Trabajo Fin de Máster - Curso 2023/2024

Comparativa de Métodos de Análisis en Transcriptómica Espacial para la Detección de Genes Espacialmente Variables

Irene Bonache Gómez

Tutores: Francisco García García, Cristina Galiana Roselló y Anabel Forte Deltell

---

Facultat de Ciències Matemàtiques

Máster Universitario en Bioestadística

Glosario de términos.

Atlas tisulares: colección sistemática y detallada de imágenes, gráficos o mapas en las cuales se muestra la distribución espacial de distintos tipos de tejidos en un organismo.

Barcode: secuencia corta de DNA o RNA diseñada para identificar de manera inequívoca una muestra o una secuencia genética en un análisis.

DNA: molécula del interior de las células, denominada como ácido desoxiribonuicleico (DNA por sus siglas en inglés), que contiene la información genética necesaria para que las personas y la mayoría de organismos se desarrollen y crezcan (National Cancer Institute, s.f.).

DNA codificante: aquel DNA que contiene los genes que producen proteínas dentro del genoma (Roura, 2024).

RNA: uno de los dos tipos de ácido nucleico que elaboran las células. El RNA contiene información copiada del DNA (el otro tipo de ácido nucleico). Las células elaboran varias formas diferentes de RNA y cada forma cumple una función específica en la célula. Muchas formas de RNA cumplen funciones relacionadas con las proteínas. El RNA también es el material genético de algunos virus en lugar del DNA. El RNA se puede producir en el laboratorio y se usa en estudios de investigación (National Cancer Institute, s.f.[d]).

mRNA: tipo de RNA que se encuentra en las células. El RNA mensajero tiene la información genética que se necesita para elaborar las proteínas y lleva esta información desde el DNA en el núcleo de la célula al citoplasma donde se elaboran las proteínas (National Cancer Institute, s.f.[c]).

cDNA: copia del RNA mensajero sintetizada en el laboratorio por medio de de una enzima denominada transcriptasa inversa.(Navarra, s.f.).

Histología: estudio de los tejidos y las células bajo un microscopio (National Cancer

Institute, s.f.[e]).

Metabolitos: sustancia que el cuerpo elabora o usa cuando descompone los alimentos, los medicamentos o sustancias químicas; o su propio tejido (por ejemplo, la grasa o el tejido muscular). Este proceso, que se llama metabolismo, produce energía y los materiales necesarios para el crecimiento, la reproducción y el mantenimiento de la salud. También ayuda a eliminar las sustancias tóxicas (National Cancer Institute, s.f.[f]).

Nucleótidos: elemento fundamental de los acidos nucleicos (DNA o RNA). Los nucleótidos se unen por sus extremos para formar el DNA o RNA. (National Cancer Institute, s.f.[b]).

Oligonucleótidos: pequeñas secuencias de DNA o RNA que se pueden unir a moléculas específicas de RNA. Esto bloquea la capacidad del RNA de hacer que una proteína funcione de otras maneras (National Cancer Institute, s.f.[g]).

Transcripción: proceso mediante el cual una célula elabora una copia de RNA de una secuencia de DNA. Esta copia de ARN, es conocida con el nombre de RNA mensajero (mRNA) (National Cancer Institute, s.f.[h]).

Transcrito: RNA producido a partir de una región específica del DNA mediante el proceso de transcripción.

***Spot*:** En transcriptómica espacial, pequeña región muestreada de un tejido constituida por un conjunto de células.

Resumen.

La transcriptómica espacial es una rama de la biología molecular en la que se integran técnicas de transcriptómica con la información espacial que proporciona la localización de las células dentro de un tejido. Gracias a esta técnica, se ha podido mapear y analizar los niveles de expresión de los genes en múltiples regiones de un tejido, sin perder la información espacial sobre la organización de las células en dicho tejido.

En este estudio se ha realizado una comparativa de tres métodos para la identificación de genes que presentan una mayor variabilidad en su expresión a lo largo de los diferentes dominios espaciales de un fragmento de tejido, mostrando las diferencias tanto a nivel metodológico como en los resultados obtenidos. La detección de estos genes, denominados como genes espacialmente variables, permite reducir la alta dimensionalidad de los datos de la transcriptómica espacial. Además, en este trabajo también se muestran las necesidades computacionales y de memoria de los diferentes métodos.

Para la identificación de estos genes se utilizó una base de datos en la que había seis muestras de tejido cerebral, tres de pacientes sanos y tres de enfermos de Alzheimer. Tras la utilización de los diferentes métodos, se obtuvieron resultados muy dispares en función del método empleado.

Por último, habiendo sido detectados los genes espacialmente variables, se empleó un algoritmo de agrupamiento para definir grupos con los genes detectados. De las tres aproximaciones evaluadas, con SPARK-X se identificaron un mayor número de genes altamente variables, definiendo posteriormente una clusterización más homogénea, en un tiempo y memoria muy inferiores a los otros métodos. Aunque se dispone de una multitud de otros métodos, en nuestro estudio determinamos a SPARK-X como el más óptimo para definir las diferentes capas histológicas en los estudios de transcriptómica espacial.

Palabras clave: Transcriptómica espacial, genes espacialmente variables, *clustering*, Alzheimer, SPARK, SPARK-X, SpatialDE.

Índice

Índice de figuras

Índice de tablas.

1. Introducción.

1.1. Motivación y objetivos.

A lo largo de este trabajo se explorarán y evaluarán los métodos de análisis de datos de transcriptómica espacial, una técnica novedosa y prometedora, capaz de dar resultados mucho más informativos y detallados que los proporcionados por las tecnologías utilizadas hasta hace muy poco, como el RNA-seq. Esta mayor precisión, se debe a que estos nuevos métodos incorporan, no solo la evaluación del nivel de expresión, sino que también se tienen en cuenta la posición espacial y la histología.

En concreto, este trabajo se centrará en la identificación de aquellos genes que muestren una variabilidad espacial significativa, es decir, los niveles de expresión de los genes varían en función de su localización espacial. La importancia de la identificación de este tipo de genes se debe a que tras numerosos estudios, se ha observado que dicha variabilidad puede ser en muchos casos, un marcador de la presencia de algún tipo de enfermedad. Por lo que es posible que gracias a esta tecnología, pueda llevarse a cabo una identificación precoz de dichas enfermedades, tratándolas antes, y consiguiendo así mejores resultados, o incluso logrando que dicha enfermedad no llegue a desarrollarse.

En un inicio, el objetivo de este trabajo era caracterizar las diferencias de sexo en enfermedades neurodegenerativas mediante los nuevos métodos de transcriptómica espacial. Sin embargo, debido la novedad de la tecnología, no ha sido posible identificar conjuntos de datos que incluyeran muestras de ambos sexos. Por lo tanto, se ha procedido con el análisis de un estudio en el que había dos grupos experimentales, sanos y enfermos, siendo ahora el objetivo del trabajo caracterizar esas diferencias utilizando esta tecnología.

1.2. Tecnologías de alto rendimiento.

Se definen como tecnologías de alto rendimiento u “ómicas”, aquellas ciencias que permiten estudiar un gran número de moléculas, las cuales están implicadas en el funcionamiento de un organismo. En las últimas décadas, se han llevado a cabo grandes avances a nivel tecnológico que han hecho posible el estudio a gran escala de muchos genes,

proteínas, metabolitos y transcritos, estableciéndose como áreas de estudio la genómica, proteómica, metabolómica y transcriptómica, entre otras. Cada una de estas áreas ha ayudado a un mejor entendimiento de la causa de ciertas enfermedades. La traslación de estas ómicas a la práctica clínica podrá utilizarse para hacer un diagnóstico más temprano o para prevenir el desarrollo de una enfermedad. Además, se pretende desarrollar una medicina personalizada, donde cada individuo llevará un tratamiento para una determinada enfermedad, acorde a su información genética y a su medio ambiente (Frigolet y Gutiérrez-Aguilar, 2017).

1.2.1. Transcriptómica.

La transcriptómica es la disciplina "ómica" que se encarga de analizar la expresión de los genes en forma de RNAs, móleculas conocidas como transcritos por derivar de la transcripción directa del DNA. El RNA es específico de cada célula y de las condiciones fisiopatológicas para un determinado momento. Por ejemplo, el RNA extraído de células del músculo será diferente a las células del hígado. A causa del carácter dinámico de la transcripción, la transcriptómica se realiza en un tejido y en un tiempo específico (Frigolet y Gutiérrez-Aguilar, 2017).

Hace un tiempo, se pensaba que una gran cantidad del DNA que no se transcribía a RNA mensajero (mRNA), el tipo de RNA más abundante, carecía de función alguna. Sin embargo, se descubrió que además del mRNA, también existían otros transcritos no codificantes como son: 1) los miRNA (microRNA, secuencias de 21-25 nucleótidos) o, 2) los lncRNA (RNA largos no codificantes, >200). Estos tienen la función de regular la expresión de los transcritos que sí son codificantes. Por lo que actualmente se sabe que estas regiones del DNA, sí que tienen una función, regular la expresión de diversos genes.

Las tecnologías que se utilizan para analizar el mRNA son: 1) microarreglos o 2) diferentes tecnologías o plataformas de secuenciación del RNA de alto rendimiento (RNA-seq por sus siglas en inglés). En los microarreglos se hibrida el mRNA de un determinado tejido a secuencias de genes previamente conocidos. De esta forma, se pueden hacer comparaciones entre casos o controles o, simplemente, ver qué genes se expresan mayoritariamente en ciertas condiciones. En cambio, el RNA-seq consiste en secuenciar todos los transcritos

presentes en esas condiciones, encontrando nuevos transcritos que no se conocían anteriormente y nuevos genes involucrados en una enfermedad (Frigolet y Gutiérrez-Aguilar, 2017).

1.2.2. Transcriptómica espacial.

La transcriptómica espacial (ST por sus siglas en inglés) es una tecnología de vanguardia que permite analizar el transcriptoma de un tejido, proporcionando además, información acerca de la localización de los genes en el mismo, en función de sus niveles de expresión. Gracias a esta técnica, se han podido mapear y analizar los niveles de expresión de los genes en múltiples regiones de un tejido, sin perder la información espacial sobre la organización de las células en dicho tejido.

Esta nueva tecnología surgió a raíz de la necesidad de comprender de qué manera la expresión génica puede variar no solo cuantitativamente, sino también en términos de localización dentro de un tejido. Las técnicas tradicionales como el RNA-seq son capaces de proporcionar información sobre los niveles de expresión a nivel de población celular, pero sin posibilidad de conocer la disposición de esta población celular a lo largo del tejido.

Como resultado de la combinación de técnicas de hibridación *in situ* con métodos de secuenciación masiva, se consiguieron los primeros avances en ST. La hibridación *in situ* es una técnica con la cual es posible visualizar la localización de ácidos nucleicos en tejidos y, gracias a ella, se forjó la base para poder incorporar la dimensión espacial a los estudios transcriptómicos.

Ha sido tal la revolución que ha ocasionado esta nueva tecnología, que en el año 2020, fue nombrada "Método del Año"por la revista *Nature methods*, título que solo se otorga a desarrollos metodológicos que tienen un gran impacto en el progreso científico (Dlongwood, 2024).

En la actualidad se disponen de hasta 7 plataformas diferentes de ST, las cuales pueden clasificarse en dos tipos en función de la tecnología utilizada:

- Tecnologías basadas en imágenes.

- Tecnologías basadas en secuenciación.

La principal diferencia entre ambas es la metodología utilizada para determinar la localización espacial de un determinado mRNA a lo largo de un tejido, así como su abundancia.

A continuación, se muestran más detalladamente los dos tipos de tecnologías mencionadas previamente.

Tecnologías basadas en imágenes.

Actualmente, las tecnologías de ST basadas en imágenes se fundamentan en el uso de hibridación por fluorescencia *in situ* de célula única (smFISH por sus siglas en inglés) como método para visualizar y detectar la presencia o la localización de secuencias concretas de DNA o RNA dentro de muestras biológicas, como pueden ser células o tejidos. El procedimiento general entre las diferentes plataformas que utilizan esta tecnología es muy similar. Este *workflow* incluye la preparación de las láminas, la permeabilización del tejido, hibridación con la sonda fluorescente, captura de imágenes, descodificación de colores y cartografía espacial (figura 1). Este procedimiento se repite secuencialmente una gran cantidad de veces, para ir así actualizando la imagen del tejido.

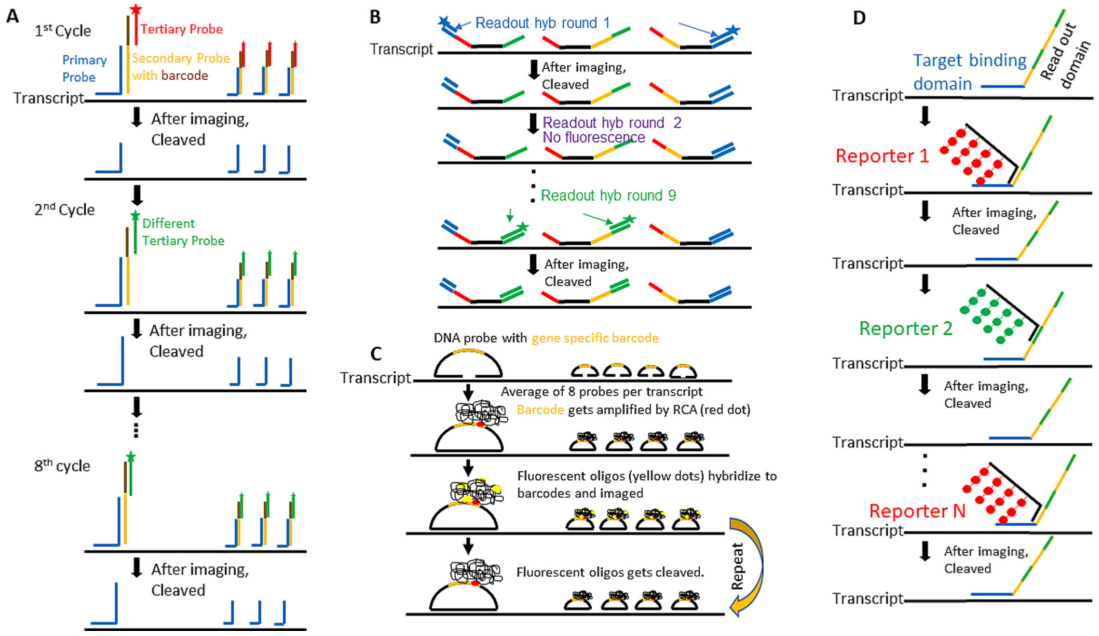


Figura 1: Tecnologías ST basadas en imágenes. Adaptada de Ye Wang et al., 2023.

Esta tecnología puede proporcionar una resolución unicelular o subcelular con una alta eficiencia de captura de RNA. Sin embargo, presenta algunas limitaciones técnicas como la aglomeración óptica y el largo tiempo de captura de imágenes para áreas pequeñas. Además, con el aumento de las rondas de hibridación se predispone a una mayor probabilidad de error de unión de las sondas. Por ello, las plataformas comerciales basadas en esta tecnología de ST se limitan a evaluar entre unos cientos y unos miles de genes.

Tecnologías basadas en secuenciación.

La mayoría de las plataformas basadas en secuenciación combinan la tecnología de *microarrays* clásicos con las tecnologías de secuenciación de nueva generación para hacer posible la localización espacial de la expresión génica. Brevemente, en el *microarray* se adhieren millones de oligonucleótidos que incluyen en su secuencia identificadores únicos (*barcodes*), para determinar su localización, así como las secuencias poli(dT) y otros dominios. Sobre el microarray se dispone la sección de tejido a visulizar. A continuación, el tejido se permeabiliza, liberándose los RNAs a la matriz donde se secuenciará. Los RNA se retrotranscribirán a cDNA, DNA de secuencia complementaria al mRNA. Finalmente, los cDNA se secuenciarán, siendo fácilmente su localización en el tejido al incluir el *barcode* en su secuencia. Actualmente, se comercializan tres plataformas basadas en secuenciación: 10× Visium, GeoMx Digital Spatial Profiler (DSP) y BMKMANU S1000, siendo la más empleada 10× Visium, y a partir de la cual se han obtenido los datos que se han utilizado en este trabajo (figura 2).

Las tecnologías de ST basadas en secuenciación proporcionan un análisis del transcriptoma completo con una gran área de imagen y a un tiempo de escaneado corto, pero su eficiencia a la hora de capturar RNA es relativamente menor. Además, aún siendo la tecnología de las plataformas más utilizadas como 10× Visium y GeoMx DSP, no pueden proporcionar una resolución de célula única. Cabe destacar, que en este año 2024 se lanzará la plataforma Stereo-seq, la cual combina los puntos fuertes de las tecnologías basadas en secuenciación e imagen, proporcionando así un análisis del transcriptoma completo con resolución unicelular.

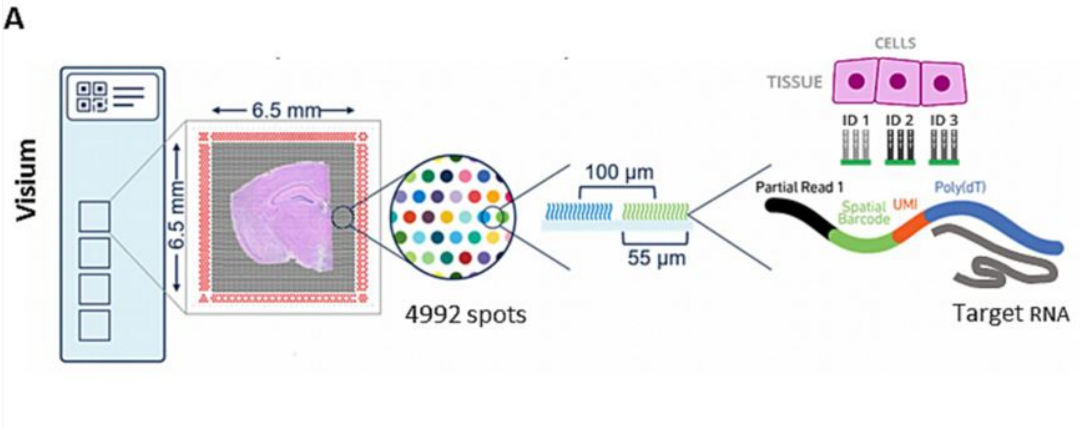


Figura 2: Tecnologías basadas en secuenciación. UMI (Unique Molecule Identifier): Identificador único de gen. Spatial Barcode: Identificador único de spot (localización). Adaptada de Ye Wang et al., 2023.

