Universidad Internacional de La Rioja

Facultad de Ciencias de la Salud

Máster Universitario en Bioinformática

Desarrollo de un *pipeline* bioinformático para el análisis de datos de GeoMx™ *Digital Spatial Profiler*

Trabajo fin de Estudio presentado por:	Carme Delicado Mercader
Tipo de trabajo:	Trabajo de Fin de Máster
Área de estudio:	Bioinformática i Biología Computacional
Ciudad:	Barcelona
Directora interna:	María Belén Lledó Bosch
Directora externa:	Mireia Ferrer Almirall
Fecha:	11/02/2026

Resumen

Es una realidad que cada vez más, gracias al avance constante de la innovación y el desarrollo tecnológico, se abordan interrogantes que en la historia resultaban impensables. En el ámbito de la investigación biomédica, han surgido nuevas tecnologías capaces de cuantificar la expresión génica, proporcionando información clave sobre la función de los genes en distintos contextos biológicos. En este trabajo se desarrolla e implementa un *pipeline* de análisis basado en herramientas existentes para el procesamiento de datos generados por la plataforma GeoMx™ DSP, con el fin de convertir datos experimentales en información significativa que permita entender el funcionamiento biológico de interés. A partir de revisar diferentes enfoques, se identifica StandR como un *pipeline* potencialmente más robusto y eficiente en comparación con GeoMxTools. La revisión y evaluación de las metodologías disponibles resulta fundamental, ya que contribuye a una comprensión más profunda de los métodos actuales y favorece el avance del conocimiento científico.

Palabras clave: GeoMx™ DSP, Transcriptómica Espacial, StandR

Abstract

Advances in innovation and technological development have increasingly enabled the exploration of questions that were previously considered unthinkable. In the field of biomedical research, new technologies have emerged that allow the quantification of gene expression, providing key insights into gene function across different biological contexts. In this work, an analysis pipeline based on existing tools is developed and implemented for the processing of data generated by the GeoMx™ DSP platform, with the aim of transforming experimental data into meaningful information that enables the understanding of the biological processes of interest. By reviewing and comparing different analytical approaches, StandR is identified as a potentially more robust and efficient pipeline compared to GeoMxTools. The review and evaluation of the available methodologies are essential, as they contribute to a deeper understanding of current methods and promote the advancement of scientific knowledge.

Keywords: GeoMx™ DSP, Spatial Transcriptomics, StandR

Índice de contenidos

Contenido

Índice de figuras

Índice de tablas

1. INTRODUCCIÓN

1.1. JUSTIFICACIÓN DEL TEMA

La elección de este tema ha surgido de la oportunidad de realizar prácticas de máster en la Unidad de Estadística y Bioinformática (UEB) del Vall d'Hebron Institut de Recerca (VHIR). La UEB es una plataforma que presta servicios a los distintos grupos del VHIR en los análisis de datos biológicos.

En enero de 2026 se adquiere en la Unidad de Alta Tecnología (UAT) del VHIR, la plataforma GeoMx™ DSP creada por la casa comercial Nanostring. La tecnología va a estar disponible para que los investigadores la implementen en sus experimentos. Este hito, conlleva la necesidad de que se desarrolle un *pipeline* para analizar los datos provenientes de este equipo.

Poderse adentrar en el mundo del análisis bioinformático de la transcriptómica, en concreto de la transcriptómica espacial, siendo esta una de las tecnologías en auge en el momento actual, es un gran reto y una enorme oportunidad para crecer y desarrollar un buen camino profesional. Rodearse de profesionales como el equipo de la UEB, para la elaboración de este trabajo, es un componente clave en la motivación personal a la vez que un gran privilegio.

1.2. PROBLEMÁTICA ACTUAL

La transcriptómica es una de las ciencias "Ómicas" que ha transformado la biología molecular y la medicina moderna desde finales del siglo XX. Se caracteriza por estudiar la cantidad de moléculas de ARN de una célula o tejido en un contexto concreto (1,2). Técnicas como la secuenciación masiva (*Next Generation Sequencing*), desarrollada en 2005 (3), revolucionó el análisis transcriptómico al permitir el desarrollo de la secuenciación de RNA (*RNA-seq*). Inicialmente se aplicaba a muestras de tejido completo (*bulk RNA-seq*, 2008) y seguidamente, apareció la secuenciación de ARN de célula única (*single cell RNA sequencing*, 2009) (4,5). Los grandes avances han permitido analizar la actividad génica de forma global y con alta resolución celular. Esto facilita la identificación de cambios genéticos asociados a enfermedades o diferenciación celular, aunque estas metodologías carecen de la información crucial sobre la ubicación espacial de las células (6).

Para superar esta limitación, en el año 2016 surgió la transcriptómica espacial, una metodología innovadora que permite analizar la expresión génica en un tejido sin perder su contexto espacial (4). Si bien esta tecnología es muy prometedora, su reciente desarrollo y la complejidad de los datos que genera han impulsado la necesidad de crear nuevos métodos y herramientas bioinformáticas para procesarlos y poder aplicar la reconstrucción espacial (7).

A lo largo de este trabajo se realiza una revisión del estado actual de las tecnologías de secuenciación espacial, describiendo las características y analizando las diferencias de estas. Se desarrolla también, una revisión del flujo de trabajo empleado para el análisis de los datos de estas técnicas, con un foco específico en los generados por la tecnología GeoMx™ DSP. Finalmente, se describe un *pipeline* para el análisis de datos de GeoMx™ DSP, se implementa con un *dataset* de ejemplo y se elabora un archivo de trabajo reproducible, escalable y colaborativo.

1.3. OBJETIVOS

El objetivo general de este trabajo es desarrollar un *pipeline* para el análisis de datos de la tecnología GeoMx™ *Digital Spatial Profiler* (DSP). Se lleva a cabo mediante los siguientes objetivos específicos.

1.3.1. Objetivo 1: Revisión de las técnicas y las metodologías de análisis de datos de la transcriptómica espacial.

- Revisión de las tecnologías actuales de transcriptómica espacial, con énfasis en GeoMx™ DSP.
- Analizar las metodologías existentes para el procesamiento y el análisis de transcriptómica espacial, especialmente para la tecnología GeoMx™ DSP.

1.3.2. Objetivo 2: Desarrollo de un *pipeline* para el análisis de datos de transcriptómica espacial de la tecnología GeoMx™ DSP.

- Familiarización con *pipelines* existentes y disponibles para reproducir el análisis de datos de GeoMx™ DSP.
- Comparativa de los *pipelines* disponibles para evaluar ventajas y limitaciones.

- Determinación de un *pipeline* robusto, escalable y reproducible a partir de las mejores técnicas encontradas, mediante el lenguaje R.

1.3.3. Objetivo 3: Implementación del *pipeline* en un conjunto de datos, mediante una metodología de trabajo reproducible.

- Implementación del *pipeline* definido en un conjunto de datos de GeoMx™ DSP.
- Adaptación del *pipeline* para su funcionamiento de acuerdo con herramientas de trabajo reproducible.

2. MARCO TEÓRICO

2.1. ESTADO DEL ARTE

2.1.1. Introducción a las transcriptómicas

La transcriptómica es el estudio del conjunto completo de los transcritos presentes en una célula o tejido (8). Las funciones biológicas del organismo están codificadas en el material genético: el ADN se transcribe a ARN, y este se traduce a proteínas, que ejecutan las rutas metabólicas que permiten el funcionamiento celular. Existen diferentes tipos de ARN: el codificable, llamado el ARN mensajero (ARNm) que da lugar a proteínas, y el no codificable, que da lugar a moléculas reguladoras (8). Mediante las técnicas transcriptómicas obtenemos la comprensión de la cantidad de ARN que se está expresando en un momento determinado, lo que permite comprender cómo se regulan los genes y cómo estas expresiones diferenciales varían entre los tejidos, condiciones o instantes en el tiempo. Este análisis ayuda a definir las funciones de los genes y a comprender los mecanismos de diversas enfermedades (9).

Existen dos técnicas contemporáneas para la obtención experimental del transcriptoma: *Microarrays* y *RNA-seq*.

La técnica de *Microarrays* permite estudiar la expresión de miles de genes simultáneamente. Su funcionamiento se basa en depositar fragmentos de ADN (oligonucleótidos) de los genes diana, cuya expresión se desea medir, en una base de cristal llamada chip. Por otro lado, se extrae el ARN de la muestra a analizar, se retrotranscribe a ADN generando un ADN copia (cADN) y se marca con una molécula fluorescente. Este conjunto de moléculas constituye la sonda, que se enfrenta al chip para que se hibride con su secuencia complementaria. Tras un lavado, se visualiza la fluorescencia de los cADNs que se han quedado fijados para determinar qué genes se están expresando (10).

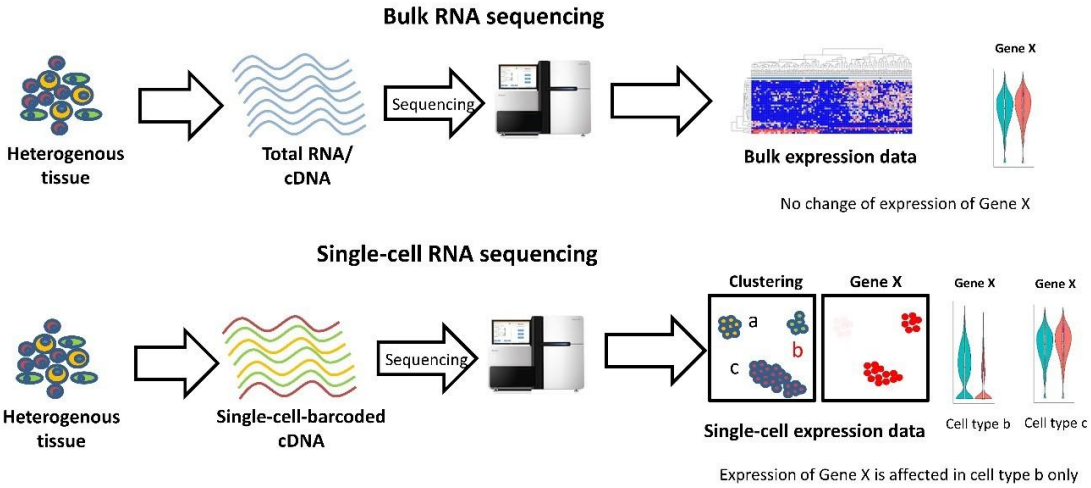
RNA-seq utiliza técnicas innovadoras basadas en secuenciación masiva. El proceso comienza con la extracción de ARN del tejido o subconjunto celular de interés, seguido de su fragmentación y conversión a cADN, generando así una librería de fragmentos de cADN. A estos fragmentos se les añaden uno o dos adaptadores (fragmentos cortos conocidos) en los extremos. Las moléculas resultantes son reconocidas por la polimerasa mediante los

adaptadores y secuenciadas por procesos de alto rendimiento, obteniendo lecturas, que son fragmentos leídos por el secuenciador. Finalmente, estas lecturas pueden alinearse al genoma o a transcritos de referencia, o bien emplearse un ensamblaje *de novo* (2), con el fin de identificar los genes expresados y cuantificar sus niveles de expresión.

Ambas técnicas, *Microarrays* y *RNA-seq*, son significativamente más rápidas, económicas, informativas, precisas y fiables que los métodos tradicionales, como las construcciones de librerías de ADN complementario, las bibliotecas de hibridación sustractiva supresora o las etiquetas de secuencias expresadas (11). Desde 2005, las técnicas de nueva generación han dejado en un segundo plano a los *Microarrays,* permitiendo el descubrimiento de nuevas especies de moléculas de ARN (12).

