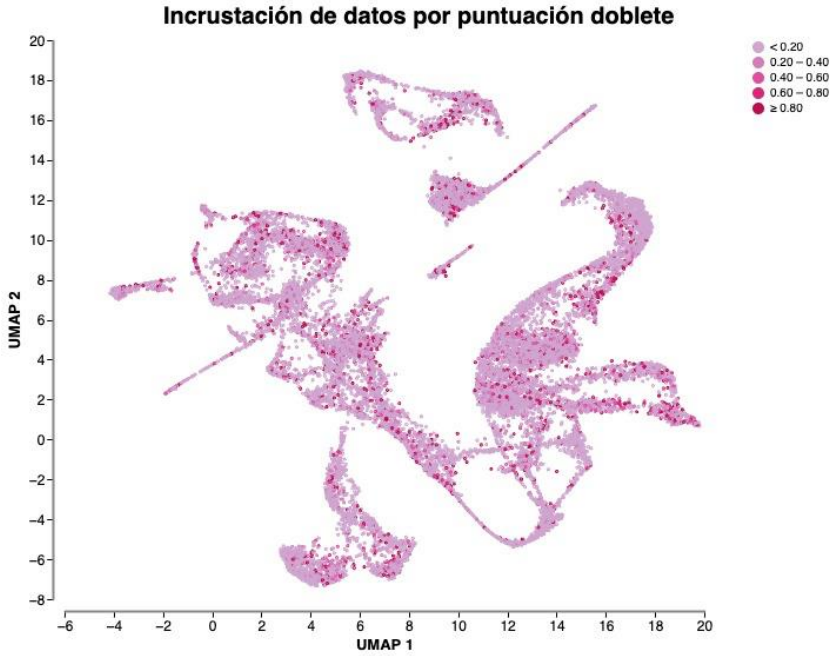

Figura 23. Configura la incrustación de todas las muestras con trama de color por puntuación de doblete

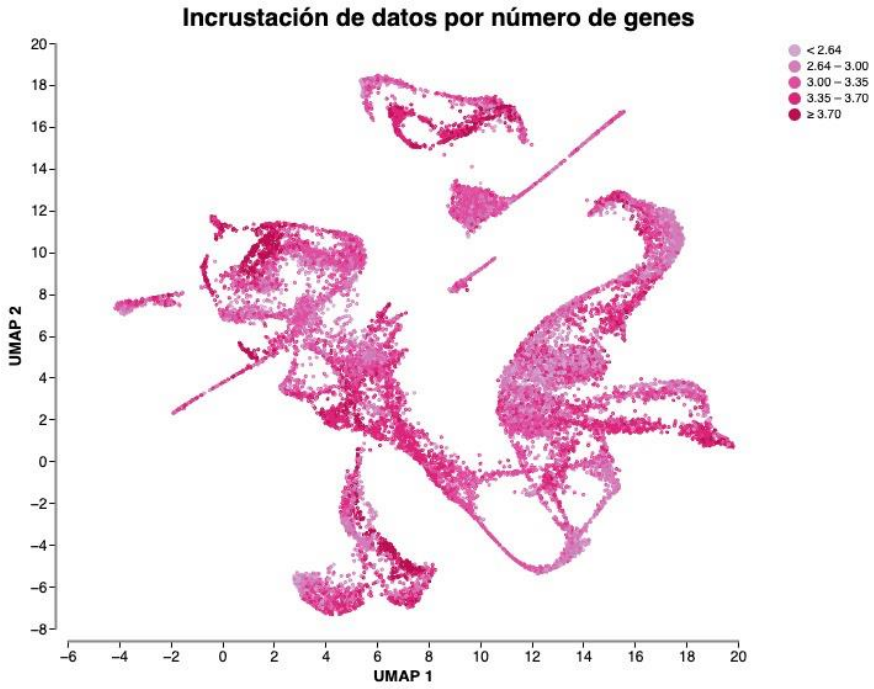

Figura 24. Configura la incrustación de todas las muestras con trama de color por número de genes

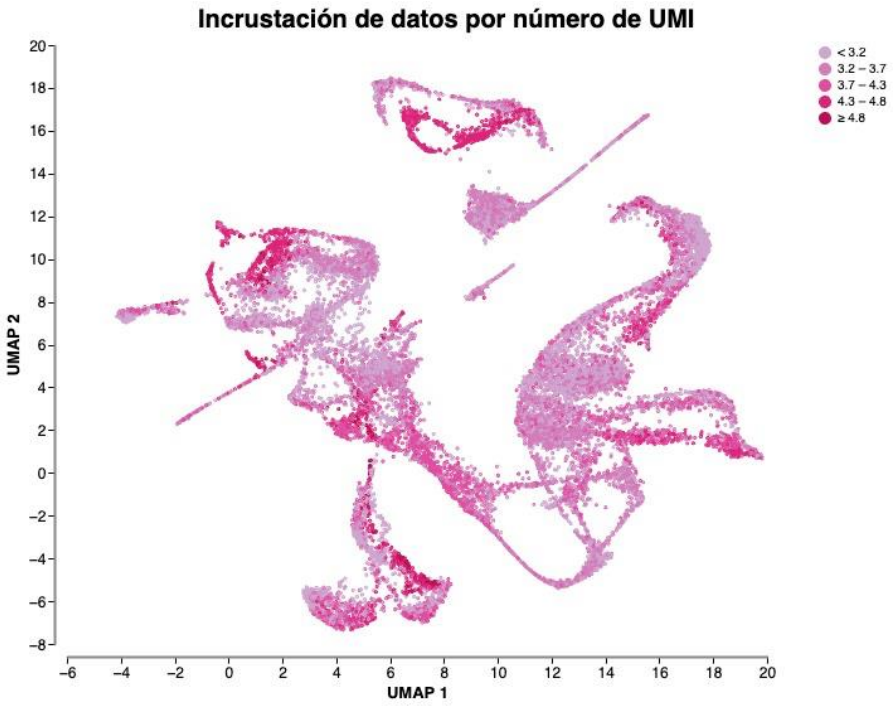

Figura 25. Configura la incrustación de todas las muestras con trama de color por número de UMI

4.4. Exploración de datos en la plataforma Cellenics

Se ha realizado una exploración profunda de datos usando una amplia variedad de funciones, permitiendo identificar los tipos de células que están representados por sus conjuntos de células, personalizando la selección del conjunto de células y generando información sobre el conjunto de datos mediante la visualización de la expresión génica y la expresión diferencial.

4.4.1. Exploración de datos del Enfoque 1: Método de Protoplastos

En la Figura 26 se muestra varias secciones, a la izquierda se encuentra la incrustación de UMAP (Agrupamientos de Lovaina), en el medio, se encuentra la lista de agrupamientos de louvain que está marcado por defecto, también muestra el conjunto de celdas personalizadas, muestras y métodos. A la derecha, se visualiza la lista de genes presentes ordenados desde los genes con alta dispersión los mismos que tienen un alto nivel de variación entre las celdas del conjunto de datos, siendo los genes XCP1, PER66, AT3G27200, ADF9 los más variados del conjunto de datos. El gen XCP1 tiene el valor máximo de dispersión igual a 20.12,

mientras que el valor mínimo mayor a 0 tiene el gen AT1G65720 igual a 0.322. En la parte inferior se encuentra el mapa de calor que muestra genes marcadores para los grupos de Lovaina de forma predeterminada.