1.2.3. ¿Cuándo utilizar la transcriptómica espacial?

La transcriptómica espacial tiene una amplia gama de aplicaciones en biología, inmunología, oncología, etc, en tejidos sanos y enfermos, tanto en humanos como de ratones u otras especies. Muchas preguntas biológicas, que son aplicables a la tecnología de secuenciación unicelular, u otras a las que ésta no puede dar respuesta, pueden ser resueltas gracias a su uso. Además, la transcriptómica espacial posee un gran potencial en aplicaciones clínicas que necesitan una mayor precisión y sensibilidad.

A continuación, se van a enumerar algunas situaciones en las cuales el uso de la transcriptómica espacial puede ser de gran utilidad. Algunos ejemplos son:

- **Análisis espacio-temporal del desarollo de tejidos.** Aunque se han aplicado muchas metodologías para estudiarla, la falta de información sobre la organización estructural celular, junto con la baja resolución espacial y el reducido número de genes, la potencia de esta tecnología se ha visto limitada. Las tecnologías ST se han utilizado ampliamente en el estudio del desarrollo de embriones, tejidos y órganos de diversas especies, destacando la embriogénesis de los mamíferos. La ST permite investigar la dinámica transcriptómica espacio-temporal en el ecosistema celular natural en relación con la especificación del destino celular, la interacción célula-célula y la formación de órganos.

- **Atlas espaciales de regiones o tejidos específicos.** Las tecnologías de ST pueden proporcionar un mapa espacial carente de sesgo de un tejido, generando varios atlas tisulares de referencia ampliables. Algunos ejemplos pueden ser: atlas de tejido renal humano en la salud y la enfermedad, mapas multiómicos espaciales del remodelado cardíaco, el pulmón humano y varios atlas de tejido cerebral de las regiones preópticas hipotalámicas, el hipocampo y el cerebelo.

- **Mecanismos moleculares de desregulación génica o celular durante una enfermedad.** Actualmente, la comprensión de los mecanismos latentes a múltiples enfermedades está limitada por la imposibilidad de examinar directamente las células que provocan dicha enfermedad y su microentorno en condiciones naturales. Es de particular interés el estudio de Galeano Niño et al., 2022 en el que se utilizó la tecnología 10x Visium para descubrir las interacciones celulares y moleculares huésped-microbio en el carcinoma oral de células escamosas y el cáncer colorrectal.

- **Heterogeneidad y microambiente celular en tumores.** La continua interacción entre las células inmunitarias extrínsecas y las células tumorales intrínsecas tiene una gran influencia tanto en la progresión del tumor como en la metástasis. La información proporcionada por el análisis cuantitativo de la heterogeneidad tumoral, y la caracterización espacio-temporal de los microambientes inmunitarios tumorales, puede ser crucial a la hora de comprender la metástasis y desarrollar nuevos tratamientos. Las tecnologías de ST permiten un análisis detallado de los perfiles específicos de expresión génica por los que se caracterizan las células tumorales y sus células asociadas, mostrando así la heterogeneidad espacial del tumor y su microentorno completo.

- **Tipos y estados celulares.** Gracias a las técnicas de transcriptómica espacial, la capacidad de las tecnologías de secuenciación unicelular para identificar y caracterizar diferentes tipos y estados celulares, han sufrido una mejora muy significativa. Para llevar a cabo la elucidación funcional de estos tipos celulares, se necesita conocer su localización física y conectividad. A partir de la dimensión espacial proporcionada por las tecnologías de ST, es posible realizar el análisis funcional de las poblaciones celulares, las interacciones entre células adyacentes y la organización estructural de las células.

- **Biomarcadores espacialmente resueltos para enfermedades y tratamientos.** Por último, el uso de estas tecnologías ha facilitado el desarrollo de biomarcadores moleculares, celulares y microestructurales más precisos y sensibles espacialmente, lo que ha permitido mejorar el diagnóstico, el pronóstico y el tratamiento de las enfermedades.

2. Preprocesado de datos.

Una vez contextualizado el trabajo con la descripción de la tecnología a partir de la cual derivan los datos, se procederá a presentar un análisis detallado del preprocesado de los mismos.

El *dataset* seleccionado proviene del trabajo de investigación de S. Chen et al. (2022) titulado "Spatially resolved transcriptomics reveals genes associated with the vulnerability of middle temporal gyrus in Alzheimer's disease". Concretamente, los datos son derivados del análisis por ST, con la plataforma 10x Visium, para 6 cortes histológicos *post mortem* de cerebro. Estas muestras derivan de 3 personas sanas, denominadas como muestra 1, 3 y 5, y de tres pacientes con la enfermedad de Alzheimer (AD, por sus siglas en inglés), muestras 2, 4 y 6. Inicialmente, las dimensiones de cada muestra fueron:

- Muestra 1: 36601 genes medidos en 3742 *spots*.
- Muestra 2: 36601 genes medidos en 4348 *spots*.
- Muestra 3: 36601 genes medidos en 4701 *spots*.
- Muestra 4: 36601 genes medidos en 3445 *spots*.
- Muestra 5: 36601 genes medidos en 4225 *spots*.
- Muestra 6: 36601 genes medidos en 4832 *spots*.

El acceso y descarga del conjunto de los datos se realizó a través del repositorio GEO (Gene Expression Omnibus), (National Center for Biotechnology Information (NCBI), s/f).

2.1. Control de calidad.

Antes de comenzar con el análisis de los genes espacialmente variables propiamente dicho, cada una de las muestras tuvo que pasar por un control de calidad para determinar que los *spots* que las componen no estuviesen en malas condiciones. Para ello, los datos de cada muestra fueron importados y organizados en un objeto de R denominado como SpatialExperiment. Este objeto S4 incluye tanto los datos de expresión o matriz

de conteos, como las localizaciones en píxeles de los diferentes *spots*. Cuando se habla de conteos de un gen, se hace referencia a la cantidad de RNA correspondiente a ese gen que está presente en un área específica o *spot* de una muestra de tejido. Para realizar este preprocesado de los datos, se adoptó el *workflow* del libro *online* **Best Practices for Spatial Transcriptomics Analysis with Bioconductor** (Weber et al., 2024).

En primer lugar, lo primero fue eliminar los *spots* que presentaran un elevado porcentaje de genes mitocondriales, ya que un nivel alto de estos puede indicar daño celular. En segundo lugar, se estudiaron otras dos variables relevantes, estas son:

- *sum*: número total de conteos por *spot*. A esto también se le denomina tamaño de *library*.
- *detected*: número de genes distintos expresados en cada *spot*.

Si alguna de estas dos variables tuviese un valor bajo, habría que eliminar ese *spot*, ya que eso podría indicar baja carga de material genético. Para determinar qué valores eran los apropiados, se realizaron 12 histogramas (figuras 13 y 14), 2 para cada muestra, siendo uno para la variable *sum* y otro para *detected*. Finalmente, los *spots* que no se tendrían en cuenta son aquellos que se encontraban en la cola inferior de las distribuciones, es decir, los que estaban situados a la izquierda de la línea vertical.

En la tabla 1 se ha recogido un pequeño resumen de las muestras. Como ya se ha mencionado, si la muestra derivaba de un paciente sano, en la columna de tipo paciente se encontrará la palabra "control". Si por el contrario la muestra derivaba de un paciente con Alzheimer, constará como "AD". Por otro lado, la tercera columna hace referencia al número de *spots* en los que alguna de las variables anteriores, *sum* o *detected*, tenían valores bajos, o existían genes mitocondriales, y que, por lo tanto, no se tendrán en cuenta durante el análisis. Por último, en la última columna se encuentra el número de *spots* que, por el contrario, sí que han pasado el control de calidad y que, por consiguiente, serán los que se utilizarán en el análisis.

	Tipo paciente	Descartados	No descartados
Muestra 1	Control	151	3591
Muestra 2	AD	142	4206
Muestra 3	Control	41	4460
Muestra 4	AD	116	3329
Muestra 5	Control	80	4145
Muestra 6	AD	99	4733

Tabla 1: Número de *spots* descartados por muestra, tras el control de calidad.

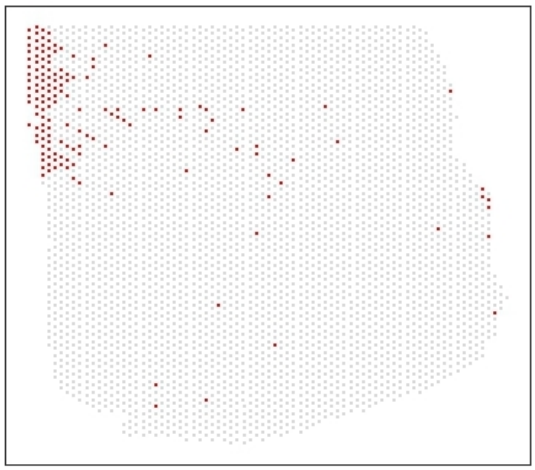
Por lo tanto, si se suman todos los *spots* que no han sido descartados, el objeto espacial final tendría 36601 genes medidos en cada uno de los *spots* de un total de 24464 para todas las muestras.

En la figura 3 se muestran todas las muestras con sus respectivos *spots*, tanto los eliminados, coloreados en rojo, como los que no lo están. Se puede observar que en todos los casos, la mayoría de los *spots* eliminados están muy próximos entre sí, lo que podría significar que por alguna razón esa parte del tejido estaba dañada, por lo que era necesario que no se tuviese en cuenta en el análisis.

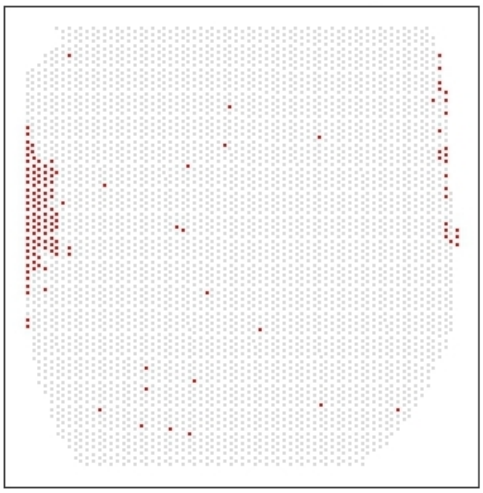
2.2. Normalización de los datos.

Una vez ya se tiene la base de datos definitiva, habrá que proceder a la normalización de los datos. La normalización es un proceso necesario para corregir las diferencias técnicas y el posible sesgo en los datos que no son de interés biológico, permitiendo así que las comparaciones sean más precisas y significativas entre las diferentes muestras.

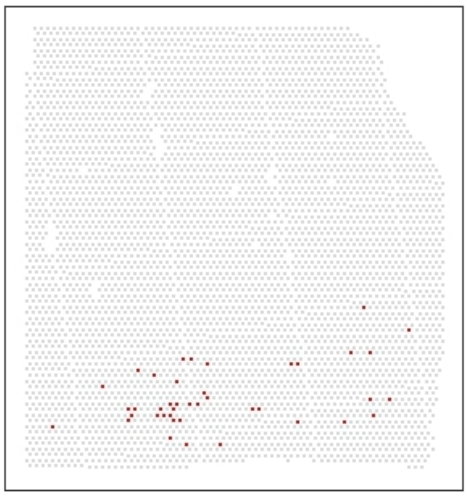
Para esta normalización, se realizará una transformación que tiene en cuenta los conteos totales o tamaño de librería de cada *spot*. Las herramientas que están más extendidas derivan de las empleadas en scRNA-seq, por lo que se asume que cada *spot* puede ser tratado como una única célula. La metodología que se ha empleado se denomina log-normalización por tamaño de librería. En primer lugar, se estima un factor de escalado ($size\ factors$) para cada uno de los *spots*, que representa el sesgo técnico que ha sufrido dicho *spot* durante su procesamiento (tabla 2). A continuación, los conteos totales que hay



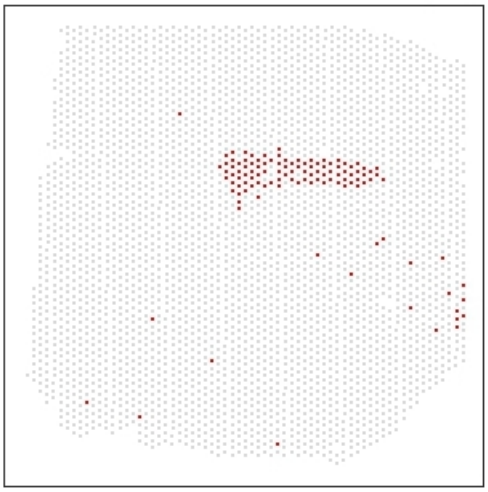
(a) Muestra 1



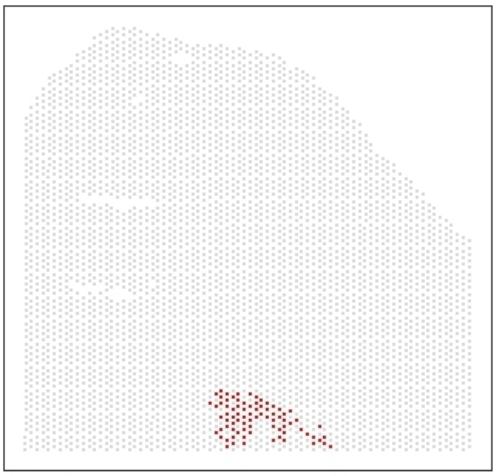
(b) Muestra 2



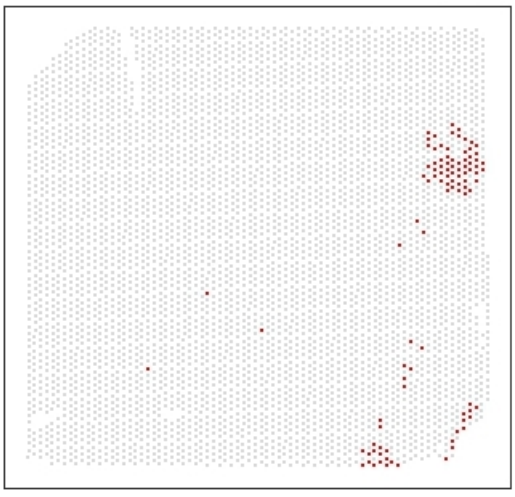
(c) Muestra 3



(d) Muestra 4



(e) Muestra 5



(f) Muestra 6

Figura 3: *Spots* descartados. Los ejes representan la posición horizontal y vertical de los *spots*.

en los diferentes *spots* se dividen por su correspondiente factor de escalado. Por último, al valor obtenido, se le suma una constante de pseudo-conteo, 1 en este caso, y se le aplica una transformación logarítmica en base 2. Esta suma de una constante se lleva a cabo para posibilitar el cálculo de los logaritmos de los conteos que son 0, algo que en este tipo de estudios es muy común. Todo lo mencionado anteriormente puede resumirse en la siguiente expresión:

$$\text{LogNorm}_{ij} = log_2\left(\frac{\text{count}_{ij}}{\text{sizeFactor}_j} + \text{pseudocount}\right)$$

Mínimo	**Primer cuartil**	**Mediana**	**Media**	**Tercer cuartil**	**Máximo**
0.0095	0.5811	0.96	1	1.3178	4.9538

Tabla 2: Descripción de los *size factors*.

Como se puede ver en la tabla 2, el 25 % de los *spots* tienen unos $size\ factors$ menores o iguales a 0.5811, lo que indica que un cuarto de los *spots* necesitan que, en promedio, sus conteos totales se reduzcan en un 41.9 % para conseguir normalizarse. Por otro lado, el valor de la mediana indica que la mitad de los datos tienen unos $size\ factors$ de 0.96, por lo que estos requerirán poca normalización, dado que están muy cercanos a la media, 1. Precisamente este valor de la media es el que generalmente se quiere alcanzar cuando se normaliza. Por último, con respecto al tercer cuartil, el 75 % de los *spots* tienen $size\ factors$ menores o iguales a 1.3178, lo cual indica que un cuarto de los *spots* necesitan que sus conteos aumenten en hasta un 31.8 % para conseguir normalizarse. Una vez se han analizado los diferentes estadísticos obtenidos en la normalización, se puede observar que los $size\ factors$ tienen mucha variabilidad de unos *spots* a otros, y que por lo tanto, en este caso, la normalización es fundamental.

A continuación, se han seleccionado varios genes con los que se pretende mostrar, a modo de ejemplificación, el comportamiento de los datos. En primer lugar, en la figura 4 se muestra cómo se distribuye espacialmente el nivel de expresión del gen ENSG00000134042 en cada una de las muestras. Se puede apreciar como, por ejemplo, en la primera muestra ese gen en concreto estaría menos expresado que en la muestra número 5.

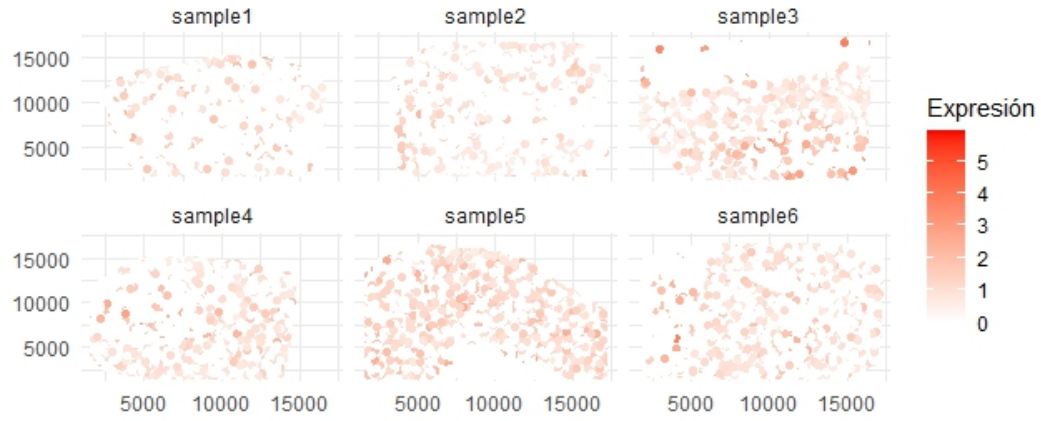


Figura 4: Expresión del gen ENSG00000134042 en cada muestra. Los ejes representan la posición horizontal y vertical de los *spots*.

Por otro lado, también hay que destacar que hay una gran cantidad de genes cuyos niveles de expresión son 0, por lo que se tendrán unos datos con mucha dispersión, y por consiguiente, unas matrices con una sobredispersión muy elevada. Para ejemplificar esto, se ha representado la distribución de los conteos del mismo gen , ENSG00000134042, a lo largo de todos los *spots* para cada una de las muestras (figura 5), viendo cómo claramente el valor que predomina es el 0. Para mostrar que esto no es fruto de la casualidad, se han representado tres genes más, los cuales se encuentran las figuras 16, 18 y 20 del anexo 3.