La primera técnica de *RNA-seq* que se popularizó fue *bulk RNA-seq*, que se basa en medir la expresión génica promedio de una población celular tomada en conjunto. Es ampliamente usada para comparar condiciones biológicas, como control frente tratamiento, permitiendo identificar genes diferencialmente expresados a nivel global. Un gran avance del *RNA-seq* en la última década ha sido el desarrollo de la técnica *scRNA-seq*, que permite secuenciar el ARN de células individuales. El proceso se inicia con una digestión del tejido donde se aíslan las células, para posteriormente analizar el ARN de cada una de ellas (13). Esta aproximación ha permitido caracterizar con mayor precisión la heterogeneidad celular y describir perfiles de expresión génica para miles de transcritos, correspondientes a aproximadamente 20.000 genes del genoma humano (13). El flujo de trabajo de ambas técnicas de *RNA-seq* se muestra en la Figura 1.

Figura 1. *Introducción a las tecnologías de RNA-sequencing*



Reproducido de https://www.singlecellcourse.org/index.html.

2.1.2. Transcriptómica espacial

Las técnicas de secuenciación de ARN mediante *NGS*, como *Bulk RNA-seq* o *scRNA-seq*, se diferencian en la resolución, lo que les confiere ventajas o inconvenientes según el experimento que se desee evaluar. Sin embargo, ninguna de las dos técnicas tiene en cuenta el contexto espacial.

Para cubrir esta necesidad, desde hace aproximadamente diez años, han emergido las técnicas de transcriptómica espacial. Estas técnicas cuantifican la expresión de transcritos o proteínas incorporando simultáneamente la información espacial de su localización, para poder generar una representación realista de los procesos biológicos de la muestra analizada. De esta forma, se puede obtener información acerca de las células y su microambiente, las señales a las que las células están expuestas y las interacciones célula-célula (11). Se han desarrollado diferentes enfoques para el análisis espacial experimental, que se diferencian por la medida del tejido analizado, que puede variar desde una superficie de 1mm² hasta un órgano entero. Actualmente, existen dos opciones para la detección espacial de ARN: tecnologías basadas en imagen y basadas en secuenciación, que se recogen de manera esquemática en la Figura 2 (14).

2.1.2.1. Métodos basados en imagen

Las tecnologías de imagen están fundamentadas en la técnica de hibridación fluorescente in-situ de moléculas únicas, *single molecule Fluorescence In Situ Hybridization* (*smFISH*), la cual permite la detección de moléculas individuales de ARN, utilizando sondas de ADN marcadas con fluorescencia (11).

El procedimiento empieza fijando el ARN del tejido o las células que se quieren muestrear, para preservar la estructura y mantener el ARN en su localización original. A continuación, se diseñan sondas complementarias a las secuencias del ARN de interés, marcadas con fluoróforos. Al ponerse en contacto, las sondas se hibridan por complementariedad con las moléculas de ARN de dentro de las células de forma que, cada molécula de ARN queda marcada con una señal fluorescente. Se capturan imágenes para poder asociar la fluorescencia emitida por una molécula individual de ARN, con su localización. Llegados a este punto, se añade una ADNasa para separar la hibridación y se repite el ciclo, donde se añade una nueva sonda con un nuevo color. Dado que el número de fluoróforos (colores) es limitado, el número de genes que pueden ser observados simultáneamente queda restringido. Para superar esta limitación, se generan diferentes rondas secuenciales de hibridación. Así, la combinación de colores a lo largo de distintas rondas generará un código de barras único para cada ARNm proveniente de cada gen. De esta forma, se agruparán las moléculas que tengan el mismo código de barras, asumiendo que son el mismo tipo de ARNm y, por lo tanto, que son expresados por el mismo gen. Este avance permite tener tantos "colores" como posibles combinaciones entre los colores de los fluoróforos iniciales, ampliando la capacidad de identificar genes.

Finalmente, hay una serie de pasos para convertir las diferentes imágenes crudas en una matriz de conteo de transcritos por gen (*gene-spot matrix*). Primeramente, se filtran las imágenes para quitar el ruido del fondo, luego se alinean las imágenes de las diferentes rondas para que la misma localización de los píxeles represente el mismo transcrito, y finalmente, las señales de cada spot se combinan para generar un código de barras que servirá para decodificar y asignar cada transcrito a un gen específico (14).

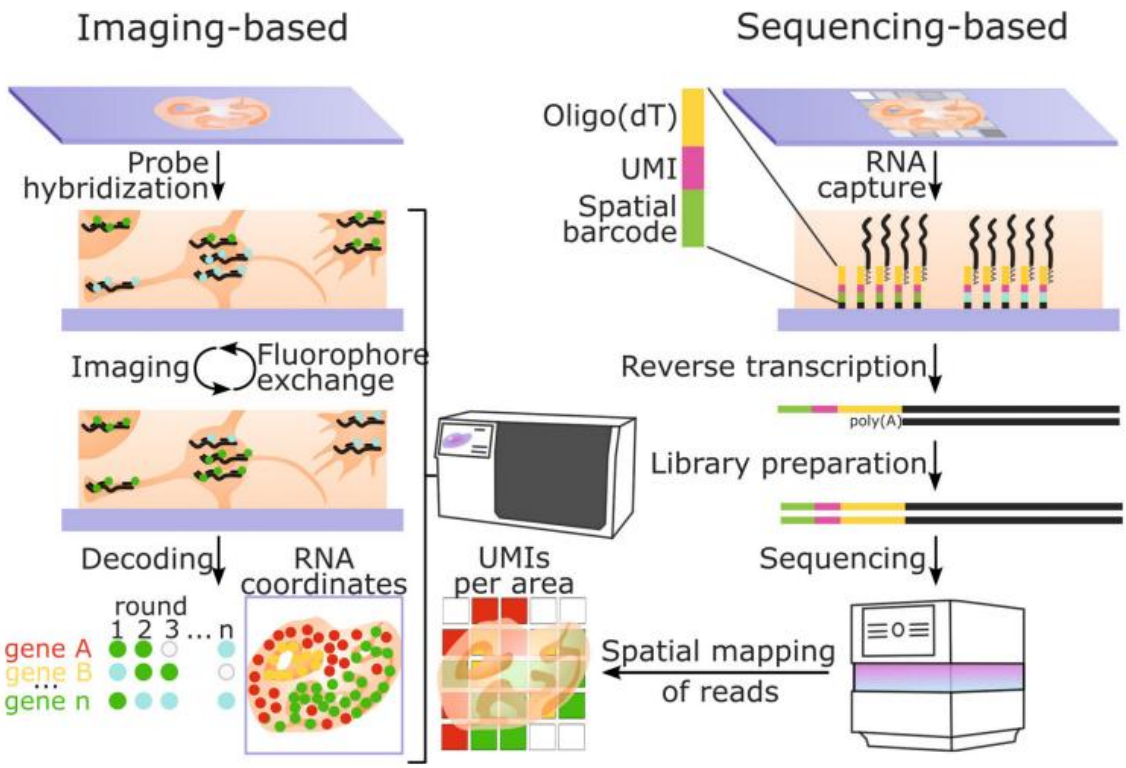
2.1.2.2. Métodos basados en secuenciación

Por su parte, las tecnologías de secuenciación mezclan la clásica tecnología de *Microarrays* con las técnicas de secuenciación de nueva generación. A diferencia de la técnica de *Microarrays* tradicional, el contenido de la matriz está compuesto por sondas de transcriptómica espacial, que son oligonucleótidos sintéticos compuestos por cuatro fragmentos. La primera parte es la secuencia de anclaje, que permite fijar la cadena al chip; luego está el código de barras espacial, que actúa como identificador único de posición; seguidamente se encuentra el UMI (*Unique Molecular Identifier*), que permite distinguir moléculas individuales de ARNm y evitar conteos duplicados tras la amplificación; y, finalmente, la cola de poliT, encargada de capturar los transcritos al complementar con la cola poliA de estos (11).

Esta tecnología se basa en seleccionar un corte fino del tejido a analizar y colocarlo sobre el chip que contiene la matriz de sondas. Cada punto, conocido como *spot*, contiene miles de sondas con un código de barras distinto, lo que permitirá saber de qué lugar exacto del tejido proviene cada molécula de ARNm capturada y secuenciada. Entonces se realiza la lisis de las células directamente sobre la matriz, permitiendo que el contenido de ARNm se hibride mediante la complementariedad de la cola poliT-poliA. Se realiza una retrotranscripción in situ, que genera moléculas de cADN que incorporan tanto el código de barras como el UMI. Seguidamente, se recolectan los cADN, se amplifican y se utilizan para construir bibliotecas que se secuenciaran mediante las máquinas de secuenciación masiva. Las lecturas extraídas se mapean contra el genoma o transcriptoma de referencia para saber a qué gen corresponde cada transcrito (11).

El resultado es una matriz de conteo de transcritos, donde las filas son los distintos genes y las columnas los *spots*. En cada spot se contabilizan las moléculas de ARNm (identificadas mediante los UMI) provenientes de cada gen. Estas matrices contienen, además, a diferencia de las matrices de *RNA-Seq*, la localización bidimensional (2D) de cada spot en el tejido gracias a los códigos de barras identificativos (11).

Figura 2. *Transcriptómica espacial.*



Reproducido de Kulasinghe (11). Esquema de los pasos de detección del ARN de la tecnología basada en imagen y de la tecnología basada en secuenciación.

2.1.2.3. Consideraciones para el uso

Al iniciarse en la transcriptómica espacial, existe el desconocimiento sobre cuándo se puede usar la técnica y si todos los tejidos son óptimos. La respuesta, por ahora, es que cualquier tejido intacto que contenga ARNm viable será adecuado para la transcriptómica espacial. Por este motivo, es una técnica ampliamente usada en neurociencia, cáncer, inmunología y biología del desarrollo (14). Ahora bien, hay una serie de requisitos o recomendaciones que es bueno tener en cuenta antes de adentrarte en esta técnica y que, a su vez, ayudan a identificar qué tecnología es la más adecuada para el estudio que desee desarrollar.

Las características intrínsecas de los tejidos son de vital importancia, ya que es necesario conocer cómo optimizar la técnica según el tejido a analizar, para ajustar los parámetros y desarrollar a medida el experimento (14). Por ejemplo, es recomendable conocer si existe autofluorescencia del tejido a analizar (14) y la calidad de este, ya que con el tiempo el ARNm se degrada y se fragmenta.

Otro aspecto a tener en cuenta es el número de muestras. Aquí se genera una distinción entre las metodologías basadas en imagen, en las que solo se analiza una imagen, y las de secuenciación, en las que pueden analizarse hasta cuatro secciones simultáneamente. La eficiencia de capturar el ARNm también es de vital importancia, donde se consigue un mejor resultado en las prácticas basadas en imagen que en las de secuenciación. Lo mismo se puede percibir en la resolución espacial. Las tecnologías basadas en imagen pueden conseguir una resolución subcelular, permitiendo el análisis de célula única, mientras que las basadas en secuenciación recogen el ARN de un área concreta, que puede ser un *spot* o una región de interés (ROIs), que son agrupaciones de células. Por último, se diferencian en el número de genes capaces de analizar: entre 1000-5000 genes en métodos de imagen y el genoma completo en los de secuenciación (14). Estas y otras características se recogen en la Tabla 1.

Tabla 1. *Tabla comparativa de las características principales de los métodos basados en imagen y en secuenciación.*

	Basados en Imagen (smFISH / ISS)	Basados en Secuenciación (ST / Visium)
Estrategia	Detección in situ con sondas fluorescentes.	Captura de ARN en chips con códigos de barras.
Resolución	Célula única o nivel Subcelular.	Spot: 1-10 células ROI: 10-100 células
Sensibilidad	Muy alta (detecta moléculas individuales).	Menor (depende de la eficiencia de captura del chip).
Target	Dirigido, captura el transcriptoma de un panel de genes predefinido.	Spot: No dirigido, captura todo el transcriptoma (Poly-A) de cada Spot. ROI: Dirigido, captura el transcriptoma de un panel de genes predefinido en cada ROI.
Output	Coordenadas exactas (x, y) de cada transcrito.	Matriz de conteo por "spot/ROI" (píxel biológico).
Modelos en el mercado	10x Xenium, Nanostring CosMx, smFISH, MERFISH, seqFISH+	10x Visium, Nanostring GeoMx

Elaboración propia.