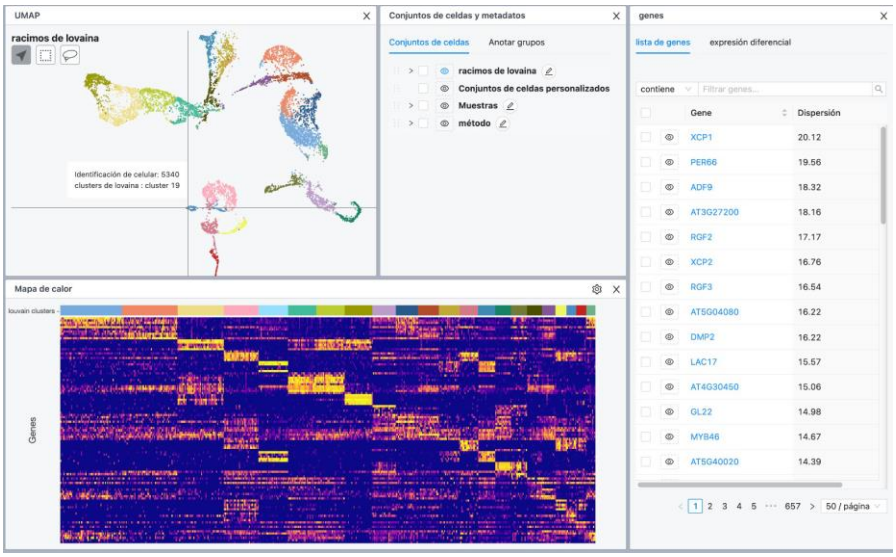

Figura 26. Exploración de datos del Enfoque 1 - Protoplastos

4.4.2. Exploración de datos del Enfoque 2: Método de Aislamiento de Núcleos

En la Figura 27 se muestran las mismas secciones del método de protoplastos, resaltando la lista de genes presentes a la derecha ordenados por dispersión, siendo los genes FAR3, AT1G03920, AT1G43020, DTX3 los más variados del conjunto de datos y diferentes a los genes reportados en la Figura 26.

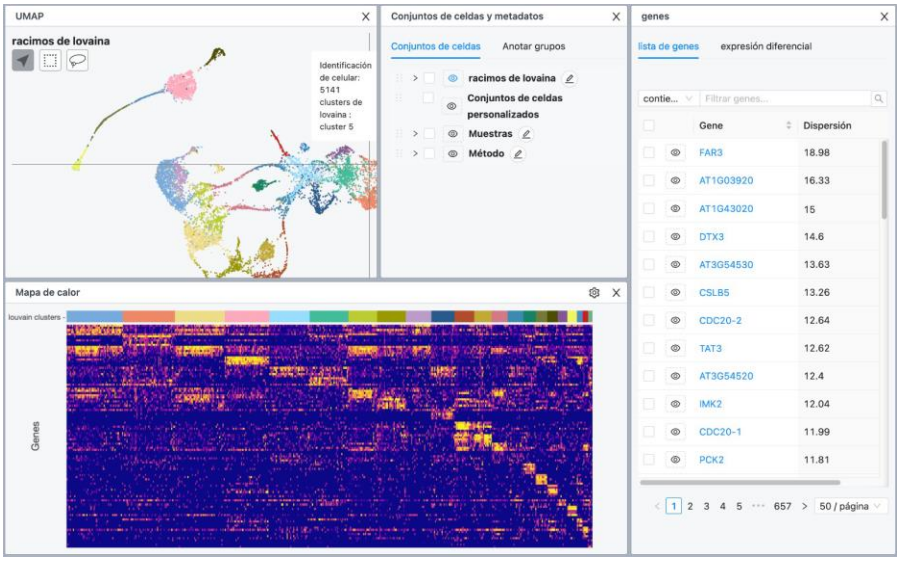

Figura 27. Exploración de datos del Enfoque 2 – Aislamiento de Núcleos

4.4.3. Exploración de datos del Enfoque 3: Combinación de Métodos (Protoplastos + Aislamiento de Núcleos)

En la Figura 28 se muestran las mismas secciones de los métodos anteriores de protoplastos y aislamiento de núcleos, resaltando la lista de genes presentes a la derecha ordenados por dispersión, siendo los genes PER66, XCP1, AT3G27200, ADF9 los más variados del conjunto de datos y coinciden con los cuatro genes con alta dispersión del método de protoplastos. Por otro lado, los cuatro genes FAR3, AT1G03920, AT1G43020 Y DTX3 correspondientes al método de aislamiento de núcleo tienen una dispesión mayor a 0. En la Tabla 16 se muestran los valores de dispersión de los cuatro genes por cada enfoque.

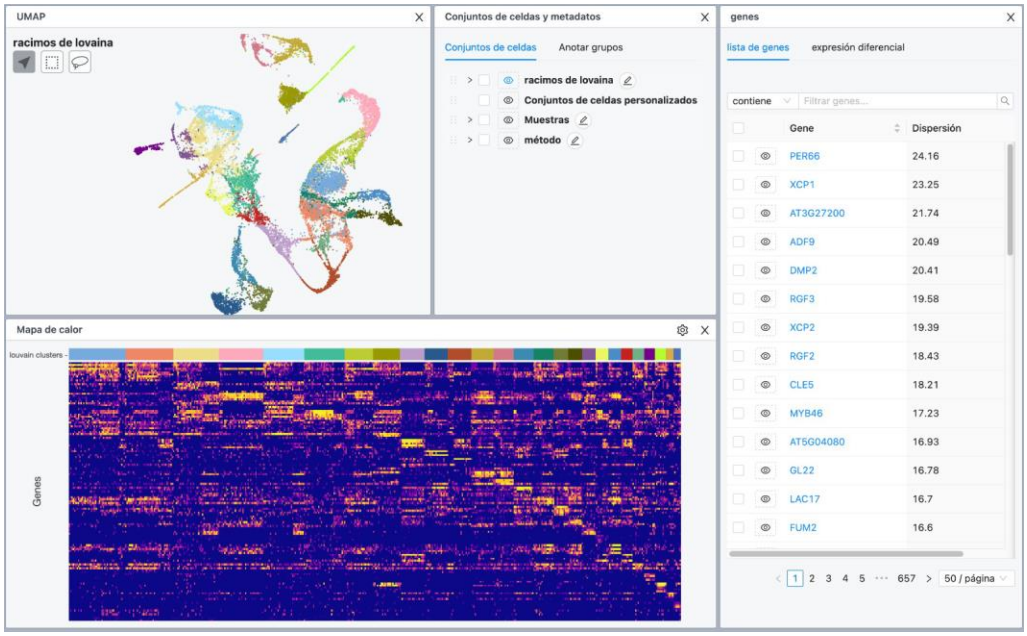

Figura 28. Exploración de datos del Enfoque 3 – Combinación de métodos (Protoplastos y Aislamiento de núcleos)

Tabla 16. Valores altos de dispersión para los métodos de protoplastos y Aislamiento de núcleos

Dispersión	Protoplastos	Aislam. núcleos	Protoplastos + Aislam. Nucleos
XCP1	20.12	9.817	23.25
PER66	19.56	6.099	24.16
ADF9	18.32	1.957	20.49
AT3G27200	18.16	6.057	21.70
FAR3	2.783	18.98	14.52
AT1G03920	4.658	16.33	8.658
AT1G43020	3.594	15	10.68
DTX3	1.002	14.6	14.16

5. Análisis de resultados

Cuando se descargaron las muestras de secuenciación de ARN de la especie *Arabidopsis thaliana*, registraban valores iniciales de pr1, pr2, mr1, mr2 y mr3 correspondientes a las muestras de los métodos Protoplastos y Aislamiento de Núcleos. Durante el procesamiento de datos, estos datos se filtraron por muestra aplicándose varios filtros como el clasificador, distribución del tamaño de celda, contenido mitocondrial, número de genes frente a UMI y doblete registrándose nuevos valores en las Tablas 17 y 18 como valores finales para pr1, pr2, mr1, mr2 y mr3, lo que indica que se eliminaron algunos datos no deseados y de mala calidad de cada muestra individual.