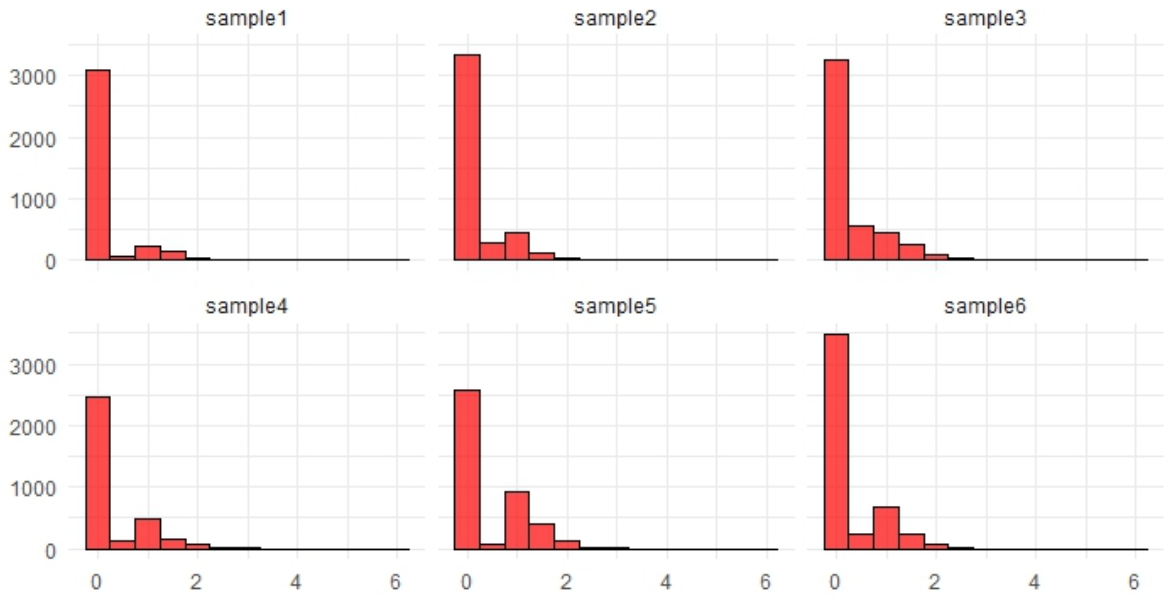


Figura 5: Distribución de la expresión del gen ENSG00000134042 en cada muestra.

3. Métodos.

Dado el gran volumen de datos que generan las técnicas de transcriptómica espacial, hubo que hacer uso del *cluster* del Centro de Investigación Príncipe Felipe (CIPF) (https://bioinfo.cipf.es/ubb/cluster/). Este *cluster* de computadores se compone de 44 nodos, gracias a los que se dispone de 600 unidades centrales de procesamiento (CPUs). En conjunto, presenta una capacidad de almacenamiento de 1 PetaByte y una memoria RAM de 11 TeraBytes. El análisis bioestadístico realizado se ha desarrollado utilizando el lenguaje de programación R (Versión 4.3.2). El código generado durante el análisis y la información del los paquetes de R se puede encontrar en el repositorio *online* de **GitHub**: https://github.com/irene27799/Spatial-transcriptomics-comparison

Como ya se mencionó en la parte introductoria del trabajo, el objetivo es la identificación de genes espacialmente variables. Para ello, se han empleado tres métodos diferentes: SPARK (S. Sun et al., 2020), su versión mejorada SPARK-X (Zhu et al., 2021) y SpatialDE (Svensson et al., 2018). En un inicio también se estudió el método Trendsceek (Edsgärd et al., 2018a), pero debido a la falta de actualización de algunos de los paquetes de R utilizados en su implementación, fue imposible utilizarlo por problemas de versiones. En esta sección se mostrarán detalladamente los desarrollos en los que están basados cada uno de los tres métodos utilizados.

3.1. SPARK

La información de este método se ha obtenido del artículo "Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies"(Sun, Zhu, Zhou, 2020). El objetivo perseguido por este método, es modelar la expresión de los genes recogidos a través de varios tipos de tecnologías basadas en secuenciación. Estas técnicas miden simultáneamente la expresión de m genes diferentes en n localizaciones espaciales en un tejido de interés, denominado muestra. Generalmente, la variable expresión génica se obtiene en forma de conteos. Estos conteos pueden definirse como el número de mRNA con *barcodes* para cualquier transcrito en una sola célula mediante técnicas basadas en smFISH, o como el número de *reads* de secuenciación asignadas a cualquier gen mediante técnicas espaciales basadas en secuenciación. El número de genes, m, varía en función de

las distintas técnicas de secuenciación espacial, puede ir desde un par de cientos (en el caso de smFISH), hasta prácticamente todo el transcriptoma (en el caso de tecnologías basadas en secuenciación). Por otro lado, se ha observado que la composición de la muestra cambia en función de la técnica utilizada. Por ejemplo, en el caso de smFISH, esta consiste en una sola célula, mientras que en la tecnología de transcriptómica espacial, las muestras están compuestas por *spots*. Durante el transcurso del experimento se recogen también las coordenadas de dichas localizaciones muestreadas. Estas localizaciones, en smFISH, pueden considerarse como aleatorias, ya que la expresión se mide en células elegidas aleatoriamente en el tejido, o predeterminadas por los investigadores antes de comenzar con el experimento, como ocurre en 10x Visium. Se denotan como $\boldsymbol{s_i} = (s_{i1}, s_{i2})$ a las coordenadas espaciales para el *spot* i, con $i = 1, \ldots, n$. Dichas coordenadas varían a lo largo de un espacio bidimensional, o lo que es lo mismo $\boldsymbol{s_i} \in \mathbb{R}^2$. Dado que los datos para este trabajo se han recogido en un espacio de 2 dimensiones, solo se tendrá en cuenta el enfoque 2D, pero cabe destacar que tanto el modelo como el método son generales, por lo que también serían capaces de manejar datos tridimensionales dónde, por ejemplo, se registra también la profundidad de la ubicación del *spot* en el tejido, o manejar casos con dimensiones aún mayores en los que también se registran otras coordenadas como por ejemplo el tiempo.

A los genes cuyos niveles de expresión muestran patrones espaciales definidos, se les llamará *SE genes*. Para identificarlos se examinan uno a uno los genes y se modela su nivel de expresión a través de las ubicaciones muestreadas utilizando un modelo espacial lineal generalizado (GLSM). Los modelos espaciales lineales generalizados son también conocidos como modelos geoestadísticos lineales generalizados o como modelos mixtos lineales generalizados espaciales. Este tipo de modelos modelan directamente datos espaciales no gaussianos y utilizan efectos aleatorios para capturar el proceso espacial estacionario latente. Además, también son utilizados comúnmente para la predicción y la interpolación de datos espaciales, con aplicaciones en *disease mapping* espacial y estudios epidemiológicos espaciales. Sin embrago, este trabajo solo se centrará en desarrollar un marco de comprobación de hipótesis para GLSM. Por lo que, para el gen de interés, se denota como $y_i(\boldsymbol{s_i})$ a la expresión génica en términos de conteos del *spot* i. Estas variables explicativas podrían contener información que sea importante ajustar durante el análisis. En este caso

se ha establecido que el factor de normalización para el *spot* i, $N_i(\boldsymbol{s_i})$, se calcule como el sumatorio del número total de conteos de todos los genes en la muestra, ya que el interés principal es analizar el nivel relativo de expresión génica, aunque cabe destacar que este enfoque no es el único posible. Para modelizar los datos se ha utilizado una distribución de Poisson con sobredispersión, es decir:

$$y_i(\boldsymbol{s_i}) \sim Poi(N_i(\boldsymbol{s_i})\lambda_i(\boldsymbol{s_i})), \;\; i = 1, 2, \ldots, n,$$

donde $\lambda_i(\boldsymbol{s_i})$ es un parámetro de tasa de Poisson desconocido que representa la expresión génica subyacente de la observación i. Además, en un ámbito espacial, $\lambda_i(\boldsymbol{s_i})$ puede considerarse como el proceso aleatorio espacial no observado en la localización $\boldsymbol{s_i}$. La variable latente $\lambda_i(\boldsymbol{s_i})$ se modeliza en escala logarítmica como la combinación lineal de tres términos

$$log(\lambda_i(\boldsymbol{s_i})) = \boldsymbol{x}_i(\boldsymbol{s_i})^T\boldsymbol{\beta} + b_i(\boldsymbol{s_i}) + \epsilon_i, \tag{1}$$

donde $\boldsymbol{\beta}$ es un vector de k componentes, incluyendo un intercepto que representa la media del logaritmo de la expresión de los genes a través de todas las ubicaciones espaciales, con $k-1$ coeficientes, uno para cada una de las covariables. ϵ_i es el error residual, el cual es independiente e idénticamente distribuido como $N(0, \tau_2)$, con varianza τ_2. $b_i(\boldsymbol{s_i})$ es un proceso gaussiano de media 0 que modeliza el patrón de correlación espacial entre las diferentes localizaciones, es decir,

$$\boldsymbol{b}(\boldsymbol{s}) = (b_1(\boldsymbol{s_1}), b_2(\boldsymbol{s_2}), \ldots, b_n(\boldsymbol{s_n}))^T \sim MVN(0, \tau_1\boldsymbol{K}(\boldsymbol{s})),$$

donde la covarianza $\boldsymbol{K}(\boldsymbol{s})$ es una función kernel de las localizaciones espaciales $\boldsymbol{s} = (\boldsymbol{s_1}, \ldots, \boldsymbol{s_n})^T$, siendo $\boldsymbol{K}(\boldsymbol{s_i}, \boldsymbol{s_j})$ el elemento ij; τ_1 es un factor de escala del kernel de la covarianza. La elección de dicha función kernel se discutirá con más detenimiento más adelante. En el modelo (1), la covarianza de las variables latentes $log(\boldsymbol{\lambda}(\boldsymbol{s}))$ es $\boldsymbol{\Sigma} = \tau_1\boldsymbol{K}(\boldsymbol{s}) + \tau_1\boldsymbol{I}$, donde $\boldsymbol{\lambda}(\boldsymbol{s}) = (\lambda_1(\boldsymbol{s_1}), \lambda_2(\boldsymbol{s_2}), \ldots, \lambda_n(\boldsymbol{s_n}))^T$, siendo $\boldsymbol{I}$ la matriz identidad de dimensión n. En estadística espacial, τ_1 es comúnmente conocida como el umbral parcial que mide la varianza de la expresión en la localización espacial i, $log(\lambda_i(\boldsymbol{s_i}))$, capturada por los patrones espaciales o la información de localización espacial. Por otro lado, τ_2 se conoce como la pepita que mide la varianza del logaritmo de la expresión del gen i, $log(\lambda_i(\boldsymbol{s_i}))$, debido al ruido aleatorio independiente de las localizaciones espaciales.

En el GLSM (1), testear si los genes muestran un patrón espacial en los niveles de expresión se puede traducir como $\tau_1 = 0$, o lo que es lo mismo, $H_0 : \tau_1 = 0$. La potencia estadística de este contraste dependerá de lo bien que se ajuste la función kernel $\boldsymbol{K(s)}$ al verdadero patrón espacial subyacente mostrado por el gen en cuestión. Por ejemplo, un kernel periódico puede ser muy útil para detectar patrones en los niveles expresión cuando estos son periódicos a lo largo de la localización espacial, mientras que el kernel Gaussiano será de gran utilidad a la hora de detectar patrones que están agrupados en zonas focales. Desafortunadamente, el patrón espacial real para cualquier gen es desconocido y puede cambiar de unos genes a otros. Para asegurar una identificación robusta de los *SE genes*, se utilizan diez tipos diferentes de kernels, cinco de tipo periódico, a los cuales se les van cambiando los parámetros de periodicidad, y otros cinco gaussianos, con valores diferentes en los parámetros de suavizado. Estos diez kernels cubren un rango de posibles patrones espaciales que se han observado de manera frecuente en bases de datos biológicas, sin embargo, tanto el método como el *software* pueden incluir otro tipo de funciones kernel que quizás se ajusten más a los datos que se quieren estudiar.

Una vez ajustado el GLSM, este se contrasta con la hipótesis nula usando uno a uno los diez kernels, para descartar así que τ_1 sea 0. En este tipo de modelos, la estimación de los parámetros y el contraste de H_0 es muy costoso, ya que la verosimilitud de un GLSM es una integral n-dimensional que no se puede resolver analíticamente. Para resolverla y permitir una estimación e inferencia escalables utilizando un GLSM, se ha desarrollado un algoritmo cuyo enfoque está basado en la pseudo-quasiverosimilitud, PQL. El desarrollo de este algoritmo se muestra en el material suplementario del artículo S. Sun et al., 2020. Con las estimaciones de los parámetros obtenidas a través de este algoritmo, se calcula un valor p para cada uno de los diez kernels usando el método de Satterthwaite a partir de los *scores* calculados, los cuales se distribuyen como una mixtura de χ^2. A continuación, se combinan estos diez valores p utilizando la regla de combinación de valores p de Cauchy, la cual se desarrolla en el anexo 2. De esta manera no se pretende encontrar un efecto espacial de una función kernel en concreto, sino si existe alguna combinación de efecto espacial de todos ellos. Para ello, se convierten cada uno de ellos en un estadístico de Cauchy, después se suman, y se vuelve a convertir esa suma en un único valor p basado

en la distribución estándar de Cauchy. La regla de Cauchy aprovecha que la suma de variables distribuidas como Cauchy sigue también una Cauchy, sin importar si estas están o no correlacionadas. Además, esta regla permite combinar múltiples valores p correlacionados en uno solo sin perder control sobre los errores de tipo I. Después de obtener m valores p a través de los m genes, la tasa de falsos positivos, FDR, se controla mediante el procedimiento de Benjamini-Yekutieli (Benjamini y Yekutieli, 2001), el cual es efectivo bajo una dependencia arbitraria entre los genes.

El método descrito anteriormente se conoce como la versión Poisson de SPARK, pero también se ha desarrollado una versión gaussiana, la cual es muy útil a la hora de modelizar datos distribuidos normalmente. Destacar que ambos métodos están incluidos en el mismo paquete de R (Shiquan Sun et al., 2021), por lo que no será necesario ningún paquete adicional si decide usarse el SPARK de tipo gaussiano.

Por otro lado, con SPARK también se pueden dividir los *SE genes* en diferentes categorías. Para ello, primero se aplicará una transformación para estabilizar la varianza de los datos crudos, denominada transformación de Anscombe Svensson et al., 2018. Una vez hecha esta transformación, se obtienen la expresión génica relativa mediante el ajuste del recuento total de conteos a escala logarítmica. A continuación, se emplea un algoritmo de *cluster* de tipo jerárquico aglomerativo, para lo cual se ha utilizado el paquete **amap** (Lucas, 2022).

3.1.1. Versión gaussiana de SPARK.

En esta ocasión los datos a modelizar, $y_i(\boldsymbol{s_i})$, ya no son datos de conteo, sino que son datos normalizados. Esta normalización puede llevarse a cabo, mediante por ejemplo, la transformación logarítmica de los datos originales. Asumiendo que los datos normalizados de expresión siguen una normal,

$$y_i(\boldsymbol{s_i}) = \boldsymbol{x_i(s_i)}^T\boldsymbol{\beta} + b_i(\boldsymbol{s_i}) + \epsilon_i, \;\; i = 1, 2, \ldots, n, \tag{2}$$

donde $\boldsymbol{\beta}$, al igual que en el de tipo Poisson, representa un vector de k coeficientes, incluyendo el intercepto que representa la media del logaritmo del nivel de expresión del gen a través de las diferentes ubicaciones, junto con k-1 coeficientes para las covariables

correspondientes; ϵ_i es el error residual independiente e idénticamente distribuido como $N(0, \tau_2)$, con varianza τ_2, y $b_i(\boldsymbol{s_i})$ es un proceso gaussiano de media 0 que modeliza el patrón de correlación espacial entre las diferentes localizaciones,

$$\boldsymbol{b}(\boldsymbol{s}) = (b_1(\boldsymbol{s_1}), b_2(\boldsymbol{s_2}), \ldots, b_n(\boldsymbol{s_n}))^T \sim MVN(0, \tau_1 \boldsymbol{K}(\boldsymbol{s})),$$

donde la covarianza $\boldsymbol{K}(\boldsymbol{s})$ es una función kernel de las localizaciones espaciales $\boldsymbol{s} = (s_1, \ldots, s_n)^T$, siendo $\boldsymbol{K}(\boldsymbol{s_i}, \boldsymbol{s_j})$ el elemento ij; τ_1 es un factor de escala del kernel de la covarianza. En el modelo 2, la covarianza de los niveles de expresión normalizados $y_i(\boldsymbol{s})$ es $\boldsymbol{\Sigma} = \tau_1 \boldsymbol{K}(\boldsymbol{s}) + \tau_1 \boldsymbol{I}$, donde $\boldsymbol{I}$ es la matriz identidad de dimensión n. De nuevo, el interés se centra en contrastar la hipótesis nula $H_0 : \tau_1 = 0$. Esto se llevará a cabo utilizando los *scores*, estadísticos utilizados para evaluar la validez de la hipótesis nula. Dichos *scores* se calculan de la siguiente manera, ignorando de nuevo la notación de la localización espacial de todas las variables por simplicidad,

$$\boldsymbol{Q}(\boldsymbol{y}) = \boldsymbol{y^T S K S y},$$

donde $\boldsymbol{S} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X}$ es la matriz de proyección del subespacio ortogonal a las variables $\boldsymbol{X}$. Los *scores* se distribuyen como una mixtura de distribuciones χ^2 bajo la hipótesis nula,

$$\boldsymbol{Q}(\boldsymbol{y})/\tau_1 \sim \sum_{i=1}^{t} \phi_i \chi^2_{1i},$$

donde $\{\phi_i\}$ son los autovalores de $\boldsymbol{SKS}$; χ^2_{1i} es una variable aleatoria independiente e idénticamente distribuida con un grado de libertad, y t es el número de valores propios no nulos. Al igual que se hizo con el SPARK de tipo Poisson, se obtiene el valor p para cada uno de los kernel y se combinan todos juntos utilizando la regla de combinación de Cauchy.

Si se compara la versión Poisson con la gaussiana, esta última es mucho más eficiente computacionalmente, además de mucho más robusta ante una especificación errónea del modelo, por lo que podría ser más eficaz en ciertas aplicaciones.