Basándonos en los factores comentados anteriormente, se sugiere que los experimentos destinados a comprobar una hipótesis (a menudo, experimentos altamente precisos para analizar la regulación de genes diferencial y rutas metabólicas) se relacionen con métodos espaciales basados en imagen. Por lo contrario, si el estudio tiene un enfoque de generación de hipótesis, se recomienda usar técnicas basadas en secuenciación, ya que el objetivo es revelar mecanismos (14).

2.1.2.4. Aplicaciones

Las aplicaciones de la técnica espacial son similares a las de las técnicas de *RNA-seq*, pero aportando el contexto espacial. Uno de los análisis más recurrentes en *RNA-seq* es la generación de clústeres, donde se agrupan las células mediante algoritmos no supervisados según sus perfiles de expresión genética. De esta forma, se acaba conociendo qué linajes celulares hay en la muestra y como se distribuyen. También se puede conocer cómo de variable es la expresión de un gen en distintas localizaciones de la muestra, porque la distribución de la expresión no es aleatoria ni tampoco lo es la distribución de los distintos linajes celulares en un tejido (15). Por estos motivos se ha convertido en una herramienta con gran potencial en aplicaciones clínicas cuando se necesita mayor sensibilidad y exactitud, incluyendo diagnóstico precoz, pronóstico, identificación de marcadores espaciales, desarrollo de fármacos y ensayos clínicos, ente otros.

En el Anexo de este trabajo, se proponen ejemplos de aplicaciones más específicos.

2.2. ANÁLISIS DE DATOS DE TRANSCRIPTÓMICA ESPACIAL

Cada vez es más claro que los análisis bioinformáticos son una parte muy significativa de la investigación biomédica, con retos con el mismo peso que el procedimiento experimental. La dificultad del análisis no solo parte de la cantidad masiva de datos a analizar, sino de los errores que se han podido introducir durante el experimento, la secuenciación, así como por la complejidad del transcriptoma (8). Dejando de lado las especificaciones de cada equipo, el objetivo principal del análisis de datos es integrar información de la expresión genética y de la localización espacial para permitir la extracción de información biológica útil a partir de datos obtenidos experimentalmente (17).

El proceso de análisis de datos se basa en dos partes: el preprocesamiento, que incluye desde la gestión de datos crudos hasta la generación de la matriz de conteo, y los análisis posteriores *(downstream analysis)* que obtienen una cantidad enorme de información de muy diversa índole (17).

2.2.1. Preprocesamiento

El preprocesamiento técnico es la única parte del análisis de datos que se diferencia entre la técnica basada en imágenes y en secuenciación. Sus diferencias se recogen en el Anexo, pero ambas generan como resultado la matriz de conteos que para cuya realización se ha tenido que identificar la ubicación, el tipo y la cantidad de cada transcrito.

Una vez conseguida la matriz de trabajo, antes de empezar con los análisis posteriores, se le aplica a ésta una serie de transformaciones estadísticas y controles de calidad para acabar de preparar los datos, proceso conocido como preprocesamiento analítico (14). Los pasos incluyen, entre otros, el control de calidad, la reducción de la dimensionalidad, la normalización y la corrección del efecto de lote. El objetivo de esta parte es asegurar que los datos que van a entrar en los procesos de análisis posteriores tienen alta calidad, ya que posteriormente, se busca descubrir sus implicaciones biológicas (18). Se detallan estos pasos en la sección Anexos.

2.2.2. Análisis posteriores *(Downstream)*

Hay un rango amplio de diferentes análisis posteriores posibles en datos de transcriptómica espacial, con diferentes objetivos y diferentes herramientas para llevarlos a cabo.

La clusterización espacial es de los análisis más importantes de esta tecnología. Este análisis tiene en cuenta tanto la expresión génica como la ubicación y agrupa los perfiles más parecidos. Para ello tiene en cuenta la vecindad física de los *spots*, de modo que los clústers resultantes representan dominios tisulares coherentes que reflejan la histología y la morfología celular para ayudar en la recuperación de la estructura celular de los datos (18,20,21).

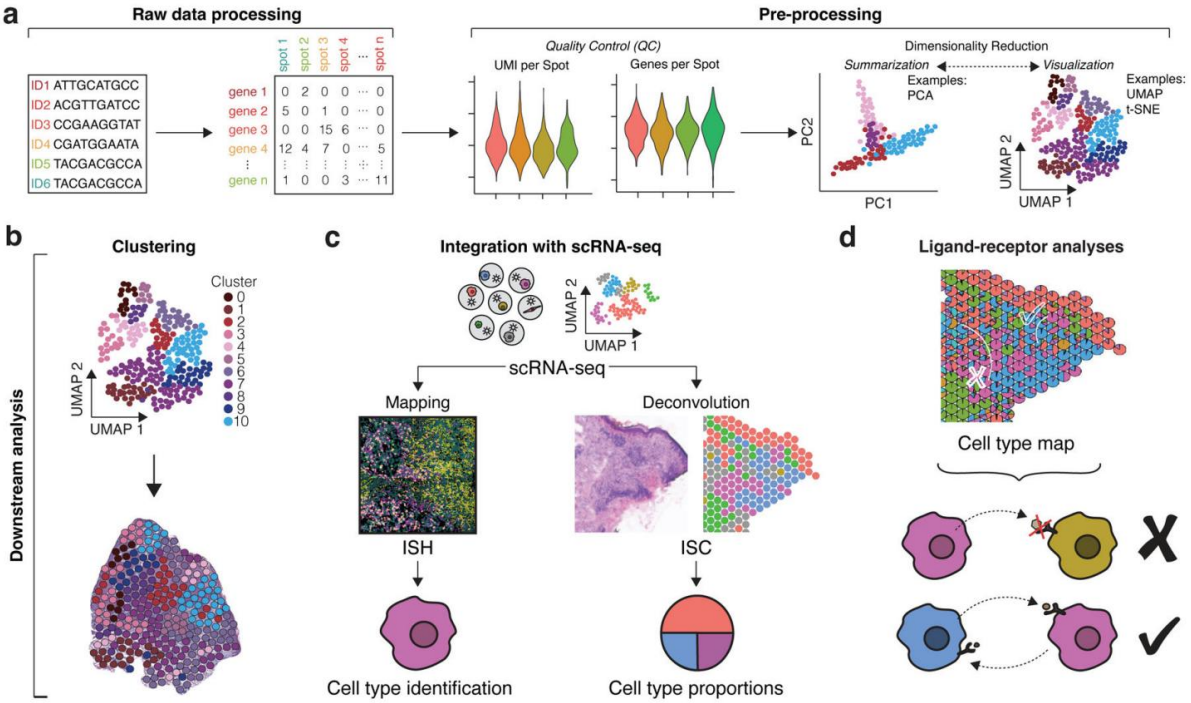
En transcriptómica espacial, para identificar genes espacialmente variables (*spatially variable genes*), se utiliza el análisis de expresión diferencial directamente entre distintas regiones del

tejido. Nuevos enfoques computacionales, que se basan en el uso de estadística, aprendizaje automatizado y aprendizaje profundo, se han desarrollado para abordar este análisis (17). Posteriormente se aplican los procesos correspondientes para encontrar el análisis de patrones de expresión genética, donde se inferirán las funciones de los genes.

También existen análisis, como el mapeo o la deconvolución, que cuentan con la integración de funciones y análisis de *scRNA-seq*. El mapeo sirve para asignar un tipo celular definido por la técnica *scRAN-seq,* a su localización. Se usan algoritmos de integración que comparan los perfiles de expresión de *scRNA-seq* con los perfiles de expresión de los *spots* de la técnica de transcriptómica espacial. El resultado es una predicción de qué células o tipos celulares se encuentran en cada ubicación espacial. Se usa principalmente en técnicas de secuenciación basada en imágenes. La deconvolución sirve para poder separar las distintas proporciones celulares según su origen celular (17). Se usa información de referencia de *scRNA-seq,* donde los perfiles de expresión de cada tipo celular son conocidos. Algoritmos de deconvolución ajustan la mezcla de expresión de los spots para asignar proporciones de cada célula.

Otros análisis como las interacciones célula-célula o gen-gen también son parte de los análisis posteriores, donde se busca interpretar la interacción molecular para entender el funcionamiento de los microambientes específicos, la comunicación y señalización inter- e intracelular y los módulos génicos, que son grupos de genes que se coexpresan: su nivel de expresión varía a la vez, ya que participan en funciones biológicas relacionadas (17). En la Figura 3 se recoge un flujo de trabajo del análisis de datos de transcriptómica espacial general.

Figura 3. *Flujo de trabajo del análisis de datos*



Reproducido de A. J. Piñeiro (18). a) Procesamiento y Control de Calidad: Los datos brutos se convierten en matrices de conteo (genes por punto de captura). Se filtran los datos por calidad (UMIs y genes por punto) y se aplican técnicas de reducción de dimensiones (PCA, UMAP o t-SNE) para facilitar su visualización. b) Agrupamiento (Clustering): Se agrupan los puntos con perfiles genéticos similares y se representan sobre la imagen original del tejido para su interpretación. c) Integración con scRNA-seq: Se combina la información de secuenciación de célula única (scRNA-seq) con los datos espaciales. El mapeo identifica tipos celulares específicos, mientras que la deconvolución calcula la proporción de diferentes células en cada punto de captura. d) Análisis de interacción ligando-receptor: Utiliza los mapas celulares para predecir la comunicación entre células basándose en su proximidad física.

2.2.3. Desarrollo de *Pipelines* y *workflows*

Los pasos que se llevan a cabo para realizar el análisis de los datos, mencionados en el apartado anterior, se ejecutan mediante *pipelines* bioinformáticos. *Pipeline* se traduce como tubería y es una secuencia automatizada de pasos para procesar datos biológicos. En sí, son comandos o scripts, en los que la salida de un proceso es la entrada al siguiente. Un *pipeline* involucra múltiples herramientas, paquetes y bibliotecas que se integran para llevar a cabo el análisis. Todo eso se recoge en un nuevo concepto, el *workflow*, que es el marco de trabajo completo que define y gestiona la ejecución de un *pipeline*.

Los *pipelines* tradicionales están muy ligados a las infraestructuras de computación locales donde se ejecutan. No permiten reanudar la ejecución de un proceso interrumpido, tienen poca documentación, no cuentan con una trazabilidad de los parámetros y versiones de

paquetes utilizados y requieren de instalación manual, lo cual impide una fácil distribución. Para solucionar los inconvenientes citados, se han desarrollado los *Workflow Managers*, gestores de flujo de trabajo. Estos permiten la utilización de *pipelines* de análisis complejos en distintos entornos de computación asegurando la máxima reproducibilidad, portabilidad, escalabilidad, robustez y modularidad de los procesos ejecutados (22).

Los *pipelines* enfocados al análisis de datos de transcriptómica espacial tienen una misma estructura o esquema, pero cada equipo comercial cuenta con su propia manera de analizar los datos, es decir, con su *pipeline* desarrollado. La mayoría de los *pipelines* son de dominio público, lo que facilita a los usuarios poderlos reproducir. Es habitual en la implementación de un pipeline, empezar con una revisión de la bibliografía asociada a la tecnología que se desea analizar. Seguidamente, diseñar un esquema con las distintas etapas del *pipeline* junto con las herramientas empleadas. En dicha representación, las viñetas corresponden a los bloques o recuadros que representan cada etapa del proceso. Estas sirven para organizar y simplificar la lectura del flujo de trabajo, especialmente en diagramas o figuras.

2.3. GeoMx™ *Digital Spatial Profiler* (DSP)

GeoMx™ DSP es un instrumento desarrollado por NanoString Technologies que permite estudiar la expresión de proteínas y genes *in situ*, manteniendo la información espacial, en regiones tisulares definidas por el usuario (23). Si bien se puede encontrar categorizada como tecnología de secuenciación, GeoMx™ DSP emplea una estrategia que combina sondas marcadas con códigos de barras con la selección de regiones de interés (ROI) para determinar la ubicación espacial de los transcritos (14). Nace como solución a las técnicas espaciales de secuenciación, aportando la capacidad de elegir la zona del tejido específica a secuenciar (24). Este trabajo se centra únicamente en la aplicación de GeoMx™ DSP en transcriptómica.