Se procedió a calcular la media y desviación estándar de las muestras para cada método, además se calcularon los límites máximos y mínimos. En la Tabla 17 los valores iniciales y finales de las muestras pr1 y pr2 están dentro del rango para el método de Protoplastos, de igual manera en la Tabla 18 los valores de la muestra mr1 están dentro del rango para el método de Aislamiento de Núcleos. Por otro lado en la Tabla 18 no todos los valores iniciales y finales de las muestras mr2 y mr3 permanecen en el rango límite mínimo y límite máximo para el método de Aislamiento de Núcleos.

Tabla 17. Valores de pr1 y pr2 iniciales y finales en el método protoplastos

	Método Protoplastos											
	Valores iniciales						Valores finales					
	pr1	pr2	Media	Desv. Estand.	Límite mínimo	Límite máximo	pr1	pr2	Media	Desv. Estand.	Límite mínimo	Límite máximo
Número estimado de células	4196	4437	4316,5	170,4	4146,1	4486,9	3871	4084	3977,5	150,6	3826,9	4128,1
Número total de genes	23815	23750	23782,5	46,0	23736,5	23828,5	23597	23577	23587,0	14,1	23572,9	23601,1
Número medio de genes por célula	1466,5	1650	1558,3	129,8	1428,5	1688,0	1339	1526	1432,5	132,2	1300,3	1564,7
Recuentos medios de UMI por celda	3908,5	4757	4332,8	600,0	3732,8	4932,7	3373	4320	3846,5	669,6	3176,9	4516,1

Tabla 18. Valores de mr1, mr2 y mr3 iniciales y finales en el método aislamiento de núcleos

	Método Aislamiento de núcleos													
	Valores iniciales							Valores finales						
	mr1	mr2	mr3	Media	Desv. Estand.	Límite mínimo	Límite máximo	mr1	mr2	mr3	Media	Desv. Estand.	Límite mínimo	Límite máximo
Número estimado de células	1744	2003	5849	3198,7	2298,9	899,8	5497,6	1608	1892	5197	2899,0	1995,2	903,8	4894,2
Número total de genes	23050	20672	24772	22831,3	2058,7	20772,6	24890,1	22725	20448	24412	22528,3	1989,3	20539,0	24517,6
Número medio de genes por célula	970	527	1707	1068,0	596,1	471,9	1664,1	918,5	518	1619	1018,5	557,3	461,2	1575,8
Recuentos medios de UMI por celda	1577	768	2684	1676,3	961,9	714,5	2638,2	1472	743,5	2499	1571,5	882,0	689,5	2453,5

Realizando el análisis de los resultados usando el gráfico de frecuencias, en las Figuras 30 y 31 muestran las proporciones de células de cada grupo en cada muestra. El eje “x” representa las muestras por las cuales se agrupan las celdas. El eje “y” representa las proporciones que son los valores de frecuencia. En la Figura 29 se observa cambios significativos en algunas y no en todas las proporciones de celdas entre las muestras pr1 y pr2 del método protoplastos. En este caso, se observa que el cluster 3 en ambas muestras son iguales. Por otro lado, en la Figura 30 se observa cambios significativos en todas las proporciones de celdas entre las muestras mr1, mr2 y mr3 del método de aislamiento de núcleos.

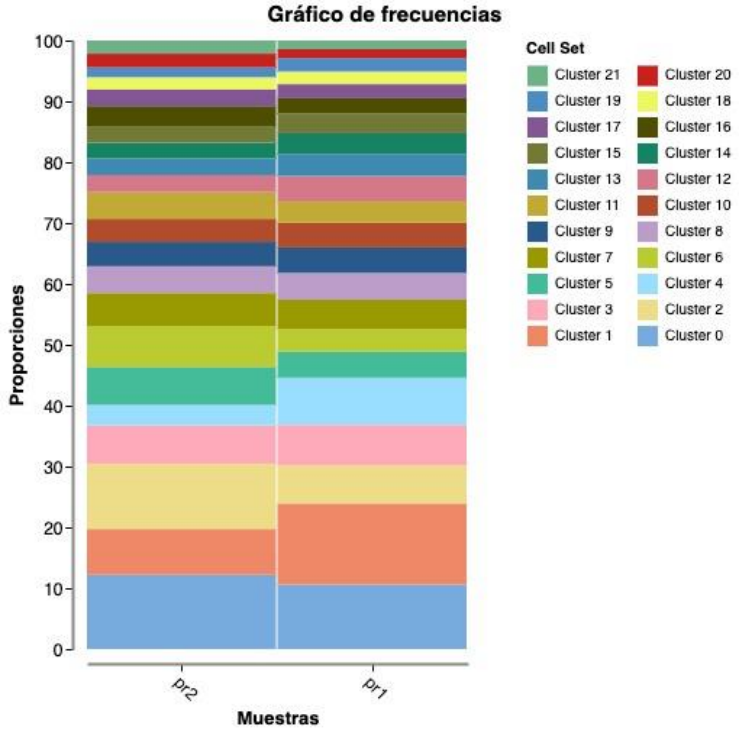

Figura 29. Gráfico de frecuencias de las réplicas pr1 y pr2 del método protoplastos

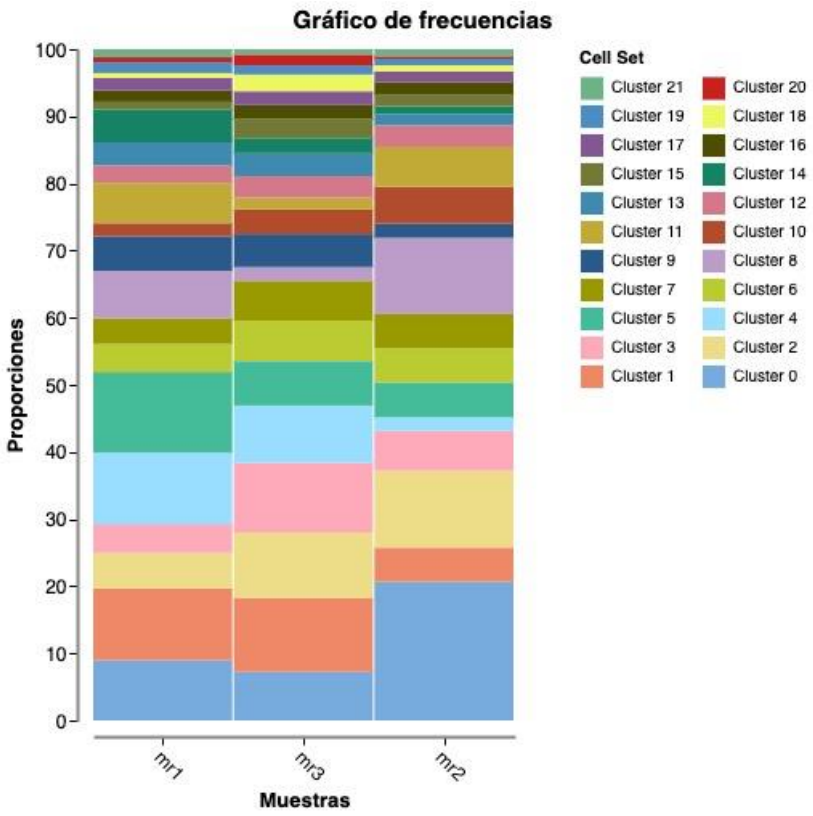

Figura 30. Gráfico de frecuencias de las réplicas mr1, mr2 y mr3 del método aislamiento de núcleos

Finalmente, analizando la Figura 31 muestra las proporciones de células de cada grupo en cada muestra pr1, pr2, mr1, mr2 y mr3 donde se observa cambios significativos todas las proporciones de celdas.