3.2. SPARK-X.

En esta sección se desarrollará una variante de SPARK, nueva aproximación para la detección de genes espacialmente variables, que al igual que el método que se mostrará a continuación, SpatialDE, con menores requerimientos computacionales, tanto en tiempo de ejecución como memoria. Además, estos datos, normalmente se recogen en forma de matrices que tienen mucha dispersión y gran cantidad de 0s, 0s que se corresponden con el nivel de expresión de ese gen en un determinado *spot*. Este tipo de matrices podrían modelizarse con una Poisson con sobredispersión. Sin embargo, modelizar paramétricamente datos con tanta sobredispersión puede ser muy complicado. Además, modelizar directamente datos de este tipo con una distribución binomial negativa o una Poisson con sobredispersión puede desembocar en problemas de estabilidad del algoritmo, algo que puede llevar a la falta de convergencia en el 90 % de los genes. Por otro lado, modelizar este tipo de datos tan dispersos utilizando una distribución normal, como ocurre en la versión gaussiana de SPARK o en SpatialDE, tampoco es lo ideal, ya que esta aproximación paramétrica puede ocasionar pérdida de potencia y fallo del control del error tipo I en valores p pequeños, los cuales son esenciales a la hora de detectar $SE\ genes$.

SPARK-X realiza un test no paramétrico a conjuntos de datos grandes de manera eficiente. Este método está construido basándose en un marco robusto de pruebas de covarianza, el cual amplía para poder introducir las funciones kernel espaciales para el modelado espacial no paramétrico del recuento de datos muy dispersos en estudios transcriptómicos de gran tamaño. Para conseguir reducir tanto la complejidad computacional como la memoria RAM necesaria para realizar un análisis de expresión espacial, se han realizado algunas modificaciones algebraicas en el método con respecto a su antecesor SPARK. Como ya se mencionó anteriormente, tanto en SPARK como en SpatialDE, la complejidad aumentaba cúbicamente con respecto al número de localizaciones espaciales, sin embargo, con los cambios realizados, este incremento producido aumentará de forma lineal. Debido a su naturaleza no paramétrica, SPARK-X es un algoritmo estable y estadísticamente robusto independiente a la estructura original de los datos. Además, proporciona un error de tipo I calibrado y mejora la potencia del método, pudiendo utilizarse para bancos de datos provenientes de numerosas plataformas.

Al tratarse de una variante de SPARK la notación de este método es muy similar a la de SPARK. El vector de localizaciones espaciales para el *spot* i se denota como s_i, con $i \in (1, \ldots, n)$, y $\boldsymbol{S} = (\boldsymbol{S_1^T}, \ldots, \boldsymbol{S_n^T})^T$ hace referencia a la matriz de dimensiones $n \times d$. En este caso en particular en el que solo se tratará con 2 dimensiones, estas coordenadas espaciales varían a través del espacio bidimensional (d=2; $\boldsymbol{s_i} = (s_{i1}, s_{i2}) \in \mathbb{R}^2$). La medida de la expresión para el *spot* i se denota como $y_i(\boldsymbol{s_i})$ e $\boldsymbol{y} = (y_1(\boldsymbol{s_1}), \ldots, y_n(\boldsymbol{s_n}))$ hace referencia a la misma medida de la expresión, pero esta vez medida en los n *spots* del estudio. Para un correcto funcionamiento del método, se asume que tanto $\boldsymbol{y}$ como $\boldsymbol{S}$ han sido centradas y escaladas para tener media 0 y desviación típica 1. Estas modificaciones no tendrán ningún tipo de influencia sobre el control del error de tipo I, pero sí que pueden afectar a la potencia estadística del método. De nuevo, en esta parte del método lo que se busca es identificar si el nivel de expresión de un gen en concreto presenta o no algún tipo de patrón espacial, o lo que es lo mismo, se quiere testear si $\boldsymbol{y}$ es dependiente de las coordenadas espaciales $\boldsymbol{S}$. Para ello, el método se basa en un conjunto amplio de pruebas de covarianza, entre los que se incluyen el criterio de independencia de Hilbert-Schmidt (Gretton et al., 2007) y la prueba de covarianza de distancias (Székely et al., 2007). Estas pruebas se utilizan para realizar análisis SE de manera no paramétrica, por lo que al no asumir una distribución determinada de los datos, proporciona una mayor flexibilidad y aplicabilidad a diversas situaciones. Este método se construye sobre la siguiente intuición: si $\boldsymbol{y}$ es independiente de $\boldsymbol{S}$, entonces la distancia entre dos localizaciones i y j también sería independiente de la diferencia de expresión génica entre las dos localizaciones. Es por eso que se pueden construir para cada uno de los *spots* dos matrices, una basada en el nivel de expresión y otra basada en la información espacial. El objetivo de la creación de estas dos matrices es examinar si son más similares entre sí de lo que cabría esperar sólo por azar.

Técnicamente, la matriz de covarianza de los niveles de expresión se construye como una matriz $n \times n$, la cual se define como: $\boldsymbol{E} = \boldsymbol{y}(\boldsymbol{y^T y})^{-1}\boldsymbol{y^T}$. Por otro lado, la matriz $n \times n$ de covarianzas de las distancias para todos los *spots* basados en las localizaciones espaciales se construye de la siguiente manera: $\boldsymbol{\Sigma} = \boldsymbol{S}(\boldsymbol{S^T S})^{-1}\boldsymbol{S^T}$. A ambas matrices se las denomina matrices de covarianzas, ya que ambas han sido generadas a partir de una función de covarianza de proyección, además ambas dos poseen las dos propiedades clave

de las matrices de covarianzas: son simétricas y semidefinidas positivas. La ya mencionada función de covarianza de proyección, ha sido ampliamente utilizada en multitud de aplicaciones en genética. Antes de empezar a utilizar ambas matrices, habrá que centrarlas como $\boldsymbol{E_C} = \boldsymbol{HEH}$ y $\boldsymbol{\Sigma_C} = \boldsymbol{H\Sigma H}$, donde $\boldsymbol{H} = (\boldsymbol{I} - \mathbf{1_n}\mathbf{1_n^t}/n)$ con $\boldsymbol{I}$ la matriz identidad $n \times n$ y $\mathbf{1_n}$ un vector de 1's de n componentes. Esta transformación no supondrá ningún tipo de alteración en los resultados al haber centrado las matrices $\boldsymbol{y}$ y $\boldsymbol{S}$ antes de construir estas matrices de covarianzas. Una vez llevado a cabo este procedimiento, se creará el siguiente estadístico,

$$T = traza(\boldsymbol{E}_C\boldsymbol{\Sigma}_C)/n.$$

Intuitivamente, cada elemento de cualquiera de las matrices de covarianza mide la similitud entre pares de localizaciones en términos de su desviación coordinada respecto a la media, es decir, cómo se desvía cada una de las localizaciones de sus respectivas medias. Cuando $\boldsymbol{y}$ y $\boldsymbol{S}$ son independientes la una de la otra, la medida de similaridad entre un par de localizaciones en función de la expresión génica, no estará correlada con la medida de similaridad del par de localizaciones en términos de distancia. En este caso, los valores del estadístico serán elevados, sucediendo lo contrario cuando $\boldsymbol{y}$ y $\boldsymbol{S}$ son dependientes. Formalmente, bajo la hipótesis nula de que $\boldsymbol{y}$ y $\boldsymbol{S}$ son independientes, el estadístico T asintóticamente sigue una mixtura de distribuciones chi-cuadrado

$$\frac{1}{n^2}\sum_{i,j}\lambda_{E,i}\lambda_{\Sigma,j}z_{ij}^2,$$

donde $\lambda_{E,i}$ representa el i-ésimo valor propio no nulo ordenado de la matriz $\boldsymbol{E}_C$; $\lambda_{\Sigma,j}$ es el j-ésimo valor propio no nulo ordenado de la matriz $\boldsymbol{\Sigma}_C$, y z_{ij}^2 son las variables independientes idénticamente distribuidas como χ_1^2. Un valor de T extremadamente grande, algo muy poco frecuente en la hipótesis nula, indica que hay evidencias en contra de H_0. Por lo que se puede calcular el valor p para medir la probabilidad de encontrar T igual o mayor que la observada en los datos basados en la distribución nula. El valor p para testear $H_0 : \tau_1 = 0$ se puede calcular utilizando el método exacto de *Davies'* (Davies, 1980).

Como ya se mencionó anteriormente, se han realizado algunas modificaciones alge-

braicas para optimizar el método. En primer lugar, destacar que los valores propios de $\boldsymbol{E}_C$ y $\boldsymbol{\Sigma}_C$ son equivalentes a los valores propios de $(\boldsymbol{y^T y})^{-1}\boldsymbol{y^T H y}$ y $(\boldsymbol{S^T S})^{-1}\boldsymbol{S^T H S}$ respectivamente. El coste computacional de la obtención de estos autovalores basado en las fórmulas anteriores tan solo es $O(nd^2)$. En segundo lugar,

$$Tr(\boldsymbol{E}_C\boldsymbol{\Sigma}_C) = Tr\left(\boldsymbol{y}(\boldsymbol{y^T y})^{-1}\boldsymbol{y^T H\Sigma H}\right) = (\boldsymbol{y^T y})^{-1}Tr(\boldsymbol{y^T H\Sigma H y}).$$

Por lo que nunca será necesario calcular $\boldsymbol{E}, \boldsymbol{\Sigma}$ ni sus versiones centradas $\boldsymbol{E}_C, \boldsymbol{\Sigma}_C$ utilizando el algoritmo. En su lugar, se calcularán $\boldsymbol{y^T y}$, $\boldsymbol{y^T H y}$, $\boldsymbol{S^T S}$,$\boldsymbol{S^T H S}$ y $\boldsymbol{y^T H\Sigma H y}$, los cuales, como mucho requieren una complejidad computacional de $O(nd^2)$ y de $O(nd)$ de memoria. Estas cantidades pueden calcularse eficientemente de la siguiente manera:

$$\boldsymbol{y^T H y} = (\boldsymbol{H y})^{\boldsymbol{T}}(\boldsymbol{H y}) = (\boldsymbol{y} - \bar{y})^T(\boldsymbol{y} - \bar{y}),$$

$$\boldsymbol{S^T H S} = (\boldsymbol{H S})^{\boldsymbol{T}}(\boldsymbol{H S}) = \left(\boldsymbol{S} - \frac{\boldsymbol{1 1^T}}{\boldsymbol{n}}\boldsymbol{S}\right)^{\boldsymbol{T}}\left(\boldsymbol{S} - \frac{\boldsymbol{1 1^T}}{\boldsymbol{n}}\boldsymbol{S}\right),$$

$$\boldsymbol{y^T H\Sigma H y} = \boldsymbol{y^T H S}(\boldsymbol{S^T S})^{-1}\boldsymbol{S^T H y} = (\boldsymbol{y} - \bar{y})^{\boldsymbol{T}}\boldsymbol{S}(\boldsymbol{S^T S})^{-1}\boldsymbol{S^T}(\boldsymbol{y} - \bar{y}).$$

Finalmente, las cantidades en las que $\boldsymbol{S}$ interviene, tanto el cálculo de $\boldsymbol{S^T S}$ y $\boldsymbol{S^T H S}$ como la descomposición propia de $(\boldsymbol{S^T S})^{-1}\boldsymbol{S^T H S}$, solo necesitan llevarse a cabo una vez, y no necesitan ser recalculadas para cada uno de los genes. Sin embargo, en las que dependen de $\boldsymbol{y}$, incluidas $\boldsymbol{y^T y}$, $\boldsymbol{y^T H y}$ y $\boldsymbol{y^T H\Sigma H y}$, podrían variar entre genes, pero pueden calcularse eficientemente basándose en la dispersión de $\boldsymbol{y}$. Esto supondría un gran ahorro en el cálculo de estas tres cantidades cuando los datos son muy dispersos, ya que la complejidad de cálculo varía de manera lineal con respecto a los *spots* con valores distintos de 0. Por lo tanto, este método tendrá una complejidad computacional final de $O(nd^2 + pn'd)$ y necesita una memoria de $O(nd^2)$, donde p es el número de genes analizados y n' es el número de localizaciones espaciales con recuentos no nulos promediados a través de los genes.

La potencia estadística de este test de covarianza va a depender de cómo esté construida la matriz de covarianzas $\boldsymbol{\Sigma}$ y de cómo coincida con el verdadero patrón espacial que presenta el gen de interés. Aunque la construcción del kernel de proyección anterior permite alcanzar órdenes de magnitud de ganancias computacionales en comparación con

otros kernels, kernels gaussianos y periódicos utilizados en SPARK, es probable que no sea óptimo para detectar todos los posibles patrones de expresión encontrados en los datos reales. Para asegurar una identificación robusta de los *SE genes* a través de diferentes tipos de patrones espaciales, se han considerado diferentes transformaciones de las coordenadas espaciales $\boldsymbol{s_i}$, y la posterior construcción de diferentes matrices de covarianza de distancias. Concretamente, se han aplicado cinco transformaciones gaussianas a las coordenadas $\boldsymbol{s_i} = (s_{i1}, s_{i2})$ para conseguir cinco conjuntos de coordenadas transformadas $\boldsymbol{s'_i} = (s'_{i1}, s'_{i2})$, con $s'_{i1} = exp\left(\frac{-s_{i1}^2}{2\sigma_1^2}\right)$ como la primera componente transformada de las coordenadas, y $s'_{i2} = exp\left(\frac{-s_{i2}^2}{2\sigma_2^2}\right)$ como la segunda componente. En esta transformación se emplean diferentes tipos de parámetros de suavizado σ_1 y σ_2 para cada conjunto, para así cubrir un amplio rango de los posibles patrones de covarianza local. Además, se han aplicado cinco transformaciones de tipo coseno sobre $\boldsymbol{s_i}$, para conseguir otros cinco conjuntos de coordenadas transformadas con $s'_{i1} = cos\left(\frac{2\pi s_{i1}}{\phi_1}\right)$ como la coordenada x transformada, y $s'_{i2} = cos\left(\frac{2\pi s_{i2}}{\phi_2}\right)$ como la y transformada para cada uno de los conjuntos. También se han empleado diferentes parámetros de periodicidad ϕ_1 y ϕ_2, para cubrir algunos de los posibles patrones de periodicidad. Los parámetros de transformación σ_1, σ_2, ϕ_1 y ϕ_2 están predeterminados, usando los cuantiles 20 %, 40 %, 60 %, 80 % y 100 % de los valores absolutos de las coordenadas de los datos x e y. Usando los cuantiles empíricos para construir diferentes matrices de covarianzas, siguiendo las ideas propuestas en SPARK (S. Sun et al., 2020) y que se mostrarán en SpatialDE (Svensson et al., 2018). Utilizar estos cuantiles frente a fijar los valores de los parámetros a unos valores determinados, tiene la ventaja de que es invariante ante cualquier tipo de transformación en la escala de los datos originales, y además permite construir las matrices de covarianza de distancia de una manera que estas matrices dependan de los datos.

Cada $\boldsymbol{s'_i}$ se utiliza para construir una matriz de covarianzas, por lo que se obtendrán diez matrices, además de las ya existentes sin transformar. De manera intuitiva, el kernel que se ha construido utilizando las coordenadas sin transformar, es útil para detectar patrones de expresión lineales, los basados en la transformación coseno serán mejores para identificar patrones de tipo periódico y los kernels basados en la transformación gaussiana, identificarán más fácilmente aquellos patrones que presenten focos. Por lo tanto, combinar las matrices de covarianza originales de los de datos y las transformadas, puede permitir

identificar un rango más amplio de patrones espaciales. Para ello, de nuevo se calcula el valor p mediante el método de *Davies'* para cada matriz de covarianzas de las distancias. Una vez calculados esos once valores p, se combinan en uno solo siguiendo la norma de combinación de Cauchy.

Hasta ahora se ha descrito el método en ausencia de covariables. En el caso de que sí que las hubiese, habría que sustituir el vector de n componentes $\mathbf{1_n}$, en la matriz $\boldsymbol{H}$ por la matriz de covariables de dimensión $n \times q$ correspondiente. Esta matriz contiene una fila de unos que representa el intercepto, y las columnas restantes las medidas de las $q-1$ variables. Por lo que, la matriz centrada pasa a ser $\boldsymbol{H} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T})$. A pesar de este cambio el resto del proceso se mantiene invariante.

3.3. SpatialDE.

En esta sección se va a presentar un método llamado SpatialDE, un test estadístico que permite identificar *SE genes* con patrones espaciales en la variación de la expresión a partir de imágenes multiplexadas o datos de RNA-seq espacial. Además, este método también implementa la "histología de expresión automática"(AEH por sus siglas en inglés), técnica de clusterización basada en los patrones de expresión espacial de los genes dentro de un tejido. La información sobre este método se ha obtenido del artículo de Svensson et al., 2018.

Un primer paso crítico a la hora de analizar los bancos de datos, es identificar los genes que presentan una variación espacial de la expresión a lo largo del tejido. Sin embargo, los enfoques existentes para identificar genes altamente variables, como los utilizados para los datos convencionales de scRNA-seq, no miden la variabilidad espacial ya que ignoran dicha información. Como una posible alternativa, los investigadores han aplicado análisis de la varianza (ANOVA) para comprobar la diferencia en los niveles de expresión entre grupos de células, usando tanto anotaciones *a priori*, como el agrupamiento o clusterización de los *spots* usando métodos en los que se incorpora información espacial. Sin embargo, estos métodos solo detectan las variaciones que captan las diferencias entre grupos discretos.

Por todos los inconvenientes mostrados anteriormente, se ha creado SpatialDE. Este

método se construye basándose en un proceso gaussiano de regresión, un tipo de modelo utilizado en geoestadística. SpatialDE descompone la variabilidad de la expresión en dos componentes, una espacial y otra no espacial, utilizando dos efectos aleatorios: un término de varianza espacial que parametriza como la covarianza entre las expresiones génicas varía según la distancia entre los *spots* en el espacio, y otro término que modeliza la variabilidad no espacial. El ratio de la variabilidad explicada por estas dos componentes cuantifica la fracción de varianza espacial (FSV). Los *SE genes* se pueden identificar comparando un modelo con estos dos términos y otro que no contenga la componente espacial.

La interpretación de los parámetros del modelo ajustado nos permite comprender mejor la función espacial subyacente, como por ejemplo su nivel de resolución. Además, como ya se ha mencionado, SpatialDE proporciona un método de clusterización espacial dentro del mismo marco de procesos gaussianos, el cual identifica conjuntos de genes que tienen patrones de expresión espacial distintos. Esto proporciona un medio para realizar AEH, que relaciona la estructura del tejido y la composición de los tipos celulares a partir de los patrones de expresión de los genes marcadores.