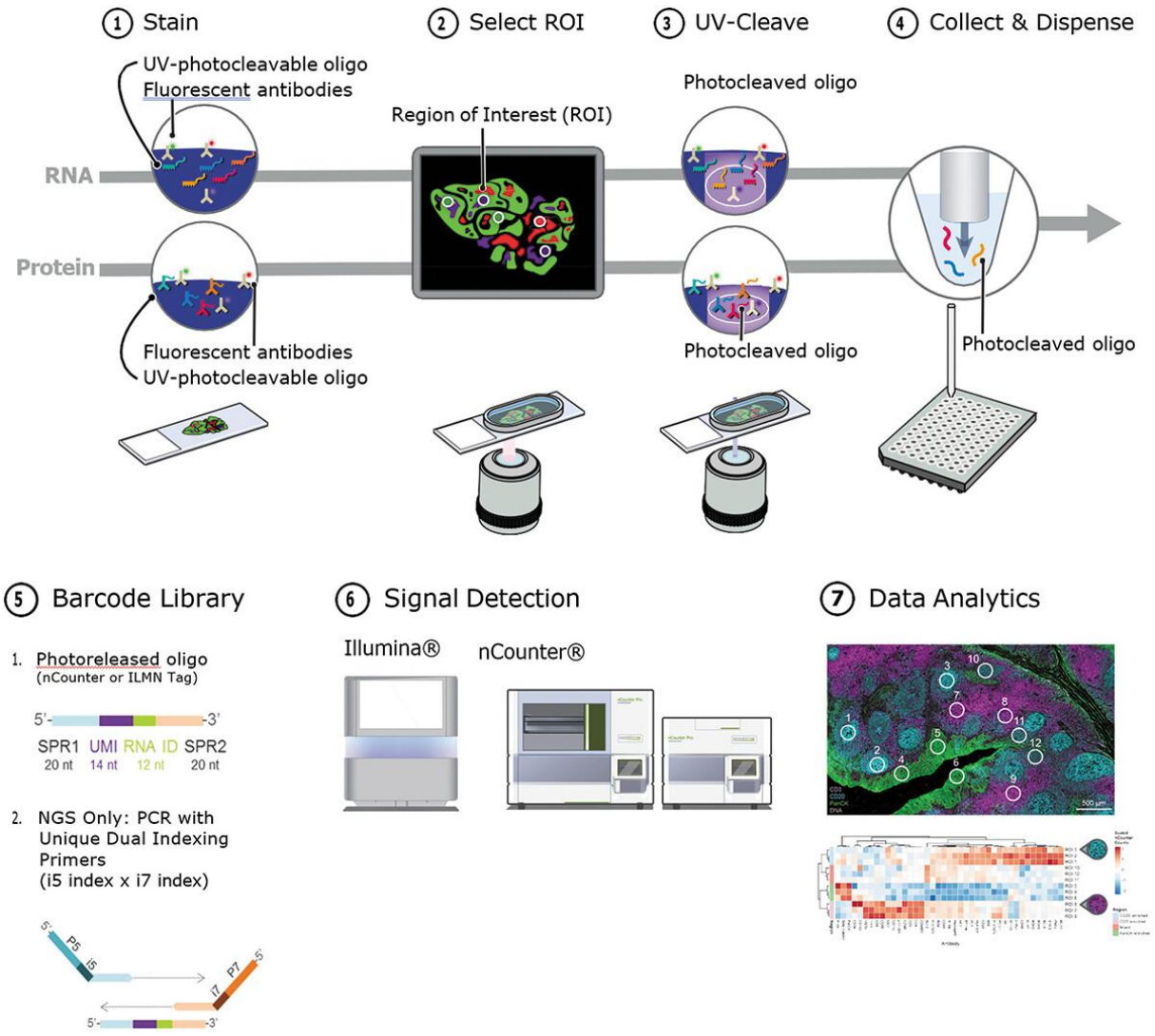
2.3.1. Procedimiento experimental

El procedimiento que se detalla a continuación se ilustra en la Figura 4. El procedimiento comienza con la selección del tejido a analizar y su colocación en un portaobjetos. Seguidamente se hace una tinción con dos tipos de reactivos. El primero es el reactivo de imagen, el cual se compone de sondas marcadas con fluorescencia que permiten visualizar la morfología del tejido y así poder elegir qué área se va a perfilar molecularmente. Es posible

usar hasta cuatro fluoróforos para teñir la muestra. El segundo reactivo es el de perfilado, constituido por sondas compuestas por una secuencia complementaria al ARNm diana, una cadena de oligonucleótidos que se utiliza como código de barras de identificación único y un enlace ultravioleta (UV) liberable que une estas dos secuencias. La exposición de luz UV comportará la liberación de los códigos de barras dada la destrucción del enlace (25).

Una vez completada la tinción, el portaobjetos se introduce en el instrumento *Digital Spatial Profiler* (DSP) donde se adquieren imágenes fluorescentes del tejido y se eligen las regiones de interés. Es importante matizar, que las regiones de interés se pueden segmentar en áreas de iluminación (conocidas como AOI o segmentos), lo que permite subdividir la ROI en secciones más pequeñas. Esta segmentación es ampliamente usada en experimentos donde se desea medir la expresión génica de un tumor, y de su microambiente. Si se segmenta la ROI, se obtiene de forma independiente la cuantificación de ARNm de diferentes zonas celulares. El proceso continúa con la exposición de las ROI seleccionados a luz UV, dando lugar a la liberación de los distintos códigos de barra específicos de esa región. Estos oligonucleótidos liberados son aspirados por microcapilaridad y depositados en un pozo individual de una placa de 96 pozos. Este proceso se repite para todas las ROI seleccionadas, generando así un pocillo por cada ROI. Finalmente, los códigos de barras recogidos se analizan mediante secuenciación de nueva generación (NGS) o mediante el sistema *nCounter Analysis System*, lo que permite cuantificar los transcritos generados en cada ROI y obtener un perfil espacial de la expresión génica (24,25). Llegados a este punto, se analizan los datos mediante programas bioinformáticos.

Figura 4. *Pasos del Flujo de Trabajo de Perfilado Espacial GeoMx™ DSP.*



Reproducido de D.Krull (23). 1. Marcaje de la muestra: Las láminas se tiñen con sondas fluorescentes (ISH) o anticuerpos para ver la morfología del tejido. Además, se añaden sondas para ARN o proteínas marcadas con un código de barras de ADN que se puede liberar mediante luz UV. 2. Selección de áreas (ROI): Se toma una imagen del tejido y se seleccionan las "Regiones de Interés" (ROI) o poblaciones celulares específicas que se desean analizar. 3. Liberación por luz UV: Cada ROI se expone a luz ultravioleta para liberar los códigos de barras de las sondas. Estas regiones pueden subdividirse aún más en "Áreas de Iluminación" (AOI) para aislar tipos celulares muy específicos. 4. Recolección: Los códigos de barras liberados se recogen en una placa de 96 pocillos. 5 y 6. Cuantificación: Los códigos de barras se cuentan fuera del instrumento, ya sea mediante secuenciación (NGS) tras preparar una librería, o mediante el sistema de conteo directo nCounter. 7. Análisis de datos: Los resultados se procesan en el software de análisis del equipo o mediante paquetes de programación en R (Bioconductor).

2.3.2. Análisis de datos con GeoMx™ DSP Analysis Suite

GeoMx™ DSP Analysis Suite es el software oficial que utiliza la plataforma de GeoMx™ DSP para poder analizar los datos generados por el equipo (26,27). Se caracteriza por tener una

interfaz gráfica muy intuitiva que permite que sea de fácil uso, ya que está pensada para biólogos e investigadores sin necesidad de que tengan un conocimiento profundo en bioinformática. Puede ser usada en versión web o versión local. Se obtiene mediante la compra del equipo GeoMx™ DSP, por lo que su uso se restringe a sus usuarios. Además, cuenta con la posibilidad de exportar a R o Python, ya que han creado el paquete GeoMxTools para poder realizar el análisis en local y complementarlo con análisis estadísticos más complejos (26,27).

Esta plataforma cuenta con un *pipeline* que se compone de las siguientes etapas (26,27), con parámetros asignados por defecto que son ajustables:

- Fase previa al análisis: cargar los conteos DCC (hay que convertir previamente de FASTQ a DCC), crear el estudio, añadir anotaciones.
- Control de calidad:
    - Visualización de los conteos de lecturas generados por el secuenciador.
    - Ajustes de los parámetros del control de calidad de los segmentos (AOI) según la señal (umbral de lecturas crudas, porcentaje de lecturas alineadas, saturación de secuenciación), según el ruido de fondo (media geométrica del conteo de sondas negativas, control sin plantilla), según los parámetros DSP (número mínimo de núcleos).
    - Ajustes de los parámetros del control de calidad de las pruebas biológicas según valores atípicos bajos (media geométrica de sondas), según valores atípicos altos (prueba de Grubbs).
- Filtrado: Se puede filtrar por segmentos (AOI) (para eliminar los segmentos con baja señal) o filtrar por objetivos (para detecta los genes muy poco expresados). Los dos filtrados tienen en cuenta dos factores:
    - Umbral de expresión: valor mínimo de expresión génica para que un gen se considere que está siendo expresado.
    - Frecuencia: porcentaje mínimo de genes que tienen que cumplir estar por encima del umbral, para que el segmento (AOI) se mantenga.
- Normalización:

  - Normalización al tercer cuartil: se calcula para cada ROI, el valor del tercer cuartil Q3 (el valor que deja por debajo el 75% de los datos) de todos los genes. Luego se divide la expresión de cada gen por ese valor Q3.
  - Normalización de fondo: es un ajuste para eliminar el ruido o señal no específica de fondo que usa el valor del conteo de las sondas negativas. Se recomienda usar esta técnica cuando el conteo de las sondas negativas de los genes es superior a 10.
  - No recomiendan usar métodos como escalado por área o escalado según el conteo de núcleos.

- Exploración de datos: Una de las formas más típicas de explorar los datos que ofrece el software es la creación de *clusterings* no supervisados jerárquicos, que agrupan los genes por patrones similares de expresión. Existen otras opciones para visualizar los datos como *volcano plot* (para comparar la expresión génica de dos grupos), *box plot* (examina el rango dinámico y la distribución de expresión de un gen concreto) o *Pathway analysis* (explora interacciones entre grupos de genes que están relacionados en sus vías biológicas).
- Tests estadísticos: El GeoMx™ DSP Data *Analysis Suite* ofrece las siguientes pruebas estadísticas, cada una de ellas será adecuada dependiendo del diseño experimental:
  - Modelo Lineal de Efectos Mixtos (*Linear Mixed Effect Model,* LMM): Se utiliza cuando hay muestras relacionadas o datos con efectos aleatorios. Ha sido diseñado para manipular datos que provienen de un mismo individuo o tejido. Es más ampliamente usado cuando se tiene una sección de un tejido por portaobjetos, múltiples tejidos por grupo de comparación o múltiples ROI por tejido.
  - T-test para datos no apareados (*Unpaired T-test*): Compara las medias de dos grupos provenientes de observaciones independientes.
  - T-test para datos apareados (*Paired T-test*):

    Se usa para comparar las medias de dos grupos relacionados o emparejados que provienen de la misma observación, como serían los segmentos originados por el mismo ROI.

- Prueba U de Mann-Whitney (Mann-Whitney U-test)
  Es la alternativa no paramétrica al t-test. No supone una distribución normal de los datos.