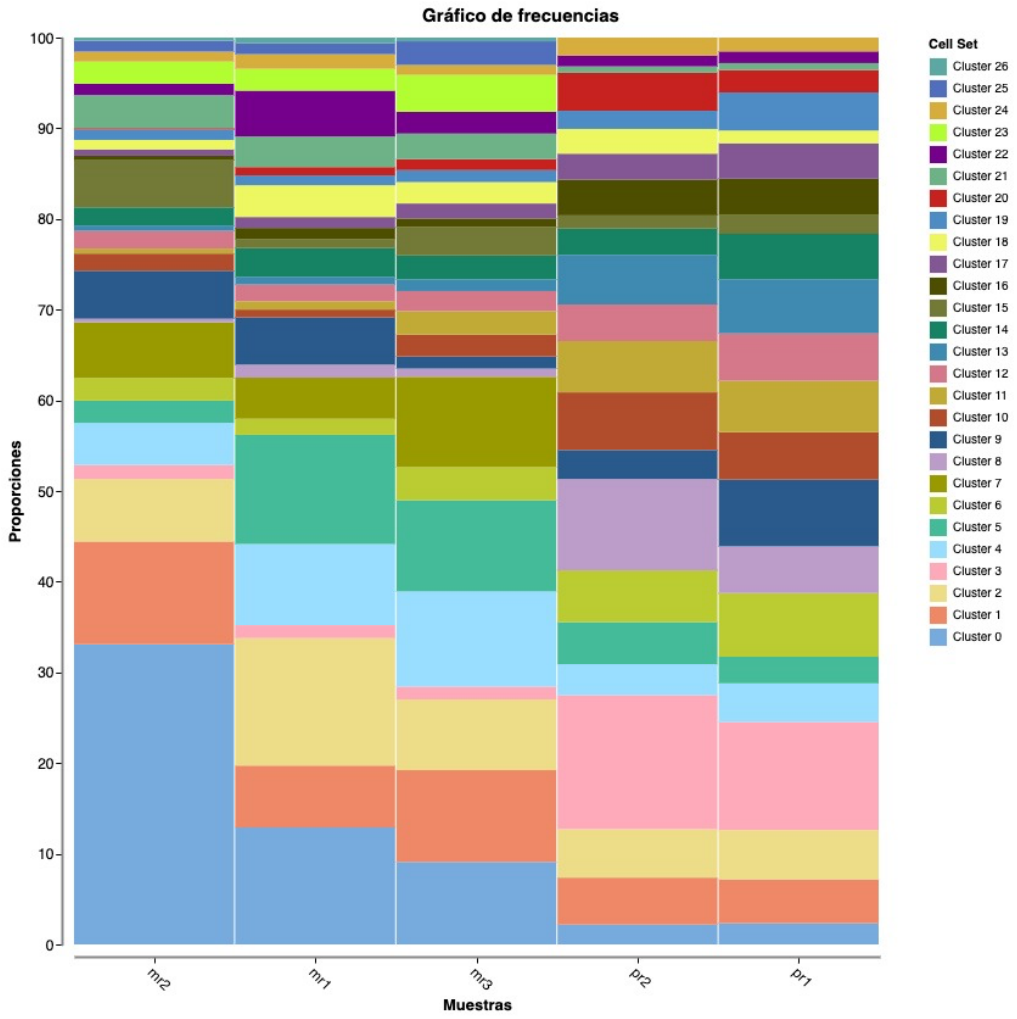

Figura 31. Gráfico de frecuencias de las muestras pr1, pr2, mr1, mr2 y mr3

El gráfico de violín es otra de las herramientas utilizadas en este trabajo que nos permite analizar la expresión génica. En la Figura 32 se observa la distribución de las expresiones normalizadas de los 4 genes del método protoplastos, XCP1, PER66, AT3G27200 y ADF9 con la mayor dispersión igual a 20.12, 19.56, 18.32 y 18,16 respectivamente como se indica en la Figura 26. En la Figura 33 se observa la distribución de las expresiones normalizadas de los 4 genes con mayor dispersión en el método de aislamiento de núcleos como indica la Figura 27, los genes son FAR3, AT1G03920, AT1G43020 Y DTX3 con 18.98, 16.33, 15 Y 14.6 respectivamente. En ambas figuras, los puntos negros representan células. Se visualizan líneas horizontales negras en la parte inferior de los granos. Estos son puntos que indican las células donde el gen no se expresa.

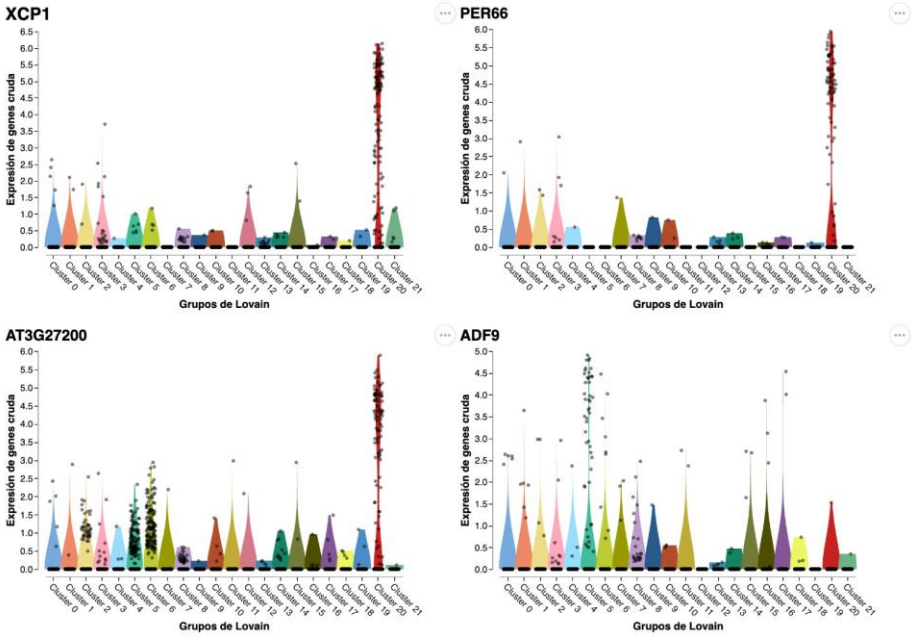

Figura 32. Expresión génica usando la técnica de Lovaina en los genes XCP1, PER66, AT3G27200 y ADF9 en protoplastos