SpatialDE también se puede aplicar a datos temporales para identificar genes que tienen una expresión dinámica en el tiempo. Aunque ya existen métodos que lo hacen, normalmente suelen ser más costosos a nivel computacional. Destacar también, que SpatialDE se puede aplicar a datos tridimensionales.

Este método generaliza enfoques previos para detectar expresiones génicas con patrones temporales y periódicos en series temporales. Aunque biológicamente es importante, la identificación de patrones periódicos tiene limitaciones técnicas, especialmente en casos límite, regiones del borde del tejido de estudio, donde el ruido puede enmascarar la significatividad estadística para patrones visualmente similares.

SpatialDE modeliza los perfiles de expresión génica $\boldsymbol{y} = (y_1, \ldots, y_N)$ para un gen dado a través de las coordenadas espaciales $\boldsymbol{X} = (x_1, x_2)$, siendo N el número de *spots*, usando una normal multivariante de la forma

$$P(\boldsymbol{y}|\boldsymbol{\mu}, \sigma_s^2, \delta, \boldsymbol{\Sigma}) = N(\boldsymbol{y}|\boldsymbol{\mu} \cdot \mathbf{1}, \sigma_s^2 \cdot (\boldsymbol{\Sigma} + \delta \cdot \boldsymbol{I})). \quad (3)$$

El efecto fijo $\mu \cdot \mathbf{1}$ tiene en cuenta el nivel medio de expresión y $\boldsymbol{\Sigma}$ denota la matriz espacial de covarianzas definida a partir de las coordenadas de entrada de los pares de celdas. SpatialDE utiliza la denominada función de covarianza exponencial al cuadrado para definir $\boldsymbol{\Sigma}$,

$$\Sigma_{i,j} = k(x_i, x_j) = exp\left(-\frac{|x_i - x_j|^2}{2 \cdot l^2}\right), \quad (4)$$

en el que la covarianza entre pares de *spots* i y j se modela para que decaiga exponencialmente con la distancia al cuadrado entre ellas. El hiperparámetro l, también conocido como el nivel de resolución característico (*characteristic length scale*) determina la rapidez con la que decae la covarianza en función de la distancia.

El segundo término de covarianza $\delta \cdot \boldsymbol{I}$ tiene en cuenta la variación independiente no espacial en la expresión génica, donde el ratio $FSV = \frac{1}{1+\delta}$ puede interpretarse como la fracción de varianza de expresión atribuible a efectos espaciales. Los parámetros del modelo se ajustan maximizando la log-verosimilitud marginal,

$$LL = -\frac{1}{2} \cdot N \cdot log(2 \cdot \pi) - \frac{1}{2} \cdot log(|\sigma_s^2 \cdot [\boldsymbol{\Sigma} + \delta \cdot \boldsymbol{I}]|) - \frac{1}{2} \cdot (\boldsymbol{y} - \mu \cdot \mathbf{1})^T \cdot (\sigma_s^2 \cdot [\boldsymbol{\Sigma} + \delta \cdot \boldsymbol{I}])^{-1} \cdot (\boldsymbol{y} - \mu \cdot \mathbf{1}). \quad (5)$$

Este problema de optimización proporciona soluciones de forma cerrada para los parámetros μ y σ_s, para determinados valores de los parámetros δ. La optimización basada en el gradiente se utiliza para determinar δ, y el hiperparámetro l se determina mediante la búsqueda en la cuadrícula. Los métodos *naive* para evaluar la probabilidad marginal en la ecuación (3) aumentan cúbicamente respecto al número de localizaciones espaciales, lo que impide su aplicación a conjuntos de datos más grandes.

Para estimar la significatividad estadística, la verosimilitud del modelo SpatialDE ajustado se compara con la verosimilitud de un modelo que corresponde a la hipótesis nula de ausencia de covarianza espacial,

$$P(\boldsymbol{y}|\mu, \sigma^2) = N(\mu \cdot \mathbf{1}, \sigma^2 \cdot \boldsymbol{I}) \quad (6)$$

A continuación, los valores p se estiman analíticamente a partir de la transformación de la distribución χ^2 con un grado de libertad. A menos que se indique lo contrario, se utiliza el método del valor q para ajustar las pruebas múltiples, controlando así la tasa de falsos positivos (FDR).

Tras la prueba de significación, los patrones de covarianza espacial identificados pueden investigarse más a fondo mediante comparaciones de modelos con funciones de covarianza alternativas. Además de la covarianza exponencial al cuadrado, SpatialDE implementa funciones de covarianza que asumen tendencias lineales, así como patrones periódicos de variación de la expresión génica, que se comparan utilizando el criterio de información bayesiano,

$$BIC = log(N) \cdot M - 2 \cdot LL. \tag{7}$$

Aquí M denota el número de hiperparámetros de un modelo dado, N el número de *spots*, y LL es la verosimilitud marginal logarítmica de los datos (5).

Para agrupar espacialmente los genes que tienen patrones espaciales de expresión similares, SpatialDE implementa un modelo de clusterización basado en el mismo proceso gaussiano que se utilizó para testear los genes espacialmente variables (3). Siendo $\boldsymbol{Y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_G})$ la matriz de expresión de G genes espacialmente variables en cada localización espacial (ahora cada $\boldsymbol{y_g}$ es un vector de N observaciones), $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ es la matriz de K patrones subyacentes de modo que el vector $\boldsymbol{\mu_k}$ representa el patrón k. Además, siendo $\boldsymbol{Z}$ una matriz indicadora binaria que asigna el gen g al patrón k si $z_{g,k} = 1$. Entonces el modelo completo a través de todos los genes quedaría así,

$$P(\boldsymbol{Y}, \mu, Z, \sigma_e^2, \boldsymbol{\Sigma}) = P(\boldsymbol{Y}|\mu, Z, \sigma_e^2) \cdot P(\mu|\boldsymbol{\Sigma}) \cdot P(Z),$$

$$P(\boldsymbol{Y}|\mu, Z, \sigma_e^2) = \prod_{k=1}^{K} \prod_{g=1}^{G} N(\boldsymbol{y_g}|\mu_k, \sigma_e^2)^{z_{g,k}},$$

$$P(\boldsymbol{\mu}|\boldsymbol{\Sigma}) = \prod_{k=1}^{K} N(\boldsymbol{\mu_k}|0, \boldsymbol{\Sigma}),$$

$$P(Z) = \prod_{k=1}^{K} \prod_{g=1}^{G} \left(\frac{1}{K}\right)^{z_{g,k}}.$$

El parámetro σ_e^2 es el nivel de ruido del modelo, y $\mathbf{\Sigma}$ es la matriz de covarianza de las coordenadas espaciales. Este modelo puede considerarse como una extensión de la mixtura modelo gaussiano clásico, con la adición de una distribución espacial a priori en los centroides de los *clusters*. Este proceso se realiza utilizando inferencia bayesiana, en particular una aproximación variacional a la inferencia bayesiana que consiste en seleccionar una distribución a priori sencilla y mediante la similitud kulba liber, conseguir que sea lo más cercana posible a la distribución a posteriori. Las distribuciones a posteriori aproximadas para $\boldsymbol{\mu}$ y z se estiman usando la ya mencionada inferencia variacional (Bishop, 2006), mientras que el nivel de ruido σ_e^2 se estima maximizando el límite variacional inferior. El hiperparámetro de resolución, l, para la covarianza $\mathbf{\Sigma}$ lo especifica el usuario, ya que es el número de patrones ajustados, K. Dicha elección puede guiarse por la l ajustada en el test de significación de SpatialDE. Para más detalles sobre la inferencia y la derivación de las actualizaciones variacionales ver la información suplementaria del artículo de Svensson et al., 2018.

Después de la inferencia, los valores esperados a posteriori de $\boldsymbol{\mu}$ y $\boldsymbol{Z}$ de los parámetros pueden usarse para visualizar cualquier patrón histológico graficando $\boldsymbol{\mu_k}$ sobre la coordenada x. La asignación más probable de un gen a un patrón individual queda determinado por el valor más alto del vector $\boldsymbol{z}_g$, el cual corresponde a las probabilidades a posteriori de que un gen determinado pertenezca a cada uno de los patrones.

El modelo de SpatialDE está basado en la asunción de que el ruido residual se distribuye normalmente y que las observaciones entre los *spots*/células son independientes. Para poder alcanzar estos dos requisitos se han llevado a cabo dos pasos de normalización. Para satisfacer la primera condición, se utiliza una transformación estabilizadora de la varianza para datos que se distribuyen como una binomial negativa, conocida como transformación de *Anscombe* (Anscombe, 1948). En segundo lugar, destacar que normalmente el nivel de expresión de un gen está correlacionado con el recuento total en una célula o locali-

zación espacial. Para asegurar que SpatialDE captura la covarianza espacial para cada gen más alla del efecto de los recuentos totales, se eliminan los valores logarítmicos del recuento total de los valores de expresión transformados por Anscombe antes de ajustar los modelos espaciales. Para más detalles sobre estas transformaciones mirar el material suplementario del artículo Svensson et al., 2018

3.4. Clusterización.

Una vez se han identificado los genes espacialmente variables, se procederá a realizar una clusterización para identificar las diferencias existentes entre los grupos que se han formado con los genes detectados con cada uno de los dos métodos. Antes de comenzar con el agrupamiento propiamente dicho, se llevará a cabo una reducción en la dimensionalidad de los datos mediante un análisis de componentes principales. Se realiza este paso previo debido a la alta dimensionalidad que poseen los datos, para poder así comprimir la información en un espacio de dimensiones más reducido y mejorar la robustez del análisis. A continuación, cuando ya se han calculado esas componentes principales, en este caso se ha decidido quedarse con las 50 primeras, se procederá al análisis de clusterización. El algoritmo que se utilizará será una variante de los K vecinos más cercanos (KNN por sus siglas en inglés), este se denomina *Shared Nearest Neighbor* (SNN por sus siglas en inglés). Este algoritmo considera no solo la proximidad directa entre los puntos, sino también la cantidad de vecinos que dos puntos tienen en común. Esta similitud basada en los vecinos compartidos por dos puntos p y q se puede definir como:

$$SNN(p, q) = |N_k(p) \cap N_k(q)|$$

donde $N_k(p)$ y $N_k(q)$ son los conjuntos de los k-vecinos más cercanos de los puntos p y q respectivamente, y $|N_k(p) \cap N_k(q)|$ es el tamaño del conjunto de intersección de $N_k(p)$ y $N_k(q)$.

4. Resultados.

En esta sección se van a mostrar los resultados que se han obtenido tras el análisis con los métodos seleccionados, mostrando los diferentes grupos de genes que se han formado, además de un posible gen marcador. Antes de comenzar a exponer los resultados de forma gráfica, puede observarse en la tabla 3 el número de genes espacialmente variables que se han encontrado con cada uno de los métodos, junto con los que coinciden en todos ellos. Destacar que para la identificación de los genes espacialmente variables con SPARK, hubo que aplicar un filtro de calidad extra para eliminar los genes que estaban muy poco expresados ya que sino el método no convergía, y por lo tanto no era posible utilizarlo.

Muestra	**SPARK**	**SPARK-X**	**SpatialDE**	**Coincidentes**
Muestra 1	4644	30929	2119	1723
Muestra 2	5918	33900	2856	2674
Muestra 3	5419	31191	11699	3691
Muestra 4	3948	30848	1321	1182
Muestra 5	2092	24609	2386	1066
Muestra 6	2871	28060	2620	1732

Tabla 3: Número de genes espacialmente variables identificados con cada uno de los métodos en las diferentes muestras, siendo de pacientes con Alzheimer las muestras 2, 4 y 6 y de pacientes control, las muestras 1, 3 y 5.

Una vez que se han identificado los genes espacialmente variables en cada método, se han evaluado aquellos genes comunes en todos los métodos empleados. Finalmente, se ha observado, que 445 genes de los 36601 existentes, han sido catalogados cómo espacialmente variables por todos los métodos empleados las seis muestras.

A continuación, en la figura 6 se muestra como se distribuye a través de las 6 muestras uno de estos genes en los que los tres métodos han detectado variabilidad espacial, gen ENSG00000118785. Se puede observar cómo en las muestras en las que el paciente no estaba enfermo, muestras 1, 3 y 5, la expresión es menor que en las que pertenecen a los individuos que sí que paceden la enfermedad. Esto puede significar que este gen sea un gen marcador, es decir, un gen cuyos niveles de expresión se asocian al desarrollo de una

enfermedad, Alzheimer en este caso.

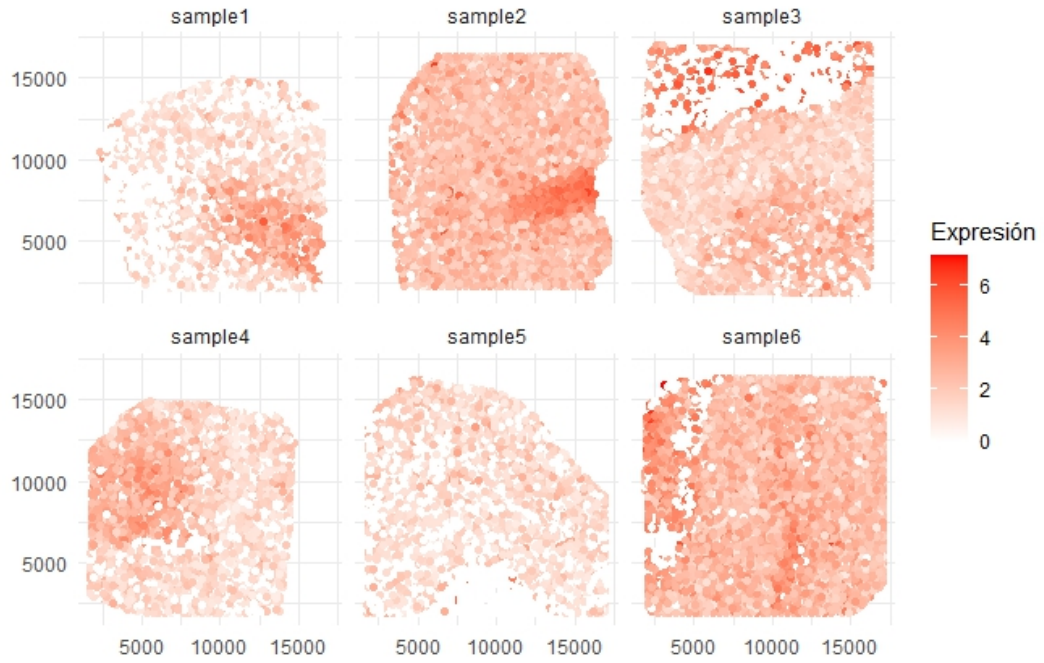
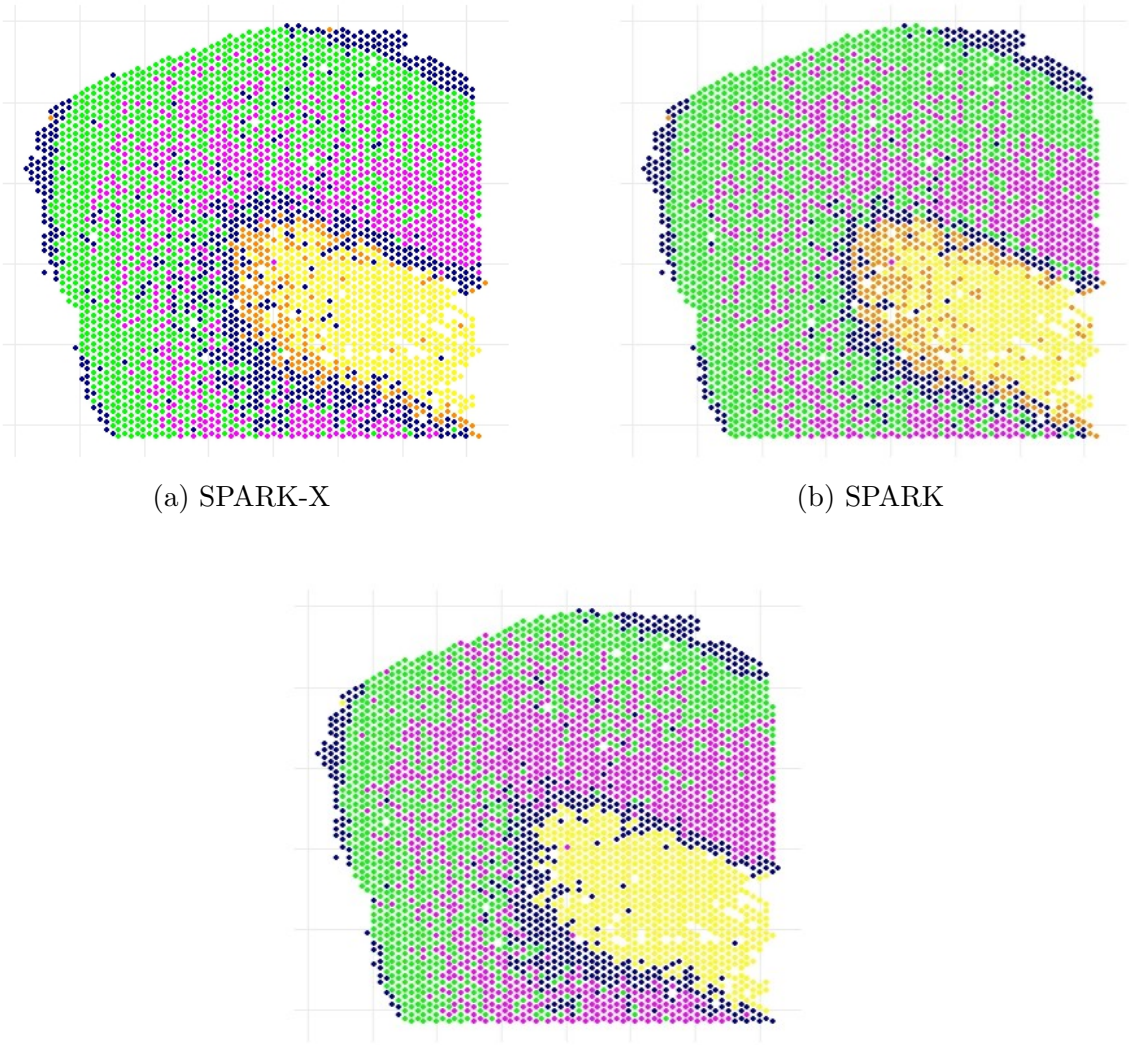


Figura 6: Expresión del gen ENSG00000118785 en cada muestra. Los ejes representan la posición horizontal y vertical de los *spots*.