2.3.3. Métodos alternativos

2.3.3.1. Limitaciones de la normalización

En 2022 el Laboratorio de Inmunología Tumoral de Róterdam, publicaba en *iScience* (28) que el *pipeline* tradicional de GeoMx™ DSP *Analysis Suite* presentaba limitaciones significativas en la normalización de los datos y la corrección de sesgos técnicos. El estudio utilizó datos de GeoMx™ DSP obtenidos de estudios apareados de tumores primarios y gliomas recurrentes, incluyendo réplicas biológicas de ocho muestras de cuatro pacientes, con análisis en 72 regiones de interés. Durante el análisis de datos crudos, los autores observaron que la intensidad de las señales y de la expresión genética, eran muy distintas entre muestras. Cuando hacían la comprobación con datos equivalentes con técnicas muy estandarizadas como *bulk RNA-seq*, obtenían distribuciones similares entre las mismas muestras, lo que les hizo plantear que esas diferencias no correspondían a cambios biológicos reales sino a técnicos (28). Utilizaron cuatro modelos de normalización (CPM ajustado, DESeq2, corrección Gamma-fit y normalización por cuantiles) para comprobar si con otro método de normalización se conseguía evitar los efectos técnicos. La normalización por cuantiles fue el único método que prácticamente eliminó la dependencia entre la señal y el ruido técnico con los cambios de expresión. Logró una buena concordancia entre los datos obtenidos y los comparativos en *bulk RNA-seq* y logró extraer conclusiones con sentido biológico.

2.3.3.2. StandR

StandR corresponde a las iniciales de *Spatial Transcriptomics Analysis of Nanostring's Dsp data in R*. Es un paquete de R que se encuentra en la librería Bioconductor que permite realizar un análisis de principio a fin de datos de GeoMx™ DSP. Surgió en 2023 por las necesidades mencionadas en el apartado anterior, ya que el *pipeline* estandarizado de GeoMxTools no suelen tener en cuenta adecuadamente la variabilidad técnica presente en los datos ni la complejidad de los diseños experimentales, lo que limita la precisión y la fiabilidad de los análisis posteriores (29). Estos sesgos ocurren, por ejemplo, en la variabilidad aportada por

los portaobjetos que se usan, la edad del material, las técnicas usadas para fijar el tejido, el tiempo de fijación, la preparación del tejido y de la muestra, creando una variabilidad entre muestras y conllevando descubrimientos falsos. Por lo mencionado anteriormente, es importante seguir buenas prácticas en el control de calidad, el filtrado, la normalización y la corrección de los efectos de lote. StandR proporciona funciones de control de calidad, normalización y evaluación y visualización para datos transcriptómicos de GeoMx™ DSP, y recomienda un flujo de trabajo que incorpora el conocido *pipeline* de expresión diferencial limma-voom, así como otros análisis posteriores, para identificar genes diferentemente expresados a partir de experimentos de GeoMx™ DSP(29).

3. MATERIALES Y MÉTODOS

3.1. MATERIALES

En este capítulo se detallan los materiales y métodos empleados para el desarrollo y la implementación de un *pipeline* de análisis de datos de la tecnología GeoMx™ DSP. Se basa en herramientas de código abierto y se evalúan las características del alcance de las técnicas.

3.1.1. Hardware

El trabajo se ha desarrollado y ejecutado de forma local en un ordenador portátil cedido por el VHIR, los detalles se recogen en la Tabla 2.

Tabla 2. *Especificaciones técnicas del ordenador.*

Modelo	Lenovo ThinkPad L15 Gen 2
Sistema Operativo	Windows 11 Pro for Workstations v21H2 (22000.2538)
CPU	11th Gen Intel® Core™ i5-1145G7 @ 2.60GHz
GPU	Intel® Iris® X e Graphics
Memoria RAM	8 GB DDR4 (7,66 GB utilizable) + 19,3 GB memoria virtual
Disco	KXG6AZNV256G TOSHIBA, 256 GB

Elaboración propia.

3.1.2. Software

Para el desarrollo del *pipeline* se usa el lenguaje de programación R (v4.5.1) (30), mediante el entorno de desarrollo integrado RStudio (v2025.09.1 401) (31). Se utiliza el formato de fichero HTML de RMarkdown (v2.30) (9) para la creación de los informes, ya que permite combinar código, resultados y texto en un único documento.

Finalmente, para poder hacer de este trabajo una herramienta estructurada, reproducible y fácil de compartir con cualquier usuario, se emplea el paquete *workflowr* (v1.7.2) (32).

3.1.3. Conjunto de datos

Para la búsqueda de un *dataset* adecuado, se ha realizado una evaluación de los distintos repositorios de código abierto existentes con datos de GeoMx™ DSP. Se han encontrado repositorios oficiales y de libre acceso como Nanostring (33), Gene Expression Omnibus (GEO) (34), ArrayExpress (4), European Genome-Phenome Archive (EGA) (35) o Zenodo (36). En los anexos se describe un breve resumen de las distintas plataformas, el número de *datasets* de GeoMx™ DSP que contiene y el tipo de tejido que se publica, para analizar si existe un patrón en los tejidos utilizados con GeoMx™ DSP.

Una vez conocido el estado actual de los repositorios de GeoMx™ DSP, se seleccionó el *dataset* más adecuado para la realización del trabajo. Para garantizar un análisis robusto y reproducible, se recomienda que el conjunto de datos incluya un número alto de ROIs y un mínimo de 3-5 réplicas por cada anotación (37). Asimismo es deseable que cuente con una cobertura amplia como el panel de *Whole Transcriptome Atlas* (WTA) (25), y disponer de datos bien anotados, lo que facilita la interpretabilidad de los resultados (38). Finalmente, se valora positivamente que el estudio asociado al *dataset* haya sido previamente publicado, ya que facilita la contextualización y permite contrastar los resultados obtenidos con una referencia bibliográfica.

Para la implementación del *pipeline* se trabaja con el *dataset* utilizado en el *paper*: Análisis transcriptómico espacial revela falta de respuesta al bloqueo de PD-1 en glioblastoma recurrente (39,40) descargado en el repositorio Zenodo. Es un estudio que busca interpretar la expresión genética de pacientes con glioblastomas tratados con inhibidores de PD-1 (inmunoterapia) y de pacientes con glioblastomas no tratados (control), para evaluar el efecto del tratamiento. Se detalla el *dataset* en el apartado 4.5.1.

3.2. MÉTODOS

3.2.1. Familiarización con Software de GeoMx™ DSP Analysis Suite

Para la familiarización de los análisis de datos de GeoMx™ DSP, se propone empezar con el software de GeoMx™ DSP *Analysis Suite* desarrollado por Nanostring. Este software es un servicio que va integrado y se descarga e instala al adquirir el equipo. Debido a que el acceso y el uso del software es privado, ya que depende de la posesión de un equipo para tener

licencia, no se ha podido emplear directamente un análisis en la plataforma. Por este motivo se han realizado tres prácticas para aprender su funcionamiento. La primera es la visualización de la sesión de *Nanostring University: GeoMx™ DSP Data Analysis Workflow Session for RNA NGS Readout* (41). Donde el doctorando Prajan Divakar guía paso a paso desde el inicio hasta el fin los pasos y etapas del análisis mediante dicho software. Seguidamente, se ha realizado el tutorial de prueba/demostración de Nanostring University, que se titula *GeoMx™ Demo Report: Diabetic Kidney Disease y SOP* accesible desde su página oficial. Por último, se ha participado en una sesión en directo retransmitida por el Laboratorio de Experiencia del Cliente en Ámsterdam: Proteómica Espacial de Alta Multiplexación con GeoMx™ DSP. Una sesión organizada también por Nanostring University donde se muestra cómo realizar un experimento mediante GeoMx™ DSP y se ejemplifica la implementación mediante un caso real basado en el estudio de las malformaciones de los vasos sanguíneos.

3.2.2. Familiarización con GeoMxTools de Nanostring

Nanostring aparte de ofrecer el software mencionado en el apartado anterior ha creado un paquete llamado GeoMxTools, que forma parte de la librería Bioconductor (42) y que permite analizar los datos en R. Este análisis se basa en el software *Analysis Suite,* pero se diferencia por desarrollarse en el lenguaje de programación R, permitiendo un mayor ajuste de los parámetros y la interoperabilidad con otras librerías para enriquecer el análisis. Se ha realizado la ejecución del *pipeline Analyzing GeoMx-NGS RNA Expression Data with GeomxTools* (43) usando los datos de *Kidney_Dataset* de Nanostring (44).

3.2.3. Familiarización con StandR

Tal y como se menciona en el apartado 2.3.3 (Métodos Alternativos) cada vez hay más consciencia sobre las limitaciones del *pipeline* de GeoMxTools, sobre todo en la normalización y en la capacidad de minimizar los sesgos técnicos. Por este motivo, se ha elegido familiarizarse con StandR en este trabajo, ejecutando el *pipeline Analysing Nanostring's GeoMx™ transcriptomics data using StandR, limma and vissE* (45) usando los datos de *Lymph_Node_Dataset* de Nanostring (46). Este *pipeline* cuenta con otro tipo de infraestructura que permite la interoperabilidad con otros *pipelines* de RNA-seq como limma (47) o edgeR (48).

3.2.4. Construcción del diagrama del *pipeline*

Para el desarrollo gráfico del diagrama del *pipeline* se ha utilizado la herramienta LucidChart de libre acceso (49). El esquema está basado en el apartado 2.2.1 y 2.2.2 de este trabajo, así como con la familiarización y comparación de los *pipelines* de GeoMxTools y StandR, para elegir qué metodología se implanta en cada fase del análisis. Se ha considerado como punto de partida tres ficheros: las matrices de conteo de datos crudos, las anotaciones de los genes y las anotaciones de las ROIs. De esta forma, se empieza el análisis de los datos con la asociación entre transcrito/sonda y gen, así como la asociación entre muestra/paciente y ROI.

3.2.5. Implementación del proyecto en *Workflowr*

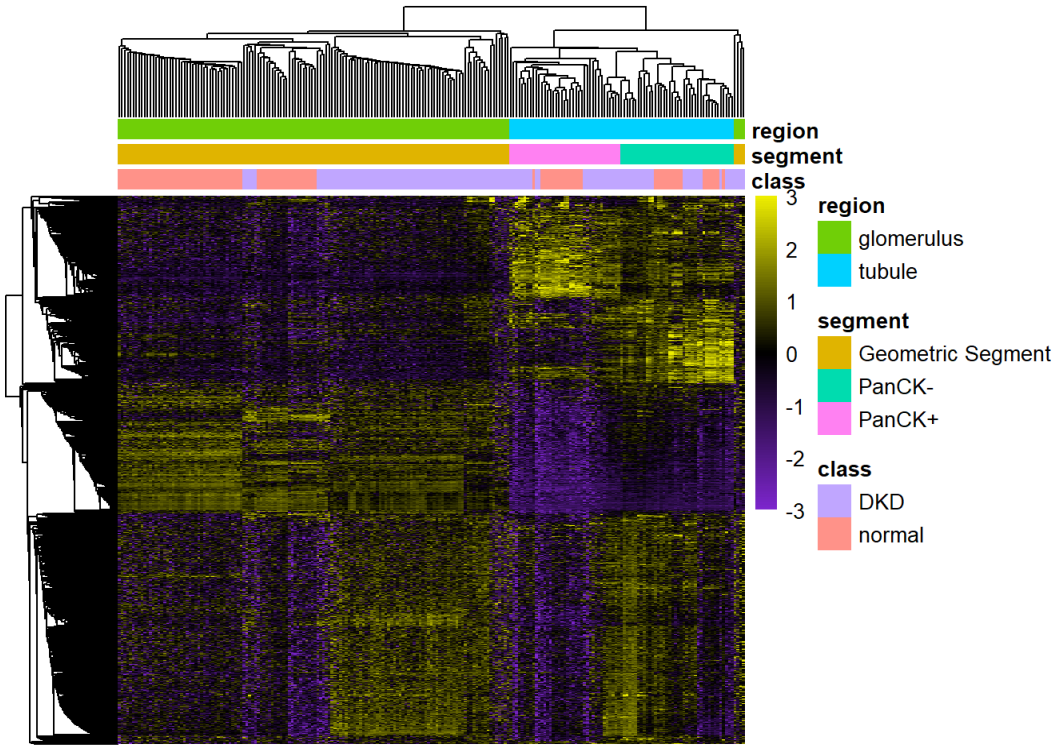
Workflowr es un paquete de R que ofrece a los investigadores la opción de organizar sus estudios de una manera que garantiza la gestión eficaz de proyectos, la reproducibilidad, la colaboración y el intercambio de resultados (32). Se basa en la construcción de un sitio web que contiene resultados con marcas de tiempo, versionados y documentados. Es una herramienta al alcance de cualquier usuario de R (32).