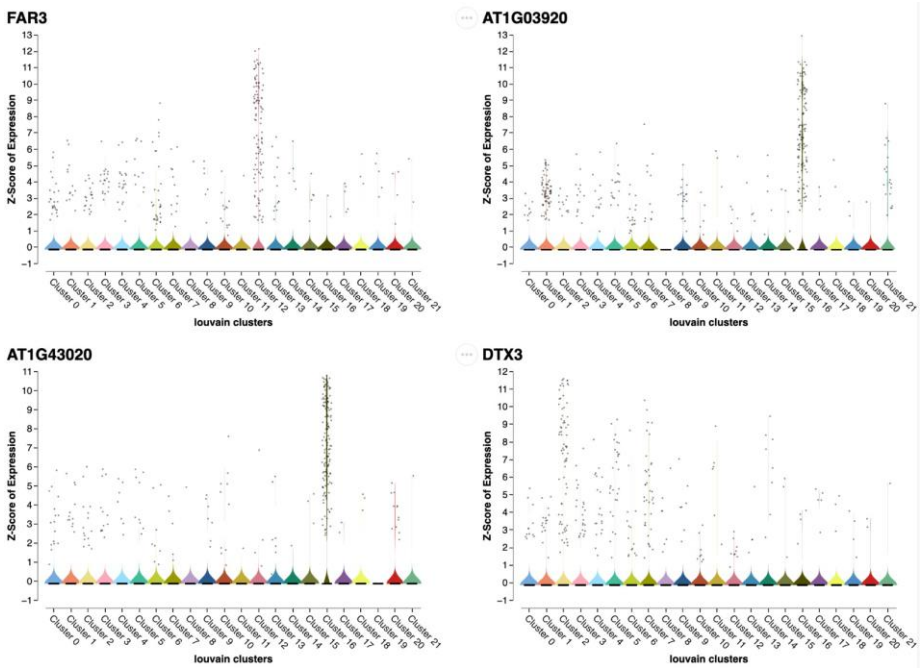

Figura 33. Expresión génica usando la técnica de Lovaina en los genes FAR3, AT1G03920, AT1G43020 y DTX en aislamiento de núcleos

Con respecto a los genes con mayor dispersión en el método protoplastos que fueron XCP1, PER66, AT3G27200 y ADF9 a continuación en la Figura 34 se muestra la expresión génica de estos cuatro genes en el método de aislamiento de núcleos.

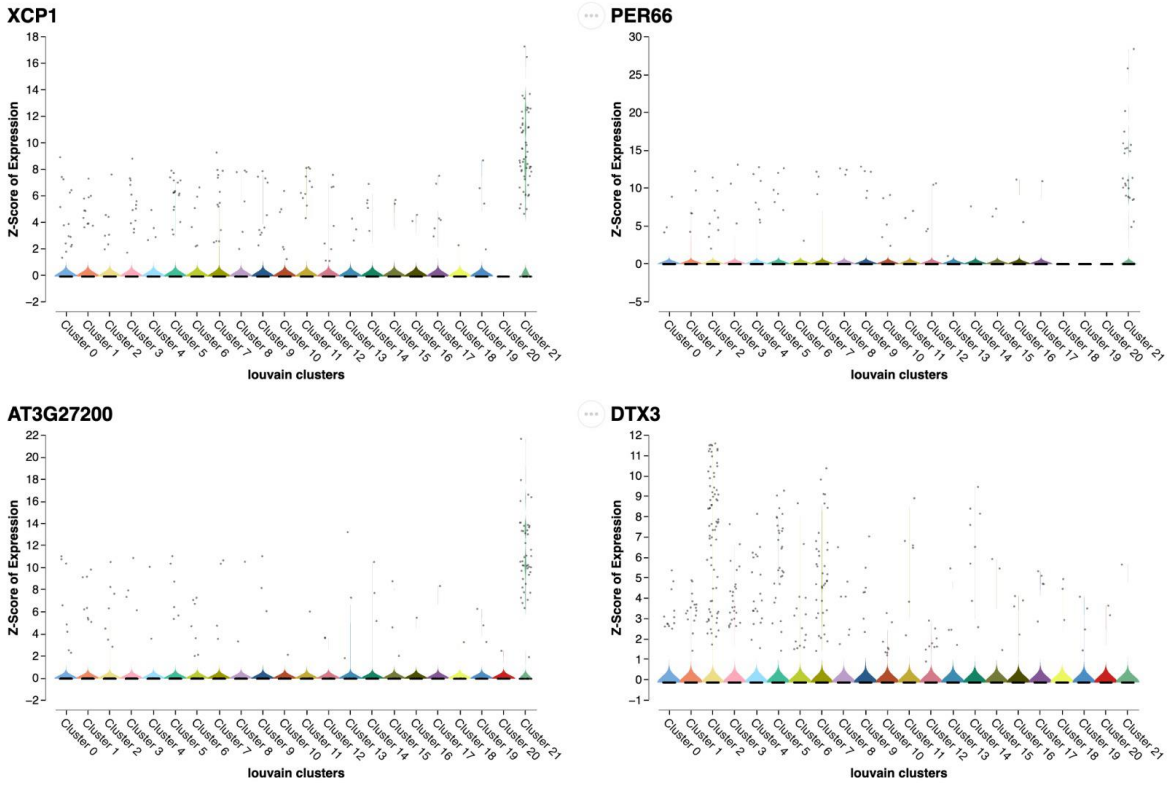

Figura 34. Expresión génica usando la técnica de Lovaina de Genes XCP1, PER66, AT3G27200 y DTX3 en aislamiento de núcleos

De igual manera, los genes con mayor dispersión en el método de aislamiento de núcleos que fueron FAR3, AT1G03920, AT1G43020 Y DTX3, en la Figura 35 se muestra la expresión génica de estos cuatro genes en el método de protoplastos.

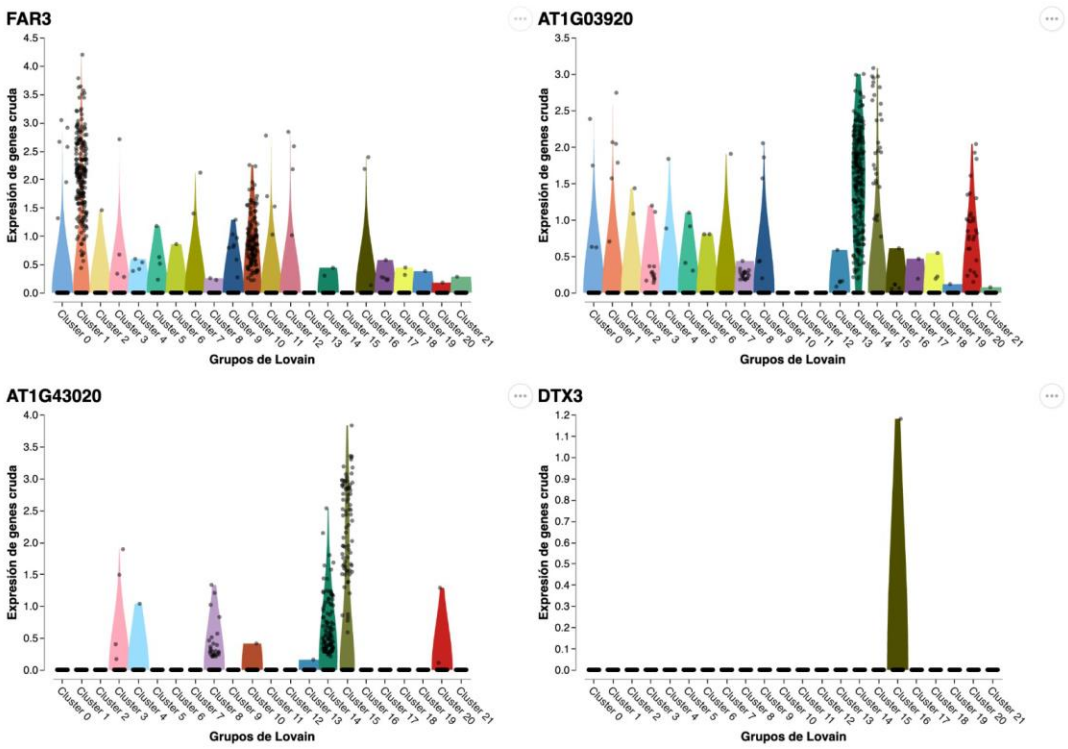

Figura 35. Expresión génica usando la técnica de Lovaina en los genes FAR3, AT1G03920, AT1G43020 y DTX3 en protoplastos

En la Figura 36 se visualiza el diagrama de volcán para representar la expresión diferencial de genes, en este caso se compara el conjunto de datos del cluster 0 con el resto de clusters louvaina dentro de los grupos de datos pr1 y pr2. En la Figura 37 se compara el cluster 0 con el resto de clusters louvaina de los grupos de datos mr1, mr2 y mr3. En ambos diagramas se utilizó un umbral equivalente a p > 1.000e-4.