Posteriormente, se mostrarán los grupos que se han formado con cada uno de los métodos en cada una de las muestras una vez se han identificado los genes espacialmente variables. El número de grupos se ha ido ajustando utilizando la metodología SNN, explicada previamente, en la que se han determinado el número de vecinos para obtener los *clusters* con mayor diferenciación.

Tal y como se puede observar en la figura 7, el número de *clusters* depende del método que se haya empleado. Mientras que con SPARK y SPARK-X, figuras 7b y 7a respectivamente, se han formado 5 grupos, con SpatialDE, figura 7c, se han formado 4. Se observa que el grupo que diferencia a SPARK y SPARK-X de SpatialDE, es el mismo, el que aparece coloreado en color naranja, entre los *clusters* azul y amarillo, aunque en SPARK-X el número de *spots* que conforman ese grupo es algo menor comparado con SPARK. Dicho esto, se puede ver que los métodos que obtienen una distribución en los *clusters* más parecida son SPARK-X y SPARK, figuras 7a y 7b respectivamente.



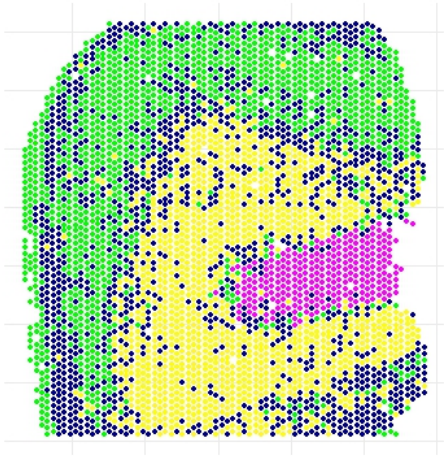
(a) SPARK-X

(b) SPARK

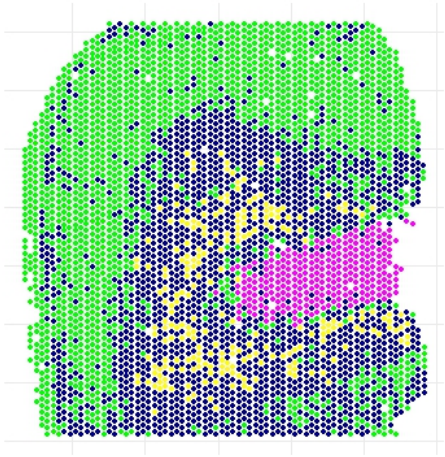
(c) SpatialDE

Figura 7: Distribución de *clusters* de los genes espacialmente variables para la muestra 1. Parte superior izquierda el método SPARK-X (a), parte superior derecha el método SPARK (b) y parte inferior, el método SpatialDE (c). Cada uno de los colores representa los diferentes grupos obtenidos. Los ejes representan la posición horizontal y vertical de los *spots*.

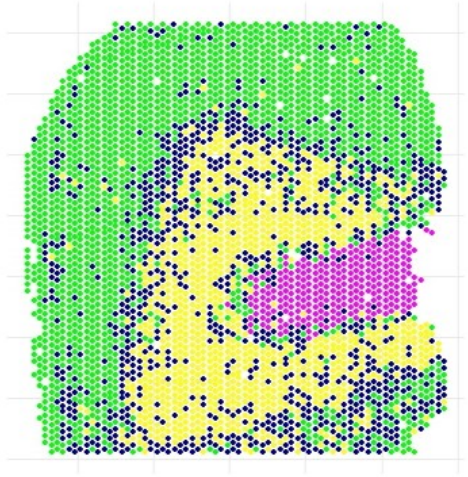
Con respecto al agrupamiento de la muestra 2 (figura 8), el número de *clusters* obtenidos es el mismo para todos los métodos, aunque su apariencia sí que tenga diferencias notables, ya que se puede apreciar fácilmente como en SPARK los *clusters* azul y amarillo no están prácticamente diferenciados entre sí, ya que ambos grupos están muy entrecruzados. Pudiendo deberse esto a que a pesar de pertenecer al mismo tipo de tejido celular, existan importantes diferencias en los niveles de expresión de los genes presentes en los *spots* de ambos *clusters*.



(a) SPARK-X



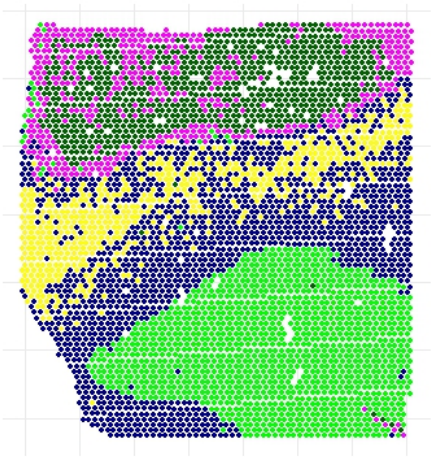
(b) SPARK



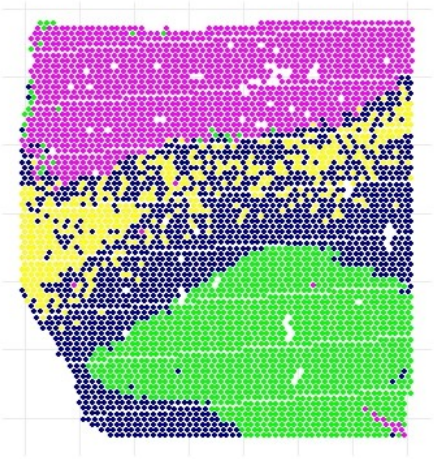
(c) SpatialDE

Figura 8: Distribución de *clusters* de los genes espacialmente variables para la muestra 2. Parte superior izquierda el método SPARK-X (a), parte superior derecha el método SPARK (b) y parte inferior, el método SpatialDE (c). Cada uno de los colores representa los diferentes grupos obtenidos. Los ejes representan la posición horizontal y vertical de los *spots*.

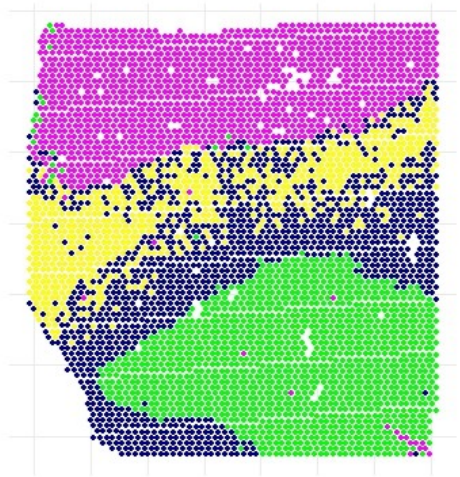
En la figura 9, se observa que en esta ocasión ha sucedido algo parecido a lo que ocurrió con la muestra 1. El número de grupos ha variado dependiendo del método utilizado. Mientras que con SPARK y SpatialDE (figuras 9b y 9c), los datos se han agrupado en 4 grupos, si se utilizan los genes obtenidos con SPARK-X (figura 9a), el número de grupos aumentaba a 5. Ese último *cluster* situado en el interior de otro grupo, puede deberse a lo que ya se comentó en la muestra anterior, *spots* que pertenecen al mismo tipo de tejido pero cuyos genes tienen niveles de expresión diferentes, creando así dos grupos diferenciados.



(a) SPARK-X



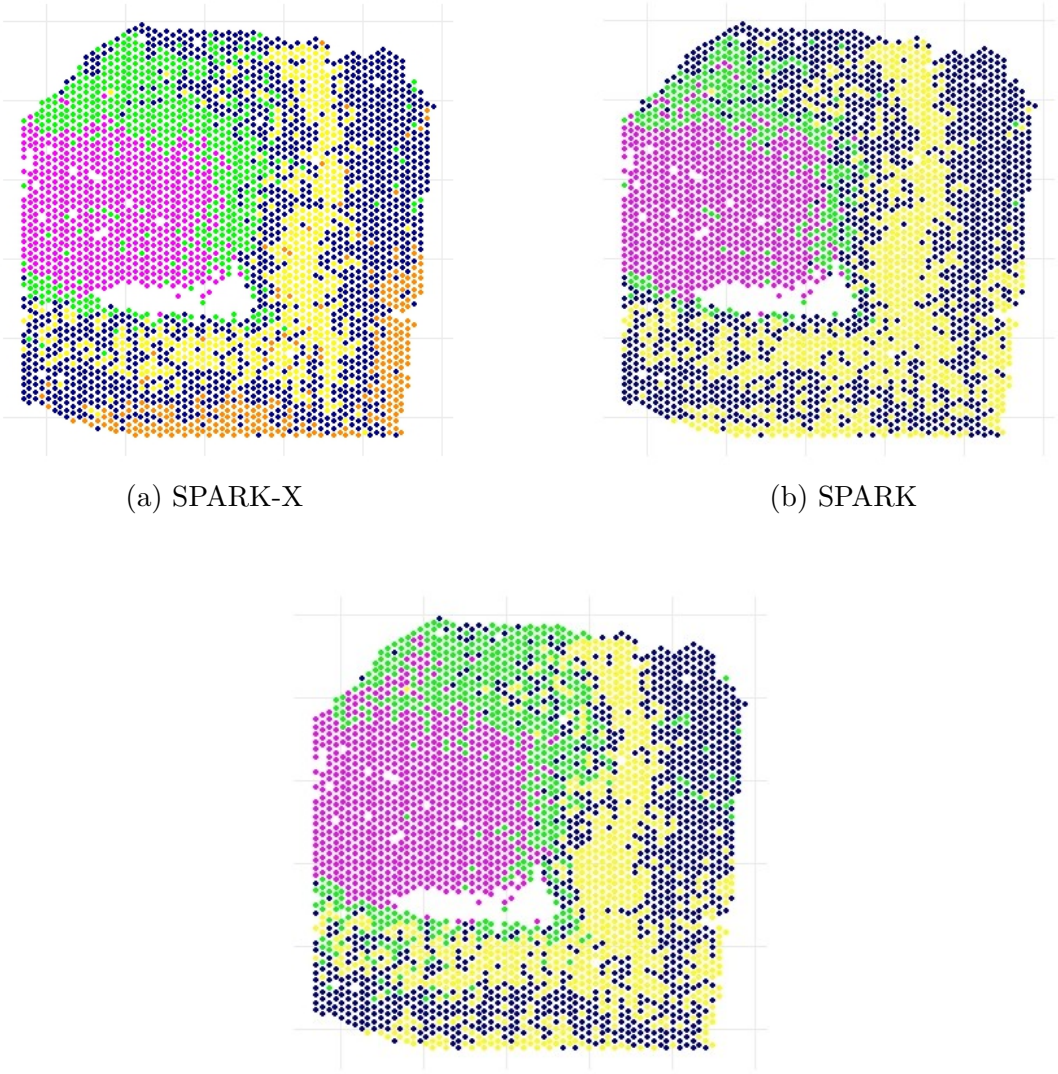
(b) SPARK



(c) SpatialDE

Figura 9: Distribución de *clusters* de los genes espacialmente variables para la muestra 3. Parte superior izquierda el método SPARK-X (a), parte superior derecha el método SPARK (b) y parte inferior, el método SpatialDE (c). Cada uno de los colores representa los diferentes grupos obtenidos. Los ejes representan la posición horizontal y vertical de los *spots*.

De nuevo, en la muestra 4 (figura 10), vuelve a suceder lo mismo que en la muestra 3. Al utilizar SPARK-X (figura 10a), se crea un nuevo grupo, en este caso el de color naranja. A pesar de que con SPARK-X se obtiene un grupo adicional, se observa que la distribución de los grupos que se han obtenido con los tres métodos son muy similares.



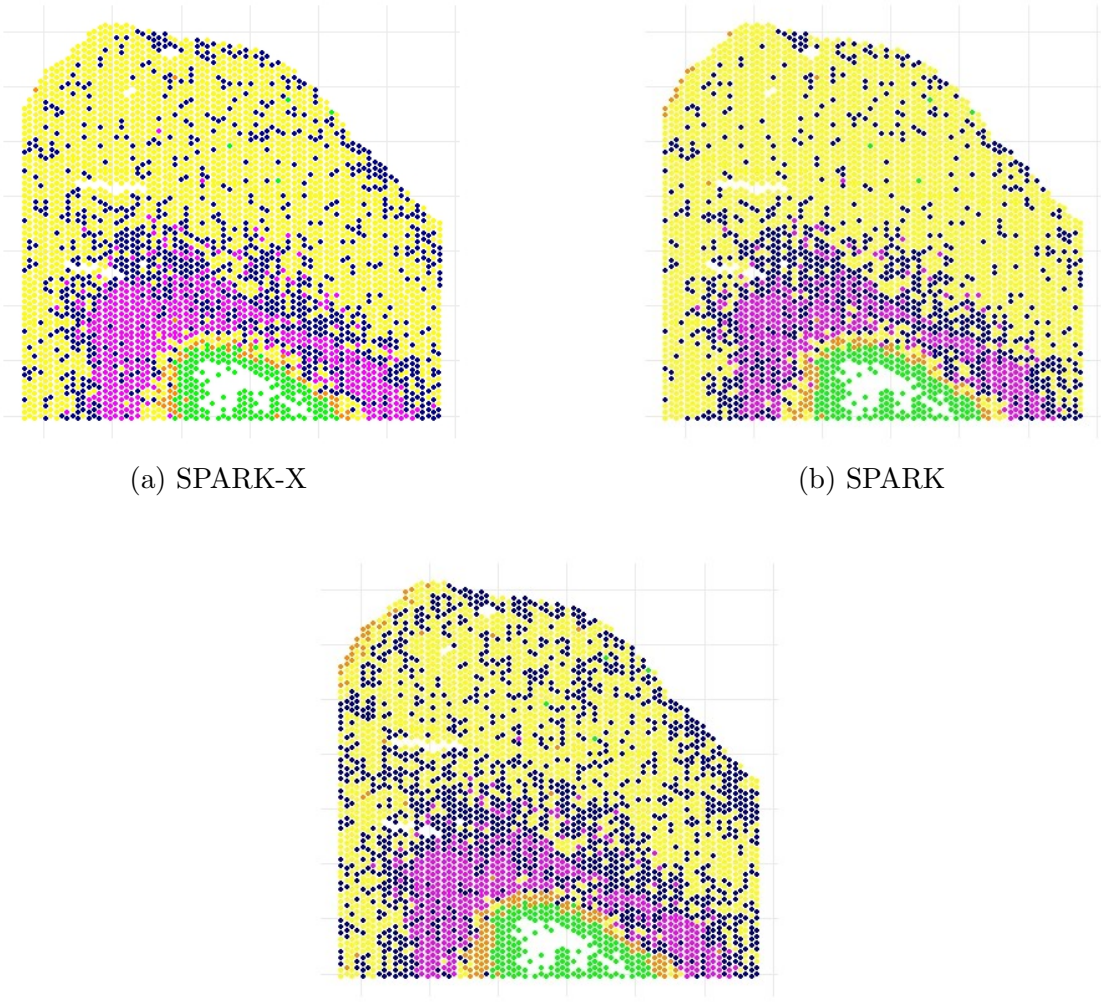
(a) SPARK-X

(b) SPARK

(c) SpatialDE

Figura 10: Distribución de *clusters* de los genes espacialmente variables para la muestra 4. Parte superior izquierda el método SPARK-X (a), parte superior derecha el método SPARK (b) y parte inferior, el método SpatialDE (c). Cada uno de los colores representa los diferentes grupos obtenidos. Los ejes representan la posición horizontal y vertical de los *spots*.

Con respecto a la muestra 5 (figura 11), aunque el número de grupos es el mismo en todos los métodos, la distribución de dichos grupos es más parecida entre los métodos SPARK-X y SPARK (figuras 11a y 11b respectivamente), ya que el grupo coloreado de color amarillo en estos dos grupos es mucho más grande y tiene menos *spots* del *cluster* azul que el de SpatialDE. En cuanto al resto de los grupos, se observa que son todos bastantes parecidos entre sí.



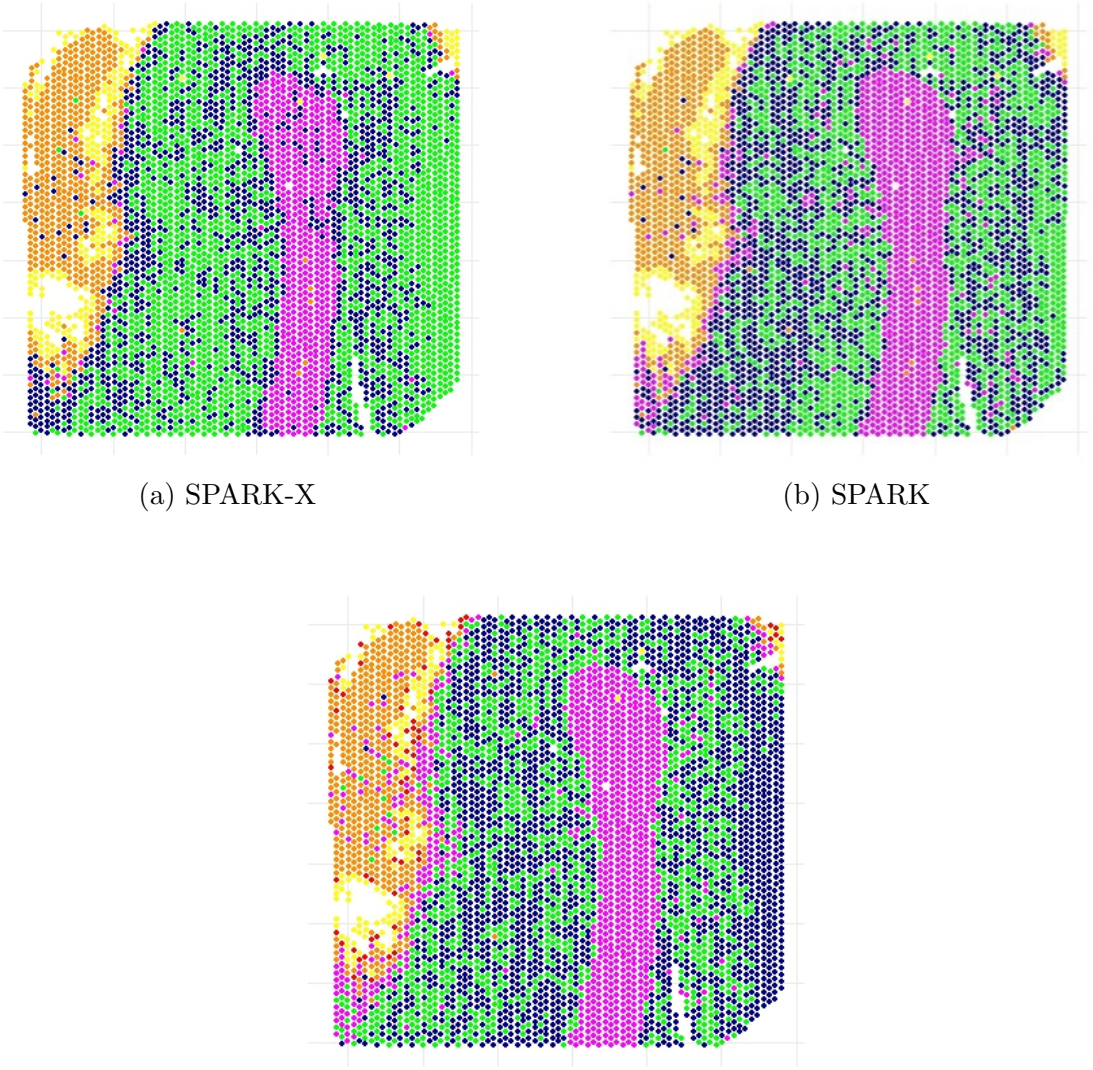
(a) SPARK-X

(b) SPARK

(c) SpatialDE

Figura 11: Distribución de *clusters* de los genes espacialmente variables para la muestra 5. Parte superior izquierda el método SPARK-X (a), parte superior derecha el método SPARK (b) y parte inferior, el método SpatialDE (c). Cada uno de los colores representa los diferentes grupos obtenidos. Los ejes representan la posición horizontal y vertical de los *spots*.