4. RESULTADOS

4.1. GeoMxTools

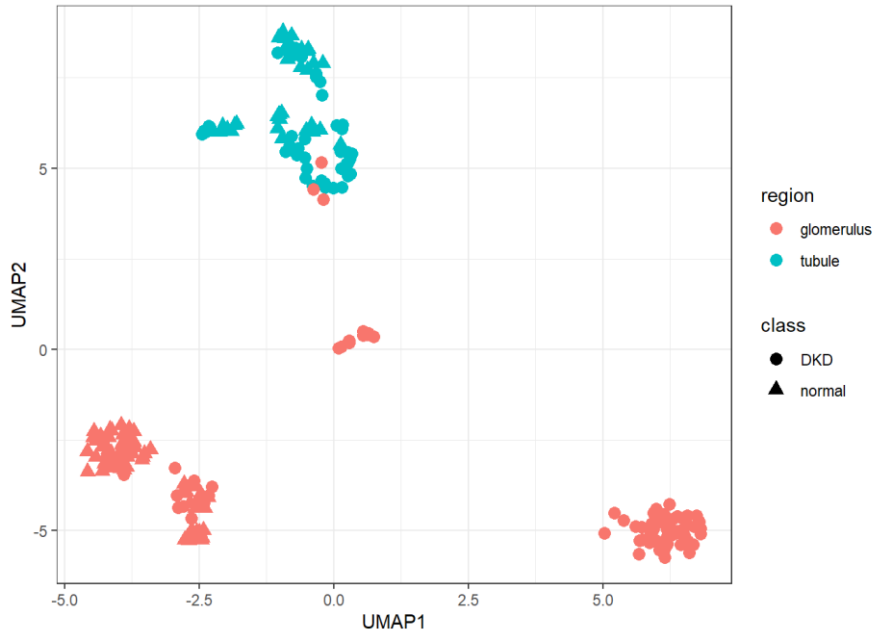
Se ha realizado la carga de las matrices de conteos, los metadatos y las anotaciones del *Kidney_Dataset* (44). Se ha hecho una primera visualización de los datos. El control de calidad se basa en una lista de parámetros establecidos (número de lecturas crudas, % Alineado, % Recortado, % Empalmado, saturación de secuenciación, conteo negativo, conteo de control sin plantilla, número de núcleos, área de segmento) con opción de ajuste. Se aplica el QC para cada ROI (donde se pasan de 235 ROIs iniciales a 229) y para cada gen (se pasan de 18642 genes iniciales a 18641). Se aplica el filtrado basado en LOQ (quedan finalmente 10131 genes y 221 ROIs). La normalización se lleva a cabo o por Q3 o por ruido de fondo. Se han realizado análisis posteriores y visualizaciones, incluyendo, análisis no supervisados como reducción de la dimensionalidad (*heatmap* (Figura 5), t-SNE, UMAP (Figura 6), y de los genes con un coeficiente de variación alto) para ver la agrupación de las células según su estructura (glomérulos y túbulos) y según su estado (enfermo o sano). Se ha ejecutado el análisis de la expresión diferencial, basado en el modelo lineal mixto y con un tiempo muy largo de compilación, tanto de las estructuras como del estado. Se han utilizado *volcano plots* para visualizarlos. También se han utilizado *violin plots* para graficar genes enriquecidos en estructura, denominados como genes de interés, y se ha realizado una comparación de ellos entre los pacientes sanos y enfermos. Para finalizar se ha realizado un *heatmap* de los genes con *p-value* y FDR más significativos. El documento HTML con todo detallado se encuentra en la bibliografía (50).

Figura 5. *Mapa de calor de expresión génica.*



Elaboración propia mediante la reproducción del pipeline GeoMxTools. Mapa de calor donde se muestran los perfiles de expresión de los genes estratificando por clase, segmento y región.

Figura 6. *Reducción de la dimensionalidad mediante UMAP.*

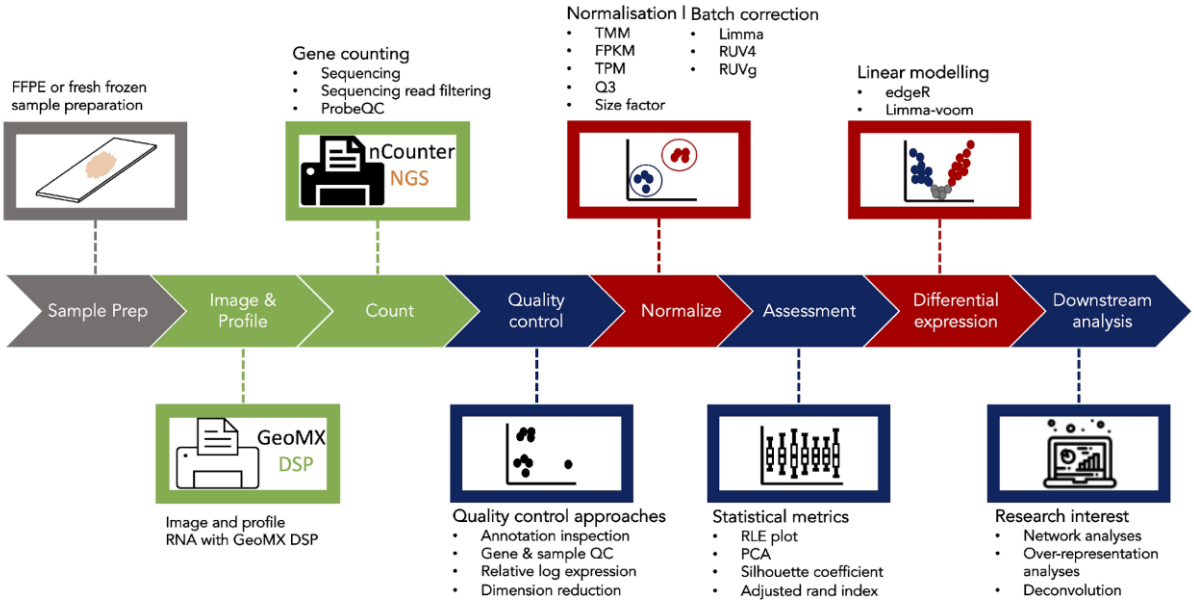


Elaboración propia mediante la reproducción del pipeline GeoMxTools. Se observa claramente la agrupación entre las distintas regiones, y dentro de cada región se intuyen también las distintas clases.

4.2. StandR

En la Figura 7 se recoge el esquema del flujo de trabajo del *pipeline* StandR donde se muestran los pasos des del procedimiento experimental (en verde) a los análisis de preprocesamiento y análisis posteriores (azul/rojo), juntamente con las herramientas utilizadas.

Figura 7. *Flujo de trabajo de StandR.*



Reproducido de GeoMXAnalysisWorkflow, Davis Laboratory. Esquema general del flujo de trabajo para el análisis transcriptómico espacial de StandR mediante GeoMx™ Digital Spatial Profiling (DSP).

Se ha realizado la carga de las matrices de conteos, los metadatos y las anotaciones del *Lymph_Node_Dataset* (46). Se ha hecho una primera visualización de los datos. El control de calidad de los genes, donde se detectan los genes que no superan el umbral mínimo de *counts*=5 y si no respetan este umbral en más del 90% de las ROIs (no se detecta ninguno). Para cada ROI se detectan los que tienen bajo número de librerías y número mínimo de 150 células (se detectan 11 por debajo del umbral) y se procede al filtrado donde se eliminan estas 11 ROIs. Para identificar si hay algún tipo de variación técnica y, por lo tanto, asegurarse de que se ha producido un buen filtrado, se realiza la distribución relativa de expresión logarítmica y la visualización de gráficos de reducción de dimensionalidad. Se ha implementado la normalización con el método TMM y corrección del efecto de lote con RUV4 y limma. Posteriormente, los análisis se realizan con el *pipeline* limma-voom (51), que permite

obtener una serie de exploraciones detalladas a continuación. Se genera la matriz de contraste para evaluar la expresión diferencial de los genes según el tipo celular, limma usa un modelo lineal y una corrección bayesiana empírica. Se calcula el coeficiente de variación biológica (CVB) para interpretar qué parte de estas diferencias son debidas únicamente al componente biológico y no al técnico, se realizan *MA plots* para la comparación de expresión de células B y células T, se realizan análisis de enriquecimiento de conjuntos de genes (GSEA) donde se grafican los 20 conjuntos de genes infra- y sobreexpresados. Este *pipeline* también cuenta con vissE, una herramienta versátil para identificar y visualizar fenotipos moleculares de alto nivel a partir del análisis de enriquecimiento funcional. Los últimos análisis que también engloba este *pipeline* es la deconvolución celular y el análisis de proporción diferencial de tipos celulares. El documento HTML se encuentra en la bibliografía (44).

4.3. Comparativa entre GeoMxTools y StandR

A partir de la familiarización de los dos *pipelines*, se ha elaborado una comparación de cada etapa, destacando las fortalezas/debilidades y diferencias/similitudes encontradas, que se tendrán en cuenta a la hora de elegir qué método se emplea en cada etapa del *pipeline* (Tabla 3). Esta comparación también está basada y respaldada con la revisión bibliográfica (29). Cabe destacar que, para realizar la comparación, se podría haber usado el mismo *dataset* en ambos *pipelines*, pero el objetivo de esta parte del trabajo era familiarizarse con las distintas metodologías, por eso se han usado los datos de demostración incluidos en las publicaciones de los *pipelines*. La comparativa entre ambos *pipelines* ya se realizó por científicos cuando lanzaron el proyecto de StandR y expusieron las mejoras que este conllevaba.

Tabla 3. *Comparativa entre viñetas, funciones, parámetros y alcance entre GeoMxTools y SandR (29).*

Etapa del análisis	GeoMxTools *Pipeline*	StandR *Pipeline*	Mejoras que aporta StandR
Control de calidad (QC)	Control de calidad con parámetros establecidos para cada ROI y para cada sonda. Cálculo de LOQ para cada ROI.	Control de calidad de genes, se calcula para cada gen si tiene mínimo x *counts*en el x% de las ROIs. Control de calidad de las ROIs según el tamaño de biblioteca y número de núcleos.	Tienen en común que se detectan los genes y ROIs que no superan un umbral establecido. El QC de StandR tiene mayor fiabilidad ya que posee un filtro adicional para las ROI de baja calidad (el núm de células y el tamaño de librerías).