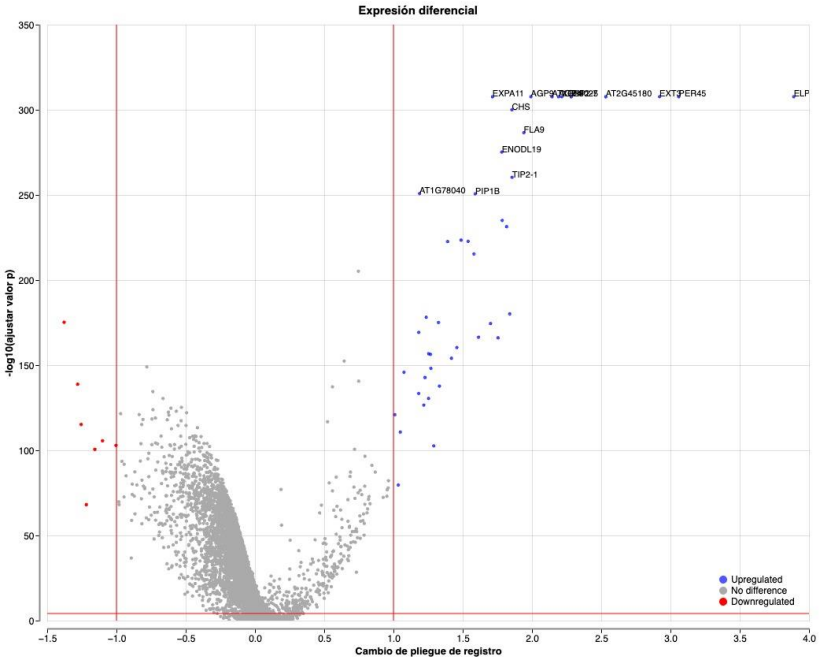

Figura 36*. Expresión diferencial entre el Cluster 0 con el resto de Clusters en el grupo protoplastos*

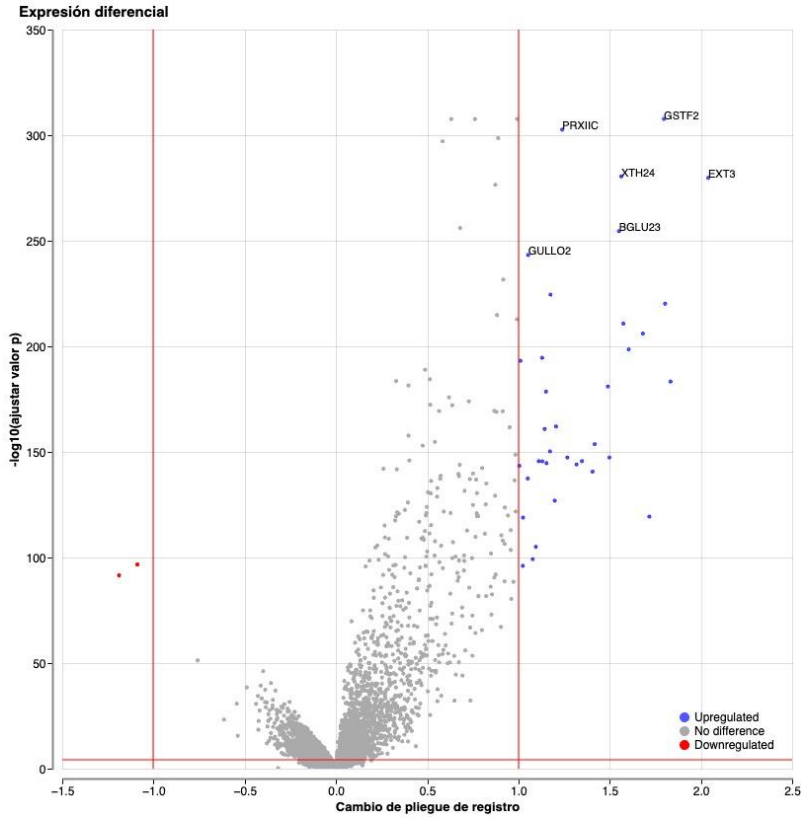

Figura 37. Expresión diferencial entre el Cluster 0 con el resto de Clusters en el grupo aislamiento de núcleos

En la Figura 38 se visualiza el diagrama de volcán comparando el conjunto de datos del cluster 0 con el resto de clusters louvaina dentro de los grupos de datos pr1, pr2, mr1, mr2 y mr3 con un umbral equivalente a p < 1.000e-4.