Por último, haciendo referencia a los *clusters* obtenidos en la última muestra (figura 12), se observa que en esta ocasión el método en el que se identifica un grupo más es en SpatialDE (figura 12c), contrario a lo que había sucedido hasta ahora. Este grupo está compuesto a penas de unos cuantos *spots*, pero al ser esta la mejor agrupación que se ha obtenido ajustando el número de vecinos, lleva a pensar que estos *spots* tienen algún tipo de diferencia con sus vecinos, a parte de sus niveles de expresión. Con respecto a la similitud en la distribución de estos *clusters*, se observa que el que tiene una agrupación algo más definida es el método SPARK-X, y que SpatialDE y SPARK son más parecidos entre sí.



(a) SPARK-X

(b) SPARK

(c) SpatialDE

Figura 12: Distribución de *clusters* de los genes espacialmente variables para la muestra 6. Parte superior izquierda el método SPARK-X (a), parte superior derecha el método SPARK (b) y parte inferior, el método SpatialDE (c). Cada uno de los colores representa los diferentes grupos obtenidos. Los ejes representan la posición horizontal y vertical de los *spots*.

5. Discusión

La identificación de genes espacialmente variables en tejidos y células es un paso esencial para los estudios de transcriptómica espacial. Dado el papel fundamental que desempeña en la interpretación de los datos, se han propuesto varios métodos para detectar este tipo de genes. Sin embargo, la falta de evaluaciones comparativas hace complicada la selección de un método adecuado.

Como ya se mencionó anteriormente, el objetivo del trabajo es comparar varias metodologías para la identificación de genes que muestran una mayor variabilidad en su expresión a lo largo de los diferentes dominios espaciales de un fragmento de tejido.

En la tabla 3, se puede ver claramente cómo el número de genes espacialmente variables detectados con los métodos SPARK y SpatialDE son muy inferiores a los detectados con SPARK-X. Esto podría deberse a que gracias al carácter no paramétrico de este último método, sea un método más flexible y por lo tanto sea capaz de adaptarse mejor a estos datos y captar así más genes espacialmente variables.

Aunque con SPARK-X se han detectado más genes espacialmente variables, las diferencias obtenidas en la distribución de los grupos formados en cada una de las muestras no son muy significativas, lo que implica que los otros dos métodos también son efectivos a la hora de detectar este tipo de genes. No obstante, la rapidez computacional y la necesidad de una cantidad de memoria muy reducida hacen que este método sea el más adecuado de los tres que se han evaluado.

En relación con los *clusters* obtenidos, a simple vista pueden no observarse diferencias entre los pacientes con Alzheimer y los sanos, ya que se ha hecho una clusterización en la que se ha elegido un número de *clusters* arbitrario. Con esta aproximación, quizás es difícil ver diferencias entre grupos. Hay que tener en cuenta que la *pipeline* de análisis continúa y serán estos análisis posteriores los que podrán proporcionar biomarcadores que discriminan entre los grupos de sujetos sanos y enfermos.

Haciendo referencia a las limitaciones de este trabajo, habría que destacar principalmente cuatro. En primer lugar, la dificultad a la hora de encontrar datos públicos adecuados para llevar a cabo este tipo de estudios, ya que muchos de los datos utilizados en los múltiples artículos consultados no eran de libre acceso, y por lo tanto, imposibles de utilizar. En segundo lugar, la necesidad de una gran capacidad a nivel tanto computacional como de memoria necesarias para poder llevar a cabo los análisis con los métodos de SPARK y SpatialDE, ya que mientras que con SPARK-X a penas eran necesarios unos minutos y unos cuantos megabytes, para poder ejecutar el código de SPARK y SpatialDE, fueron necesarios días y varios gigas de almacenamiento, algo que hace que este tipo de análisis sea muy complicado llevarlo a cabo utilizando ordenadores corrientes. En tercer lugar, como ya se mencionó, se intentó utilizar un cuarto método, Trendsceek, el cual por falta de actualizaciones acabó descartándose. Esto probablemente ha sido provocado por el desarrollo tan rápido que están experimentando estas técnicas, debido a sus grandes utilidades y aportaciones en estudios de este tipo. Por último, también cabría destacar que no todos los métodos de transcriptómica espacial se encuentran implementados en R, por lo que también ha sido complicado encontrar varios enfoques diferentes que además estuviesen implementados en este lenguaje.

6. Conclusiones.

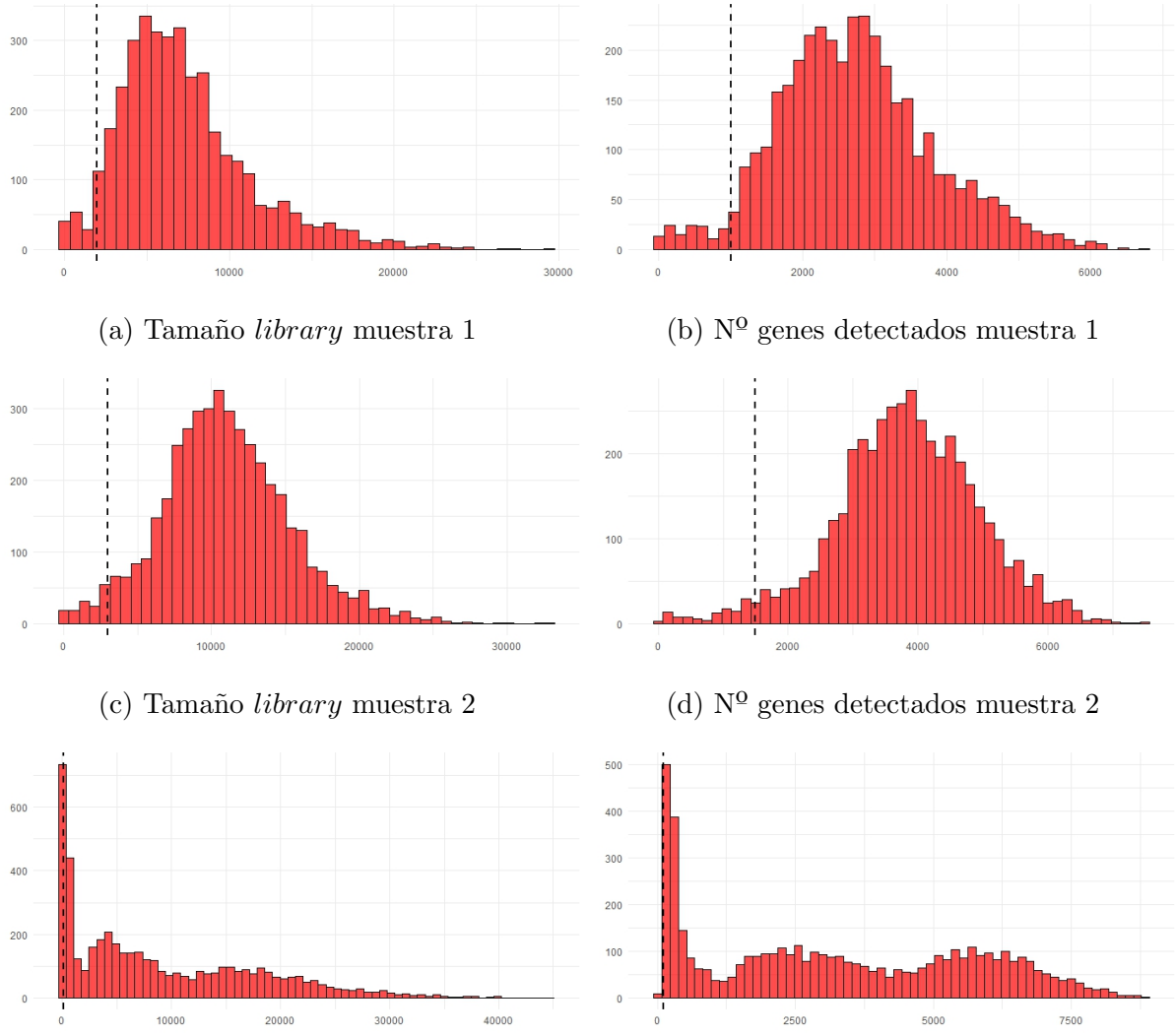
Las técnicas de transcriptómica espacial son una tecnología novedosa y prometedora, capaz de proporcionar resultados que hasta ahora, con las tecnologías de alto redimiento existentes eran impensables. A pesar de las grandes ventajas que proporciona a los diferentes estudios, al ser tan nueva aún tiene un gran margen de mejora y desarrollo, por lo que probablemente sea capaz de proporcionar incluso más información de gran valor.

A pesar de haber analizado tres métodos para la identificación de genes espacialmente variables, ya que hay más de una veintena de ellos, se ha podido comprobar que uno de ellos es muy eficiente, ya que tanto la memoria, como el tiempo de ejecución empleados han sido muy reducidos, todo ello sin obtener peores resultados. Estas características hacen que SPARK-X sea un método con el que es posible trabajar desde un ordenador corriente, sin tener que hacer uso de *clusters* o servidores externos como ha sucedido con los otros dos métodos.

Como líneas futuras o ampliación del estudio, podría plantearse la realización de un análisis de expresión diferencial con todos los genes espacialmente variables para detectar así patrones espaciales de expresión que diferenciasen a pacientes enfermos de los que no lo están. Por otro lado, se podría replicar el estudio con datos en los que la variable sexo estuviese presente, para corroborar con datos las diferencias que se sabe que existen a nivel biológico.

7. Anexo.

7.1. Histogramas *spots* eliminados.



(a) Tamaño *library* muestra 1

(b) Nº genes detectados muestra 1

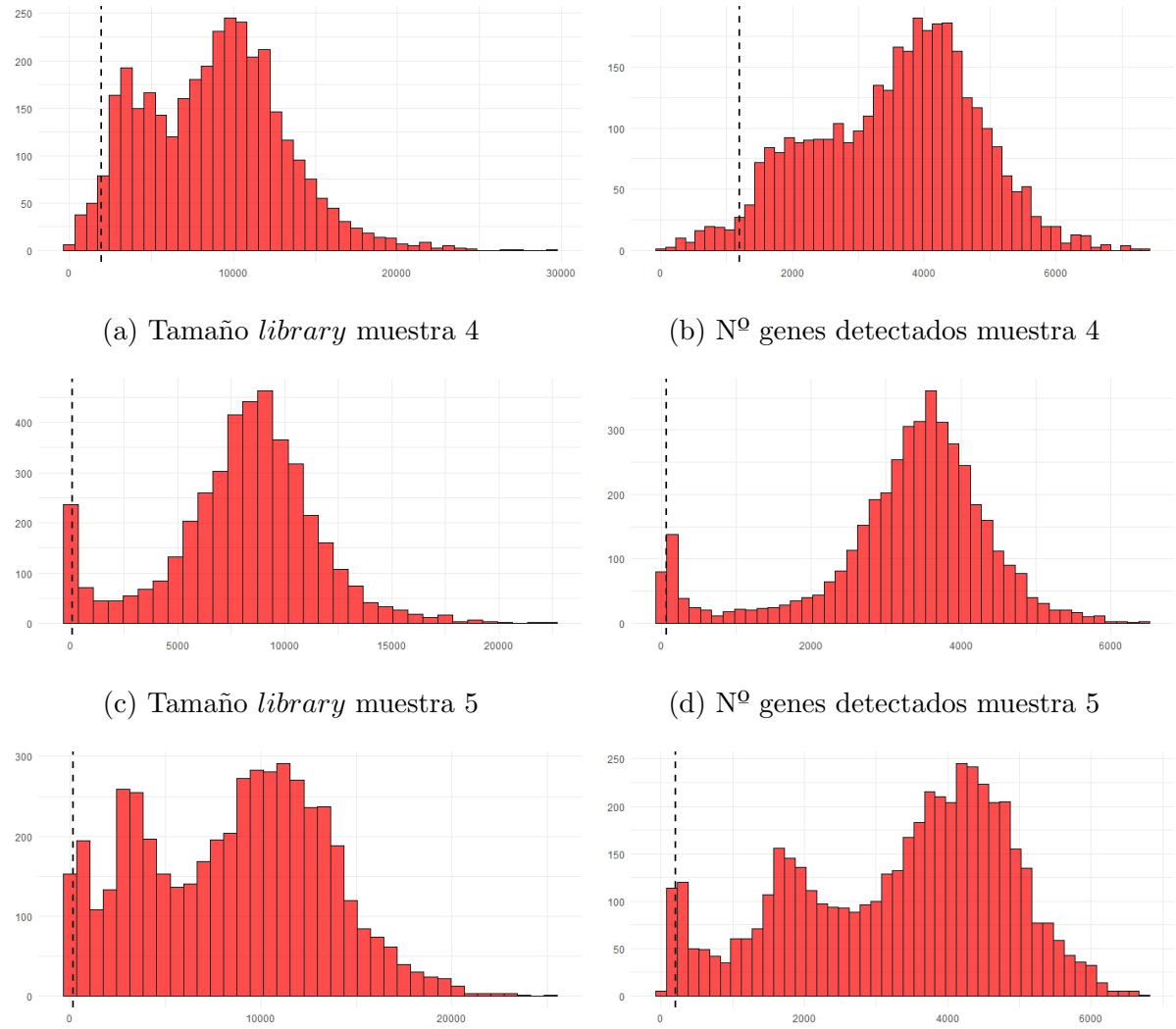
(c) Tamaño *library* muestra 2

(d) Nº genes detectados muestra 2

(e) Tamaño *library* muestra 3

(f) Nº genes detectados muestra 3

Figura 13: Spots eliminados en las muestras 1, 2 y 3.



(a) Tamaño *library* muestra 4

(b) Nº genes detectados muestra 4

(c) Tamaño *library* muestra 5

(d) Nº genes detectados muestra 5

(e) Tamaño *library* muestra 6

(f) Nº genes detectados muestra 6

Figura 14: Spots eliminados en las muestras 4, 5 y 6.

7.2. Score Test y regla de combinación de Cauchy.

Para cada uno de los diez kernels, se obtienen parámetros a través del algoritmo de PQL Svensson et al., 2018. Después se construye un estadístico de *score*

$$S_0 \equiv \frac{1}{2}\tilde{\boldsymbol{y}}^T \boldsymbol{PKP}\tilde{\boldsymbol{y}}.$$

Dado que bajo la hipótesis nula, $H_0 : \tau_1 = 0$, cada $\tilde{y}_i$ se distribuye como $N(x_i^T\beta, g'(\mu_i)+\tau_2)$, el estadístico de *score* también bajo la hipótesis nula sigue una mixtura de χ^2 $\sum_{i=1}^{t} \psi_i \chi^2_{1i}$, donde ψ_i son los autovalores de la matriz $\boldsymbol{PK}/2$, los cuales sirven como pesos para la mixtura de la distribución y t es el número de autovalores no nulos; mien-

tras χ^2_{1i} es una variable aleatoria de la distribución chi-cuadrado con un grado de libertad. Los valores p se calculan basándose en el *score* test usando el método de Satterthwaite. Este método se usa para probar la significancia estadística de los términos de efecto fijo. Esta mixtura de distribuciones chi-cuadrado de S_0, se aproxima por una distribución chi-cuadrado escalada, $k\chi^2_\nu$, utilizando el método de *matching moments*, donde k es el parámetro de escala y ν los grados de libertad. Si las media y la varianza de ambas distribuciones coinciden, obtenemos

$$\hat{k} = \frac{I_{\tau_1\tau_1}}{2e_s}, \;\; \hat{\nu} = \frac{2e_s^2}{I_{\tau_1\tau_1}},$$

donde $e_s = tr(\boldsymbol{PK})$ y $I_{\tau_1\tau_1} = tr(\boldsymbol{PKPK})/2$ son la media y la varianza de S_0 respectivamente.

Con $\hat{k}$ y $\hat{\nu}$ se obtiene un valor p para el estadístico *score* basado en la definición de valor p. A través del procedimiento anterior se obtienen diez valores p , uno para cada tipo de kernel. En este caso, también se empleará la regla de combinación de Cauchy, pero en este caso, lugar de sumar se realizará la media de dichos valores p. A la hora de calcular este valor p combinado, hay que tener en cuenta que debido a la precisión de la función de densidad acumulada de la distribución de Cauchy, cuando este tengan un valor inferior a $5{,}5e^{-17}$, tendrán un valor de 0. Finalmente, al igual que cuanod se trataba de la suma, se obtendrán m valores p, controlando la tasa de falsos positivos a través del procedimiento de Benjamini-Yekutieli, el cual es efectivo bajo la existencia de una dependencia arbitraria entre los genes.