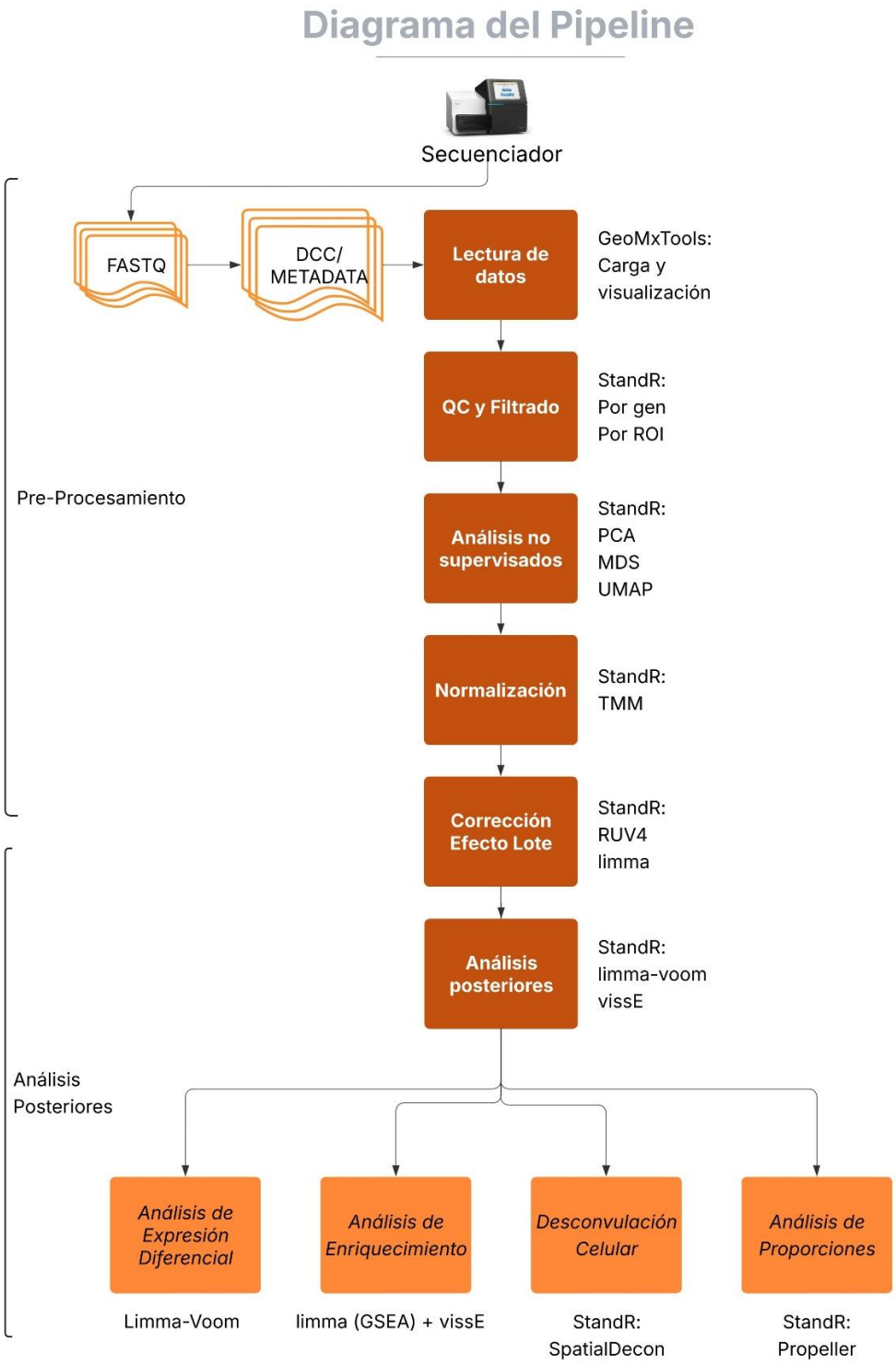
Filtrado de genes	Elimina los genes con expresión < LOQ y sin presencia mínima en el x% de ROIs.	Elimina los genes y ROIs que no pasen el control de calidad anterior. Para las ROIs de baja calidad que no han sido detectados, se realizan gráficos de reducción de dimensionalidad donde aparecerán como clústers aislados.	Filtra menos genes, reduce el riesgo de eliminar información biológica valiosa. Elimina outliers y no genes relevantes del tejido o con expresión media. (26)
Normalización	Normalización Q3 o de fondo.	Normalización con TMM, RPKM, TPM, CPM, cuartil superior o por factor. Visualizaciones de las variaciones técnicas mediante RLE o PCA.	Permite ajustar mejor los datos preservando las variaciones biológicas y minimizando las técnicas.
Corrección de efectos de lote	No se aplica.	Usa métodos estadísticos como RUV4 o limma y una posterior evaluación con una puntuación para cada método.	Es único de StandR. Reduce la variabilidad técnica entre *slides* y preserva la variabilidad biológica.
Modelado estadístico para el análisis DE	T-Test o Modelo Lineal Mixto.	Modelos lineales avanzados y corrección bayesiana empírica (limma-voom o edgeR).	Permite que se tenga en cuenta el efecto de lote, la varianza técnica y obtener estadísticos más robustos. Encuentra un número mayor de genes y de colección de genes con expresión diferencial relevante. (26)
Visualización y diagnóstico	Reducción de la dimensionalidad, mapa de calor para visualizar expresión génica, análisis de expresión diferencial entre estructuras celulares y entre estados (enfermo/sano).	PCA, RLE plots, heatmaps y métricas de varianza, análisis de expresión diferencial, análisis de enriquecimiento de genes, vissE para visualizar fenotipos, deconvolución celular y análisis de proporciones celulares.	Ambas ofrecen gráficos parecidos para la visualización de genes de DE. Aun así, StandR ofrece un número mayor de análisis, por lo tanto, una mayor comprensión de los datos.

Elaboración propia.

4.4. DIAGRAMA DEL *PIPELINE*

Se elabora un esquema de los pasos que contendrá el *pipeline* desarrollado, juntamente con las herramientas que se aplican a cada paso, recogido en la Figura 8.

Figura 8. *Esquema de la estructura del pipeline desarrollado.*



Elaboración propia mediante la herramienta LucidChart.

4.5. RESULTADOS DEL *PIPELINE* APLICADO A UN *DATASET*

El *pipeline* construido consta de cinco fases principales, representadas en naranja oscuro en la Figura 8, cuyo código ha sido desarrollado en un documento en formato RMarkdown (.Rmd) y se encuentra recogido en el archivo PIPELINE.html (52). A lo largo de este capítulo, se describen los aspectos más importantes de cada fase, como las funciones y paquetes más relevantes, o las decisiones metodológicas tomadas durante el desarrollo. Adicionalmente, acompañando a las explicaciones, se han incluido algunas figuras resultantes de la ejecución del *pipeline* con el *dataset* elegido, a modo de representación gráfica de los pasos más importantes del proceso.

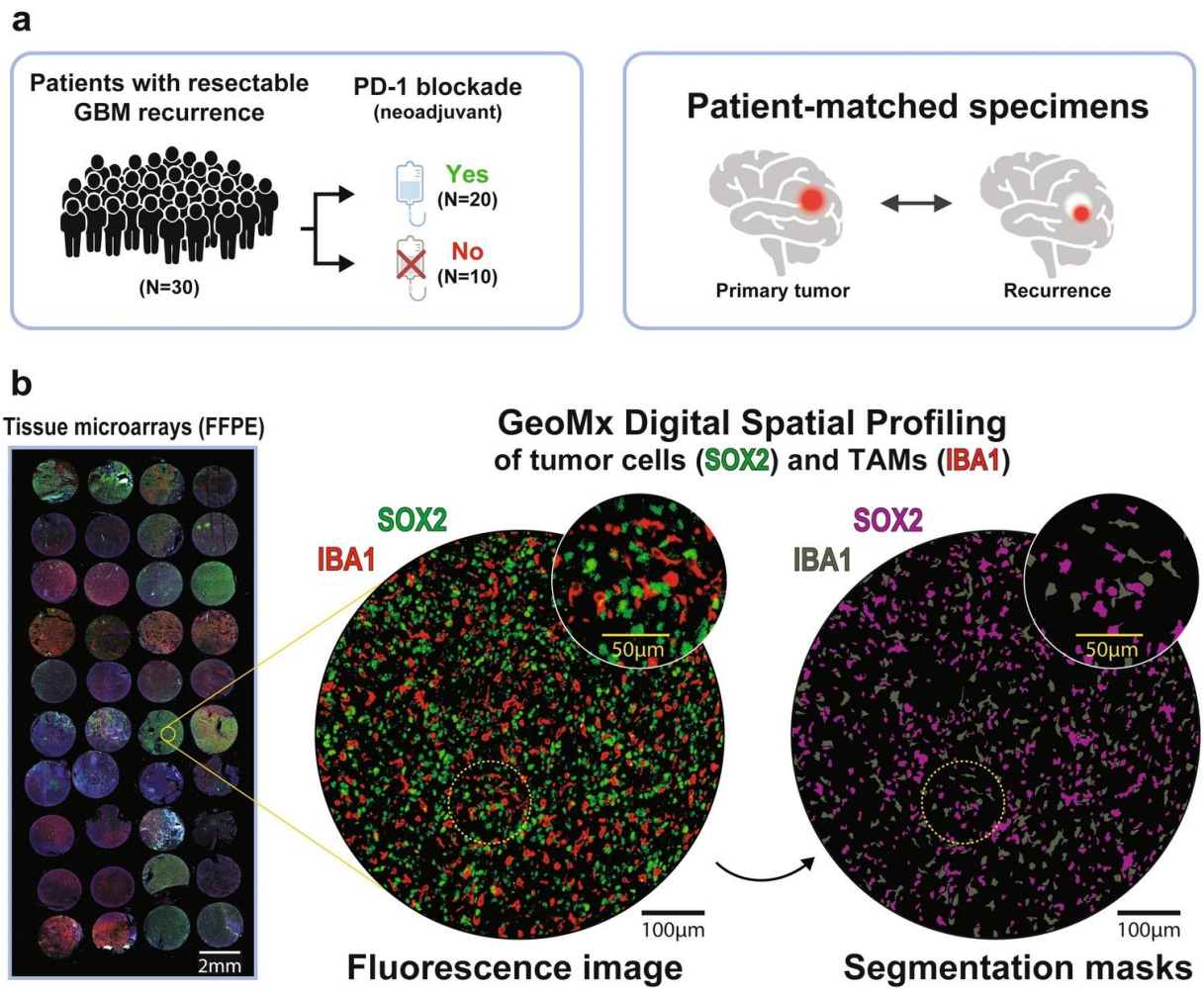
4.5.1. Lectura de los datos

El glioblastoma (GBM) es el tumor cerebral primario maligno más agresivo y común entre la población adulta. Tiene un crecimiento muy rápido y presenta resistencia a las terapias estándar, por lo que en la mayoría de los casos los pacientes experimentan recurrencia del tumor. La inmunoterapia ha tenido un gran éxito en el tratamiento de varios tipos de cáncer, impulsando respuestas inmunitarias antitumorales. Un ejemplo es el bloqueo entre el PD-1 y su ligando PD-1L. Cuando se bloquea esta interacción, los inhibidores de PD-1 restauran la actividad de las células T, permitiendo que las células T citotóxicas eliminen las células tumorales.

El objetivo del estudio es utilizar la técnica GeoMx™ DSP para evaluar el efecto del bloqueo de PD-1 en células tumorales y *Tumor Associated Macrophages* (TAM), las dos poblaciones celulares más abundantes de los Glioblastomas (GBM).

Se seleccionan 30 pacientes con GBM recurrente, de los cuales, 20 pacientes reciben Nivolumab el neoadyuvante que bloquea la interacción PD-1 con su ligando, y 10 pacientes son controles no tratados. Se seleccionan zonas con alta densidad de SOX2+, que son indicadores de células tumorales, y de IBA1+, que son indicadores de TAM, como las AOIs de este experimento (Figura 9).

Figura 9. *Planteamiento del proceso experimental de la base de datos.*



Reproducido de S.B. Artzi (39). Análisis transcriptómico espacial de células tumorales y macrófagos/microglía asociados al tumor (TAMs) en GBM recurrente con o sin bloqueo neoadyuvante de PD-1. a) Descripción general de las muestras de tejido analizadas en este estudio. b) Imagen de fluorescencia representativa obtenida mediante perfilado espacial digital (GeoMx) de un microarreglo tisular FFPE, que muestra células tumorales SOX2+ (verde) y TAMs IBA1+ (rojo). Se aplicaron máscaras de segmentación para definir las regiones tumorales e inmunes basándose en marcadores morfológicos, lo que permitió el perfilado transcriptómico espacial de poblaciones celulares distintas. Barras de escala: negro, 100 µm; amarillo, 50 µm.