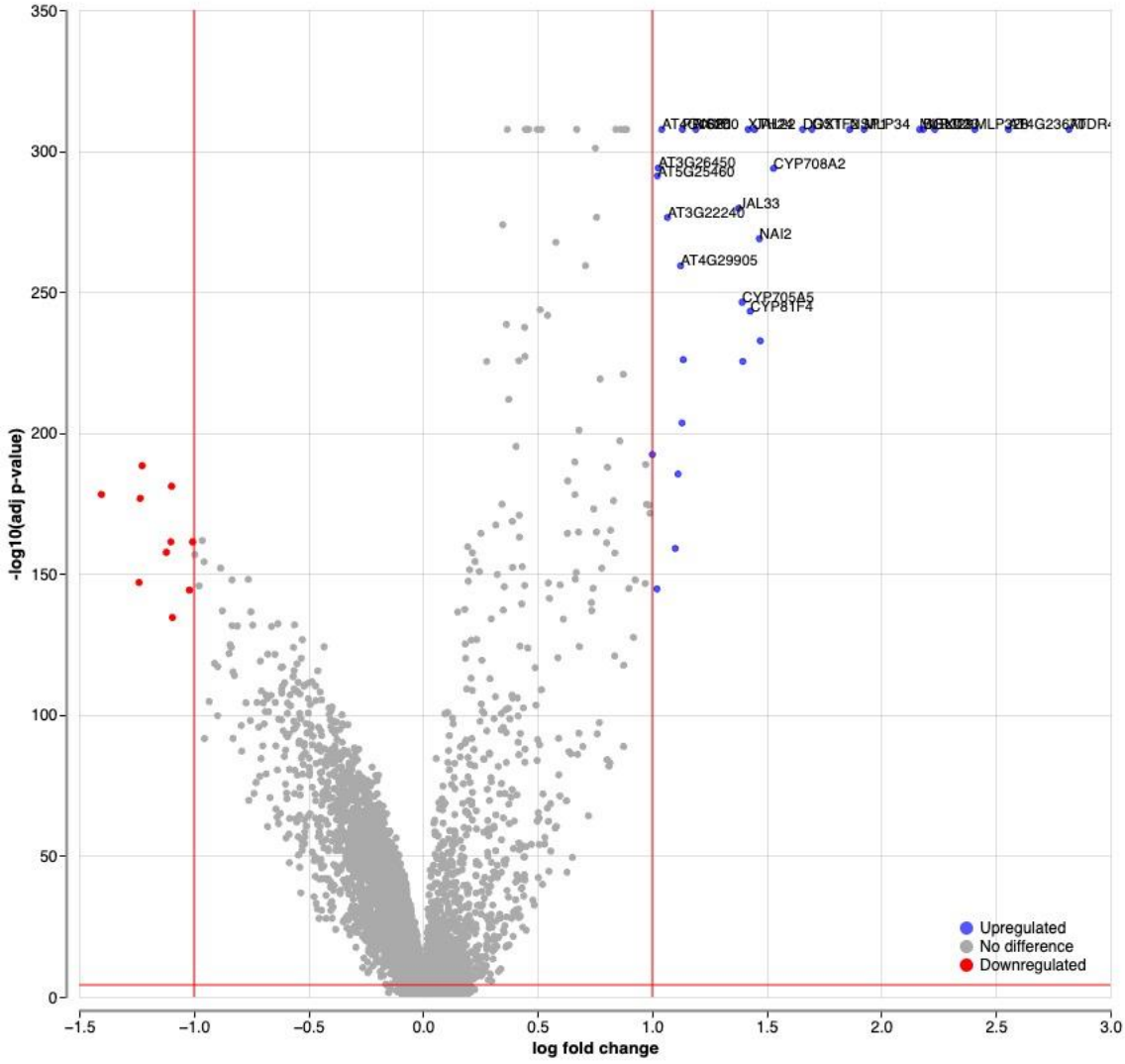

Figura 38. Expresión diferencial entre el Cluster 0 con el resto de Clusters en los 2 métodos protoplastos y aislamiento de núcleos

6. Conclusiones y Recomendaciones

Inicialmente se cargó un promedio estimado de 4316 y 3198 células para los métodos de protoplastos y aislamiento de núcleos respectivamente. De igual manera se cargó un promedio de 23782 y 22831 de genes para los mismos métodos. Se aplicó varios filtros como el clasificador, distribución del tamaño de celda, contenido mitocondrial, número de genes frente a UMI y doblete. Después de aplicar los 5 filtros, se eliminaron algunos datos no deseados y de mala calidad de cada muestra individual. Finalmente se trabajó con un promedio estimado de 3977 y 2899 células para protoplastos y aislamiento de núcleos respectivamente. Se trabajó con un promedio de 23587 y 22528 genes para los mismos métodos.

Se cargó un promedio estimado del número medio de genes por célula de 1558 y 1068 para los métodos de protoplastos y aislamiento de núcleos respectivamente. De igual manera se cargó un promedio estimado de recuentos medios de UMI por celda de 4332 y 1676 para los mismos métodos. Se aplicó los 5 filtros: clasificador, distribución del tamaño de celda, contenido mitocondrial, número de genes frente a UMI y doblete, esto provocó el decremento de estos valores debido a que en el procedimiento de datos se eliminaron algunos datos no deseados y de mala calidad de cada muestra. Finalmente se trabajó con un promedio estimado de 1432 y 1018 del número medio de genes por células para protoplastos y aislamiento de núcleos respectivamente. Se trabajó con un promedio de 3846 y 1571 genes para los mismos métodos.

El procedimiento de datos de la herramienta Cellenics consta de 7 pasos. En los pasos 1 y 2 sobre el filtro de clasificador y el filtro de distribución del tamaño de celda no se modificaron las muestras cargadas inicialmente en la herramienta, así como también, el paso 3 sobre el filtro de contenido mitocondrial no se decrementó el total porque no habían datos de secuencias mitocondriales. En los pasos 4 y 5 sobre el número de genes frente a UMI y el filtro doblete si presentó un decremento en los datos procesados. En el paso 6 se integraron varios conjuntos de datos de muestras para eliminar los efectos del lote mediante el método Seurat V4 y se realizó también la reducción de dimensionalidades. Finalmente el paso 7 configuró la incrustación usando el método estocástico UMAP para visualizar los grupos de puntos de datos y sus proximidades

relativas con una distancia mínima de 0.3, valor que permitó que los puntos incrustados se distribuyan de manera más uniforme.

Se ha observado una correlación entre dos variables: el número de genes y el número de moléculas UMI en diferentes muestras en los métodos de protoplastos y aislamiento de núcleos. La correlación es cercana a 1, porque hay una fuerte relación positiva entre estas dos variables, es decir, si el número de genes aumenta en una muestra, también tiende a aumentar el número de moléculas UMI, y viceversa. Concluyéndose que los resultados de los dos métodos están muy relacionados entre sí y que los datos son consistentes y confiables.

El proceso de filtro de doblete afecta las muestras pr1, pr2, mr1, mr2 y mr3, reduciendo el número de células, eliminando genes y cambiando las estadísticas de mediana de genes y recuentos de UMI por célula. En el histograma de frecuencias del paso de filtro de doblete del método de protoplastos se muestran semejantes de acuerdo al umbral de probabilidades para las muestras pr1 = 0.55240 y pr2 = 0.50592 siendo estos resultados importantes para comprender cómo se están procesando y filtrando los datos en el estudio. Por otro lado para el método de aislamiento de núcleos los umbrales de probabilidad para las muestras mr1 = 0.86447, mr2 = 0.68752 y mr3 = 0.55810 son diferentes para cada muestra, lo que podría influir en las diferencias observadas en los resultados.

En la visualización de la integración de datos de los dos métodos de manera individual, se concluye que las muestras pr1 y pr2 están agrupadas de manera similar. Por otro lado, las muestras mr1, mr2 y mr3 no están agrupadas totalmente, hay una mayor concentración de mr2 en una sola área. Al integrar los datos de los dos métodos de protoplastos y aislamiento de núcleos en un solo conjunto de datos se concluye que las muestras no están agrupadas en su totallidad de manera similar, mr3 sigue una mayor concentración en tres diferentes áreas esta vez.

En el proceso de configuración de la incrustación se utilizó el método UMAP para reducir la dimensionalidad de los datos y visualizar grupos de puntos de datos, empleando el método de agrupación Lovaina para agrupar los puntos de datos en clusters. Los parámetros como la distancia mínima en UMAP y la resolución en Lovaina se ajustaron

para controlar la forma en que se realizan estas transformaciones y agrupaciones, lo que ayuda a obtener una representación efectiva de la estructura subyacente de los datos y a identificar grupos de interés en el análisis de scRNA-seq. En este proceso se ha usado la distancia mínima de 0.3 como un parámetro para controlar cuán fuertemente se permite que la incrustación comprima puntos, es decir, la representación de los datos en un espacio de baja dimensión. Por otro lado se usó el parámetro de resolución de 0.8 para controlar cuántos grupos se obtienen. Usando esta configuración en los dos métodos protoplastos y aislamiento de núcleos se han creado 21 agrupaciones para cada método. Observando las figuras, en el primer método los grupos están más definidos que en el segundo método.

Se ha creado el gráfico de codo para la integración de todos los datos de las muestras de ambos métodos protoplastos y aislamiento de núcleos. Se ha definido la configuración predeterminada para la cantidad de Componentes Principales (PC) igual a 30 y un porcentaje de variación explicada del 89.34%. El gráfico muestra una curva de forma de codo, donde la variación explicada aumenta rápidamente al principio y luego se estabiliza. En este caso, con 30 PC, se ha logrado una variación explicada del 89.34%, lo que sugiere que estos PC son una representación sólida de la variabilidad en el conjunto de datos.