7.3. Distribuciones de la expresión de varios genes.

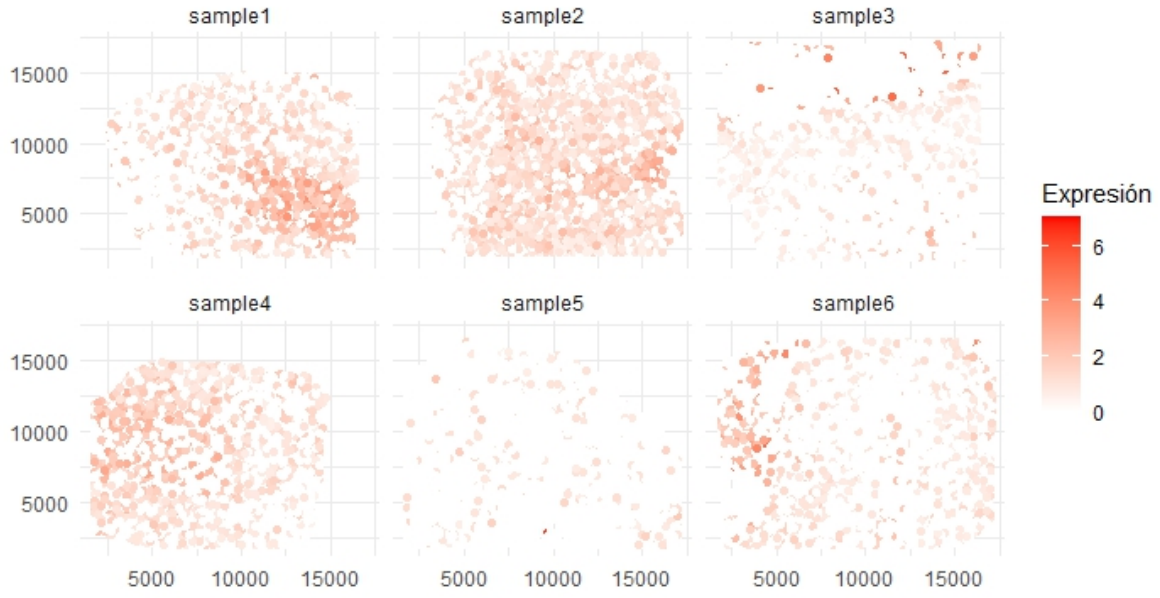


Figura 15: Expresión del gen ENSG00000150656 en cada muestra. Los ejes representan la posición horizontal y vertical de los *spots*.

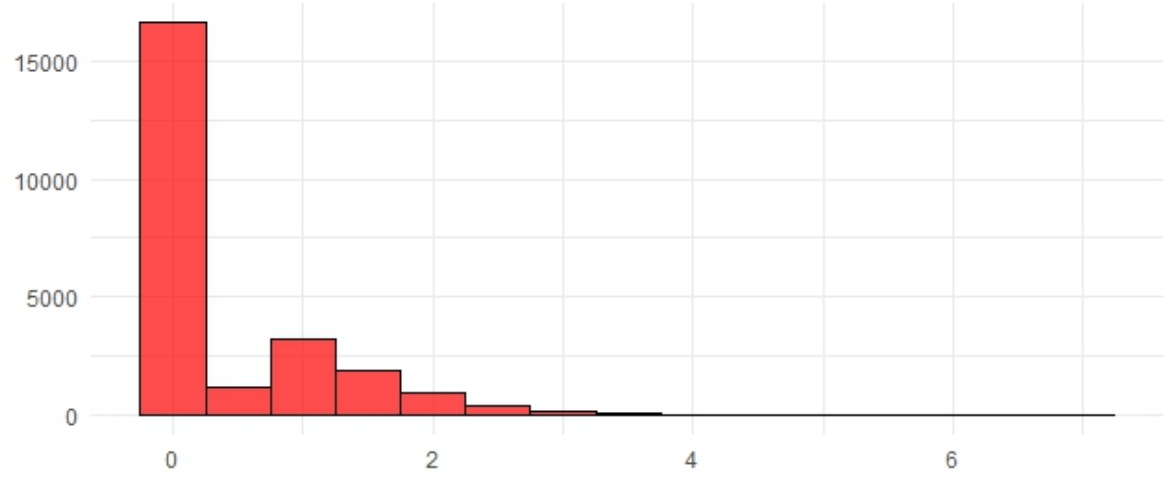


Figura 16: Distribución de la expresión del gen ENSG00000150656 en cada muestra.

Figura 17: Expresión del gen ENSG00000182578 en cada muestra. Los ejes representan la posición horizontal y vertical de los *spots*.

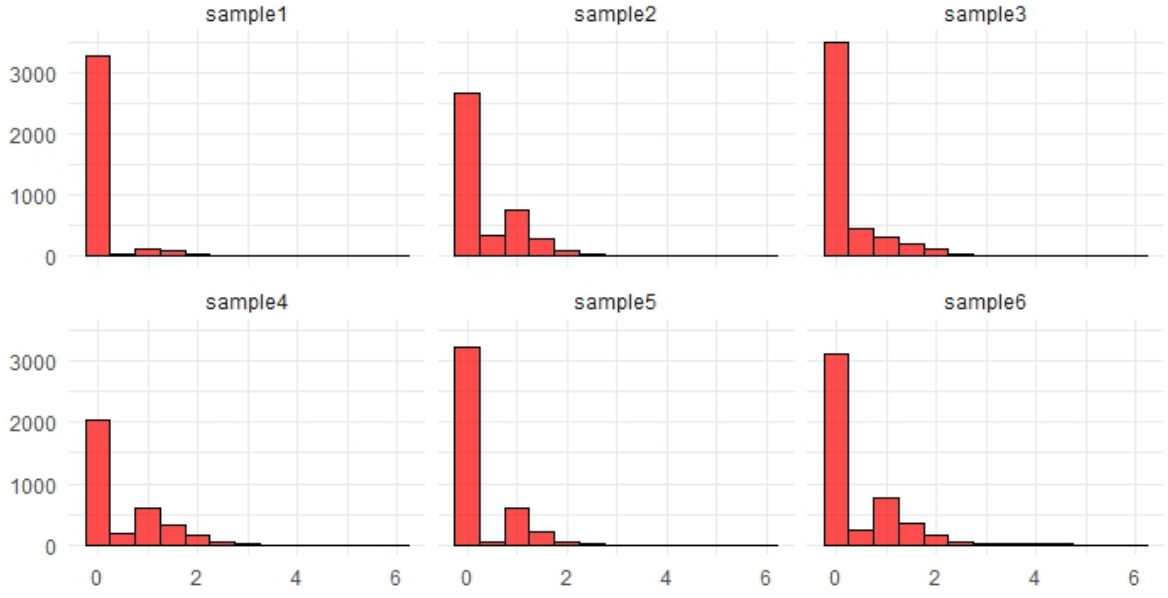


Figura 18: Distribución de la expresión del gen ENSG00000182578 en cada muestra.

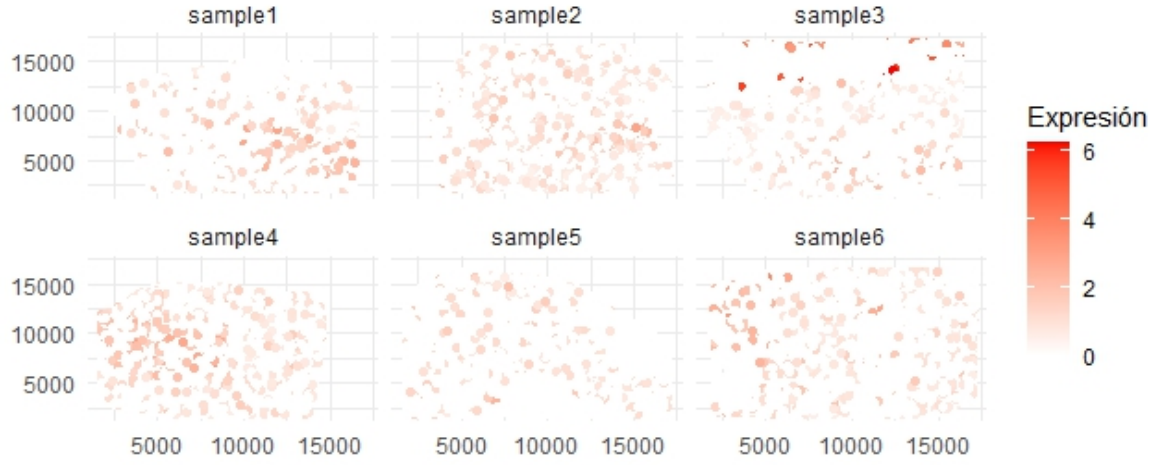


Figura 19: Expresión del gen ENSG00000224924 en cada muestra. Los ejes representan la posición horizontal y vertical de los *spots*.

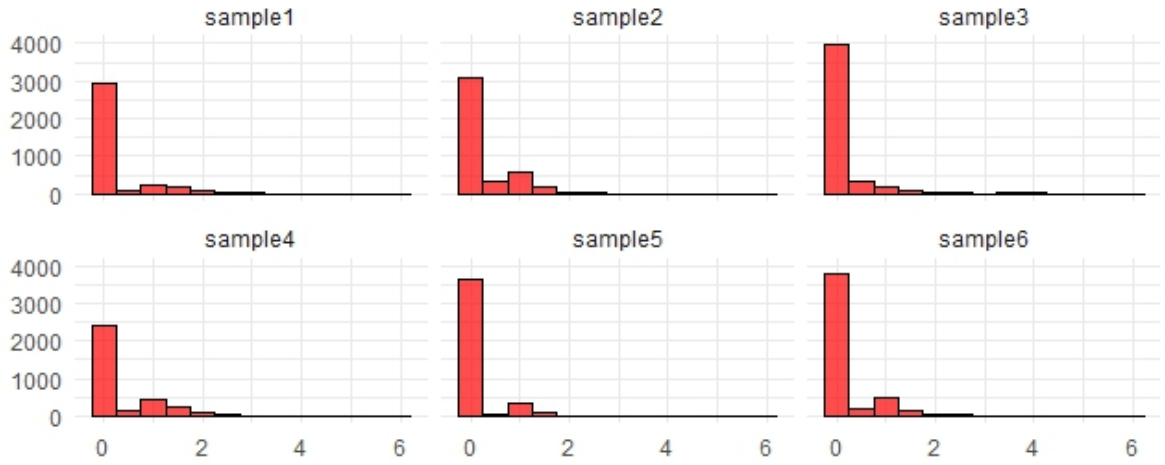


Figura 20: Distribución de la expresión del gen ENSG00000224924 en cada muestra.

Referencias

Anscombe, F. J. (dic. de 1948). «The Transformation of Poisson, Binomial and Negative-Binomial Data». En: *Biometrika* 35.3-4, págs. 246-254. DOI: 10.1093/biomet/35.3-4.246. URL: https://doi.org/10.1093/biomet/35.3-4.246.

Benjamini, Yoav y Daniel Yekutieli (2001). «The control of the false discovery rate in multiple testing under dependency». En: *The Annals of Statistics* 29.4, págs. 1165-1188.

Bishop, Christopher (ene. de 2006). «Pattern Recognition and Machine Learning». En: DOI: 10.1117/1.2819119.

Chen, S, Y Chang, L Li, D Acosta et al. (2022). «Spatially resolved transcriptomics reveals genes associated with the vulnerability of middle temporal gyrus in Alzheimer's disease». En: *Acta Neuropathologica Communications* 10.1, pág. 188. DOI: 10.1186/s40478-022-01462-3. URL: https://doi.org/10.1186/s40478-022-01462-3.

Chen, Tsai-Ying, Li You, Jose Angelito U. Hardillo y Miao-Ping Chien (2023). «Spatial Transcriptomic Technologies». En: *Cells* 12.16. ISSN: 2073-4409. DOI: 10.3390/cells12162042. URL: https://www.mdpi.com/2073-4409/12/16/2042.

Davies, Robert B (1980). «Algorithm AS 155: The distribution of a linear combination of χ^2 random variables». En: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 29.3, págs. 323-333.

Dlongwood (2024). *La transcriptómica espacial nombrada método del año por Nature Methods*. https://www.dlongwood.com/la-transcriptomica-espacial-nombrada-metodo-del-ano-por-nature-methods/. [Fecha de acceso: 25/05/2024].

Edsgärd, Daniel, Per Johnsson y Rickard Sandberg (2018a). «Identification of spatial expression trends in single-cell gene expression data». En: *Nature Methods* 15.5, págs. 339-342. DOI: 10.1038/nmeth.4634.

— (2018b). «Identification of spatial expression trends in single-cell gene expression data». En: *Nature Methods* 15.5, págs. 339-342. DOI: 10.1038/nmeth.4634. URL: https://doi.org/10.1038/nmeth.4634.

Frigolet, María E. y Ruth Gutiérrez-Aguilar (sep. de 2017). «Ciencias "ómicas", ¿cómo ayudan a las ciencias de la salud?» En: *Revista Digital Universitaria* 18.7. URL: https://revista.unam.mx/vol.18/num7/art54/index.html.

Galeano Niño, J. L., H. Wu, K. D. LaCourse, A. G. Kempchinsky, A. Baryiames, B. Barber, N. Futran, J. Houlton, C. Sather, E. Sicinska, A. Taylor, S. S. Minot, C. D. Johnston y S. Bullman (2022). «Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer». En: *Nature* 611.7937, págs. 810-817. DOI: 10.1038/s41586-022-05435-0. URL: https://doi.org/10.1038/s41586-022-05435-0.

Gretton, Arthur, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf y Alexander J Smola (2007). «A kernel statistical test of independence». En: *NIPS*. Vol. 20. Citeseer, págs. 585-592.

Lucas, Antoine (2022). *amap: Another Multidimensional Analysis Package*. R package version 0.8-19. URL: https://CRAN.R-project.org/package=amap.

National Cancer Institute (s.f.). *ADN*. Diccionario del Cáncer. Accedido: 26 de junio de 2024. URL: https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/adn.

— (s.f.[a]). *Hibridación fluorescente in situ*. Diccionario del Cáncer. Accedido: 26 de junio de 2024. URL: https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/hibridacion-fluorescente-in-situ.

— (s.f.[b]). *Nucleótido*. Diccionario del Cáncer. Accedido: 26 de junio de 2024. URL: https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/nucleotido.

— (s.f.[c]). *Resultados de búsqueda para ."RN mensajero"*. Sitio web del Instituto Nacional del Cáncer. Accedido: 26 de junio de 2024. URL: https://www.cancer.gov/espanol/buscar/resultados?swKeyword=ARN+mensajero.

— (s.f.[d]). *Resultados de búsqueda para ."RN"*. Sitio web del Instituto Nacional del Cáncer. Accedido: 26 de junio de 2024. URL: https://www.cancer.gov/espanol/buscar/resultados?swKeyword=ARN.

— (s.f.[e]). *Resultados de búsqueda para "histología"*. Sitio web del Instituto Nacional del Cáncer. Accedido: 26 de junio de 2024. URL: https://www.cancer.gov/espanol/buscar/resultados?swKeyword=histolog%C3%ADa.

— (s.f.[f]). *Resultados de búsqueda para "metabolito"*. Sitio web del Instituto Nacional del Cáncer. Accedido: 26 de junio de 2024. URL: https://www.cancer.gov/espanol/buscar/resultados?swKeyword=metabolito.

National Cancer Institute (s.f.[g]). *Resultados de búsqueda para .ºligonucleótido no codificante"*. Sitio web del Instituto Nacional del Cáncer. Accedido: 26 de junio de 2024. URL: https://www.cancer.gov/espanol/buscar/resultados?swKeyword=oligonucle%C3%B3tido+no+codificante.

— (s.f.[h]). *Transcripción*. Diccionario del Cáncer. Accedido: 26 de junio de 2024. URL: https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/transcripcion.

National Center for Biotechnology Information (NCBI) (s/f). *Gene Expression Omnibus (GEO)*. https://www.ncbi.nlm.nih.gov/gds. Accedido el 21 de junio de 2024.

Navarra, Clínica Universidad de (s.f.). *cDNA. Diccionario Médico*. Recuperado de https://www.cun.es/diccionario-medico/terminos/cdna.

Roura, Santiago (2024). *¿Por qué el porcentaje de ADN codificante es tan bajo en el genoma humano?* Consultado el 26 de junio de 2024. URL: https://metode.es/los-porques-de-metode/por-que-el-porcentaje-de-adn-codificante-es-tan-bajo-en-el-genoma-humano.html#:~:text=En%20primer%20lugar%2C%20se%20debe,no%20se%20traducen%20en%20prote%C3%ADnas..

Sun, S., J. Zhu y X. Zhou (2020). «Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies». En: *Nature Methods* 17, págs. 193-200. DOI: 10.1038/s41592-019-0701-7. URL: https://doi.org/10.1038/s41592-019-0701-7.

Sun, Shiquan, Jiaqiang Zhu y Xiang Zhou (2021). *SPARK: Spatial Pattern Recognition via Kernels*. R package version 1.1.1.

Svensson, Valentine, Sarah A Teichmann y Oliver Stegle (2018). «SpatialDE: identification of spatially variable genes». En: *Nature Methods* 15.5, págs. 343-346. DOI: 10.1038/nmeth.4636.

Székely, Gábor J, Maria L Rizzo y Nail K Bakirov (2007). «Measuring and testing dependence by correlation of distances». En: *Annals of Statistics* 35.6, págs. 2769-2794.

Transcriptómica Espacial 10x Genomics Visium (2024). https://www.bmkgene.com/es/10x-genomics-visium-spatial-transcriptome-product/. Accedido: 19-05-2024.

Wang, Y., B. Liu, G. Zhao, Y. Lee, A. Buzdin, X. Mu, J. Zhao, H. Chen y X. Li (2023). «Spatial transcriptomics: Technologies, applications and experimental considerations».

En: *Genomics* 115.5, pág. 110671. DOI: 10.1016/j.ygeno.2023.110671. URL: https://doi.org/10.1016/j.ygeno.2023.110671.

Wang, Ye, Bin Liu, Gexin Zhao, YooJin Lee, Anton Buzdin, Xiaofeng Mu, Joseph Zhao, Hong Chen y Xinmin Li (2023). «Spatial transcriptomics: Technologies, applications and experimental considerations». En: *Genomics* 115.5, pág. 110671. ISSN: 0888-7543. DOI: https://doi.org/10.1016/j.ygeno.2023.110671. URL: https://www.sciencedirect.com/science/article/pii/S0888754323001155.

Weber, Levi, Leonardo Collado-Torres y Stephanie C. Hicks (2024). *Best Practices for Spatial Transcriptomics Analysis with Bioconductor*. https://lmweber.org/BestPractice. Bioconductor.

Zhu, J., S. Sun y X. Zhou (2021). «SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies». En: *Genome Biology* 22, pág. 184. DOI: 10.1186/s13059-021-02404-0. URL: https://doi.org/10.1186/s13059-021-02404-0.