El gráfico de frecuencias coloreado por las muestras pr1, pr2, mr1, mr2 y mr3 en los 26 clusters se ha utilizado para mostrar cómo cada muestra está distribuida en diferentes clusters, revelándose cuántas muestras pertenecen a cada uno de los clusters, en este caso, las muestras pr1, pr2 y mr3 tienen la mayor contribución, esto se debe a que el número estimado de células y el número total de genes en estas tres muestras son las más altas.

Los genes XCP1, PER66, AT3G27200, ADF9 con mayor dispersión para el método protoplastos no son los mismos genes para el método de aislamiento de núcleos siendo estos FAR3, AT1G03920, AT1G43020, DTX3, sin embargo, los 4 genes PER66, XCP1, AT3G27200, ADF9 con mayor dispersión al combinar los dos métodos coinciden con los 4 genes con mayor dispersión del método protoplastos.

Los gráficos de violín mostraron la distribución de las expresiones de varios genes en dos métodos diferentes protoplastos y aislamiento de núcleos. Estos gráficos ayudan a visualizar cómo se distribuye la actividad genética en las células y a identificar células en las que un gen específico no se expresa. En el caso del primer método, se observa que los genes XCP1, PER66, AT3G27200, ADF9 no se expresan en algunos clusters. Por ejemplo, el gen XCP1 no se expresa en los clusters 7, 11, 16 y 18. En el segundo método los genes FAR3 y DTX3 se expresan en todos los clusters, mientras que los genes AT1G43020 y AT1G03920 solamente en un cluster no se expresan, custer 19 y 18 respectivamente.

Se ha documentado y se ha justificado adecuadamente los pasos para el procesamiento de datos, se ha comunicado claramente los resultados finales teniendo en cuenta las limitaciones y consideraciones relacionadas con estos pasos. Esto ha ayudado a garantizar la transparencia y la validez del análisis.

Se recomienda comprender cómo la aplicación de los diferentes filtros para futuras investigaciones afecta a los resultados obtenidos en este trabajo. Es posible que la eliminación de datos de baja calidad o datos no desados sea necesaria para garantizar resultados precisos, pero también es importante ser conscientes de cualquier posible sesgo introducido por los filtros.

Se recomienda realizar un análisis adicional que permita evaluar la calidad de los datos después de aplicar los filtros debido a la eliminación de datos presentes en las muestras de los métodos protoplastos y aislamiento de núcleos. Se puede usar métricas de control de calidad para confirmar que los datos finales con los que se analizó los resultados cumplen con los estándares necesarios.

Es importante mencionar las herramientas y métodos que fueron utilizados en el procesamiento de datos, como por ejemplo, Seurat V4 que permite realizar diversas tareas en el análisis de datos de ScRNA-seq incluyendo la normalización de datos, agrupación de célilas, análisis de expresión diferencial, vizualización de datos hasta el análisis de trayectoria celular y la técnica UMAP que permite la reducción de dimensionalidad y visualización utilizada en el análisis de datos de alta dimensionalidad, incluyendo la secunciación de ARN de célula única.

Se recomienda mencionar posibles sesgos en el proceso de filtro de doblete y como podría afectar los resultados. En este estudio el porcentaje del decremento de los valores antes y después de aplicar este proceso son mínimos.

Se recomienda realizar experimentos de validación en el laboratorio para corroborar las diferencias observadas en la agrupación de las muestras entre los métodos. Este proceso podría ayudar a determinar si las diferencias son biológicamente significativas o si puede ser debido a variaciones técnicas. Ayudaría a identificar muestras atípicas.

Se recomienda usar otros valores para los Componentes Principales (PC) y el porcentaje de variación explicada ya que los valores 30 y 89.34% registrados en este trabajo fueron definidos por el usuario. Realizando un análisis de codo con diferentes valores de PC y el porcentaje se podría comprender si estos valores fueron la elección óptima.

Debido a la cantidad de información precisa obtenida de la herramienta de visualización y análisis de datos scRNA-seq de código abierto llamado Cellenics, se recomienda ampliar la investigación a nuevas áreas y estudios.

El uso de la herramienta Cellenics no requiere que el usuario esté familiarizado con lenguajes de programación ni que tenga hardware especializado ni sistemas informáticos complejos, tan solo se necesita de la conectividad a internet porque está basada en la nube para datos scRNA-seq.

7. Referencias

DeLaughter, D. M. (2018). The Use of the Fluidigm C1 for RNA Expression Analyses of Single Cells. *Current Protocols in Molecular Biology*, *122*(1), e55. https://doi.org/10.1002/cpmb.55

Gao, C., Zhang, M., & Chen, L. (2020). The Comparison of Two Single-cell Sequencing Platforms: BD Rhapsody and 10x Genomics Chromium. *Current Genomics*, *21*(8), 602-609. https://doi.org/10.2174/1389202921999200625220812

Gierahn, T. M., Wadsworth, M. H., Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., Fortune, S., Love, J. C., & Shalek, A. K. (2017). Seq-Well: Portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods*, *14*(4), Article 4. https://doi.org/10.1038/nmeth.4179

Goetz, J. J., & Trimarchi, J. M. (2012). Transcriptome sequencing of single cells with Smart-Seq. *Nature Biotechnology*, *30*(8), Article 8. https://doi.org/10.1038/nbt.2325

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, *2*(3), 666-673. https://doi.org/10.1016/j.celrep.2012.08.003

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., & Amit, I. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*, *343*(6172), 776-779. https://doi.org/10.1126/science.1247651

Kim, J., & Marignani, P. A. (2022). Single-Cell RNA Sequencing Analysis Using Fluidigm C1 Platform for Characterization of Heterogeneous Transcriptomes. *Methods in Molecular Biology (Clifton, N.J.)*, *2508*, 261-278. https://doi.org/10.1007/978-1-0716-2376-3_19

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., & Kirschner, M. W. (2015). Droplet Barcoding for Single-Cell

Transcriptomics Applied to Embryonic Stem Cells. *Cell*, *161*(5), 1187-1201. https://doi.org/10.1016/j.cell.2015.04.044

Liu, C., Wu, T., Fan, F., Liu, Y., Wu, L., Junkin, M., Wang, Z., Yu, Y., Wang, W., Wei, W., Yuan, Y., Wang, M., Cheng, M., Wei, X., Xu, J., Shi, Q., Liu, S., Chen, A., Wang, O., … Liu, L. (2019). *A portable and cost-effective microfluidic system for massively parallel single-cell transcriptome profiling* (p. 818450). bioRxiv. https://doi.org/10.1101/818450

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202-1214. https://doi.org/10.1016/j.cell.2015.05.002

M. Klein, A., & Macosko, E. (2017). InDrops and Drop-seq technologies for single-cell sequencing. *Lab on a Chip*, *17*(15), 2540-2541. https://doi.org/10.1039/C7LC90070H

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, *10*(11), Article 11. https://doi.org/10.1038/nmeth.2639

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C., Schroth, G. P., & Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, *30*(8), Article 8. https://doi.org/10.1038/nbt.2282

Sheng, K., Cao, W., Niu, Y., Deng, Q., & Zong, C. (2017). Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nature Methods*, *14*(3), Article 3.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature, 408(6814), 796–815. https://doi.org/10.1038/35048692