La fase de lectura de datos está diseñada para importar los archivos planos generados por GeoMx™ DSP, realizar una exploración y preprocesamiento iniciales, y crear un objeto del tipo `NanostringGeoMxSet` a partir de ellos. Como entrada, se requieren tres archivos: la matriz de expresión (.dccFiles), los metadatos de los genes (.pkcFiles) y los metadatos de las AOIs (phenoDataFile), que mediante la función de `readNanosStringGeoMxSet()` se guardan en el objeto de trabajo.

Accedemos a los datos en Zenodo y descargamos y guardamos los archivos en nuestro directorio de trabajo:

- RAW.tar.gz: archivos dcc (GeoMx™ DSP output) de todas las AOIs (i.e., raw data)
- metadata.csv.tar.gz: archivo separado por comas de las AOIs utilizados en este análisis.
- sampleAnno.xlsx.tar.gz: archivo Excel que contiene la anotación del fichero dcc.

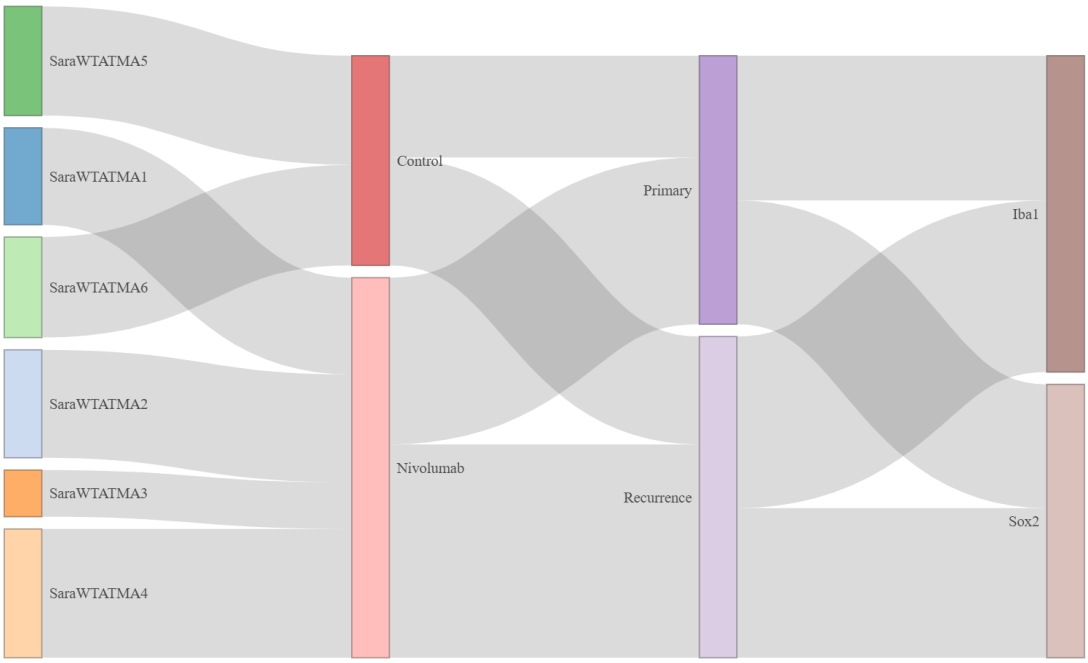
Se empieza el análisis con 481 AOI y 80 variables de ambos metadatos.

Como se utilizan mayoritariamente funciones de StandR, es necesario cambiar la clase de los datos de un objeto `NanostringGeoMxSet` a un objeto `SpatialExperiment` para poder aplicarle las funciones de StandR.

En los Anexos se ha detallado una tabla con la función de conversión entre objetos, así como las funciones para manipular objetos de `NanostringGeoMxSet` y sus correspondencias para objetos de clase `SpatialExperiment`.

Se genera un diagrama de Sankey, ilustrado en la Figura 10, para visualizar el flujo de muestras que pertenecen a cada condición y entender qué anotaciones caracterizan este experimento.

Figura 10. *Diagrama de Sankey.*



Elaboración propia. Distribución de las muestras (SaraWTATMA1-6), la clase de tratamiento (Control/Nivolumab), el tipo de tumor (Primary/ Recurrence) y la zona celular (Iba1/Sox2). Disponible la descarga de la versión interactiva en: https://github.com/carmedelicado/GeoMx_Pipeline_CDM/blob/main/sankey.html

4.5.2. QC y filtrado

El control de calidad en StandR se realiza a dos niveles, a nivel de gen y a nivel de ROI.

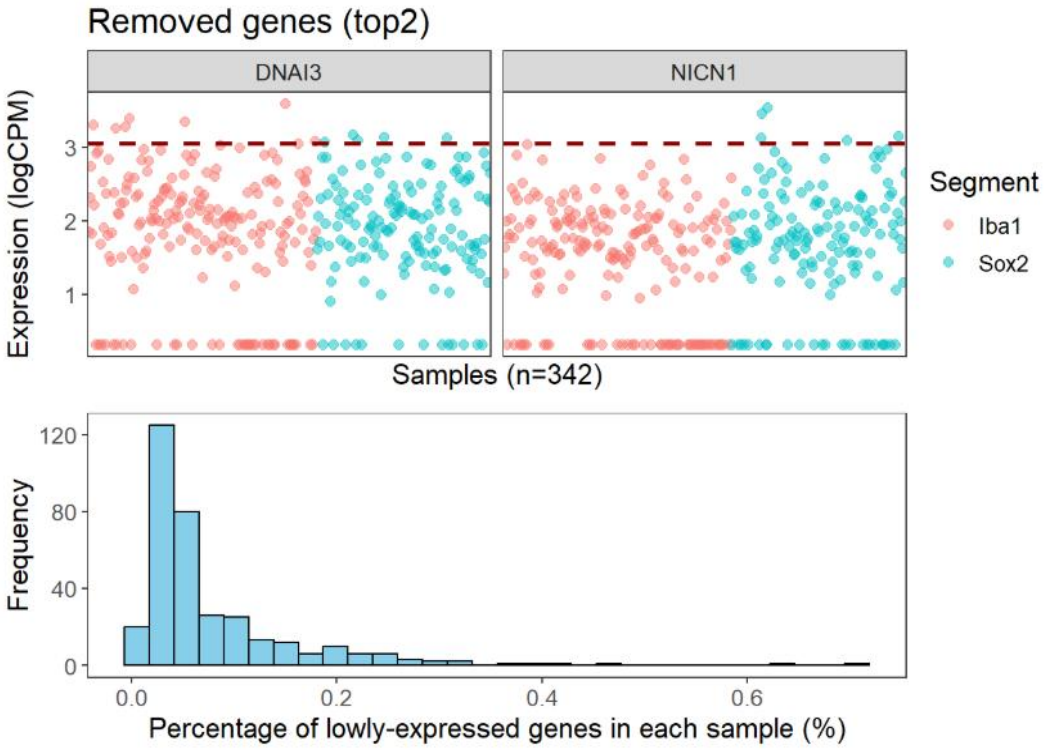
Antes de empezar con el QC se debe comprobar la columna *QCFlags* que indica las muestras de tejido de mala calidad según su paso preliminar de control de calidad. Si los datos superan correctamente su QC habrá valores *None Avaliable* (NA) o celdas vacías en esta columna.

Se filtran las muestras que contienen algún problema de calidad detectado por el secuenciador obteniendo 18677 genes y 342 AOI para empezar.

4.5.2.1. QC gen

El control de calidad a nivel de gen tiene el objetivo de eliminar los genes que no se expresan en más del 90% de las AOIs e identificar AOIs con muy pocos genes expresados. Se calcula para cada gen si tiene como mínimo 5 *counts* en el 90% de las AOIs. Es una forma de revisar la calidad de cada fila de la matriz de conteos. Se obtienen 18675 genes a diferencia de los 18677 iniciales, por lo que se han eliminado 2 genes. Con la función `plotGeneQC()` se pueden evaluar las expresiones en *logCountPerMilion* (logCPM) de los genes que fueron eliminados en las distintas AOIs (Figura 11).

Figura 11. *Genes Filtrados.*



Elaboración propia.

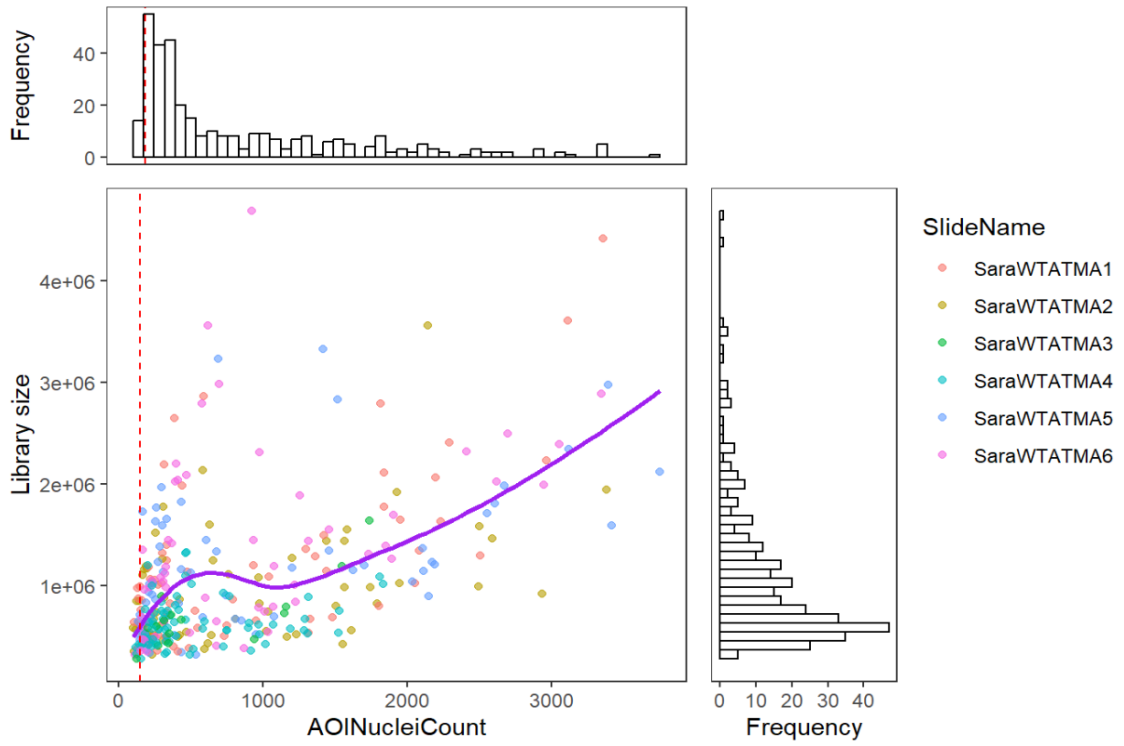
El gráfico superior muestra los genes eliminados, perteneciendo cada punto a la expresión de ese gen en una AOI (n=342). La mayoría de los puntos están por debajo del umbral de expresión (línea roja), lo que contribuye a que el gen sea eliminado. No se observa ninguna diferencia aparente entre los segmentos Iba1 y Sox2. El gráfico inferior muestra para todas las AOIs el porcentaje de genes que no tienen expresión. Se observa como la mayoría de AOIs se acumulan mayoritariamente entre el 0-0,2% indicando que la mayoría de los genes se están expresando.

4.5.2.2. QC ROI

El control de calidad a nivel de ROI, busca identificar ROIs con bajo tamaño de biblioteca y/o bajo recuento de células, ya que se consideran muestras de baja calidad por insuficiente profundidad de secuenciación o falta de ARN en la región seleccionada. Es una forma de analizar la calidad de las columnas de la matriz de conteos. Como en este experimento se han recogido los conteos por AOI, se realiza el control de calidad de las AOIs.

Se evalúa el gráfico de distribución del tamaño de la librería frente al número de núcleos en la Figura 12. Al observar el diagrama de dispersión, realizado mediante la función `plotROIQC()`, se espera que los tamaños de librería estén correlacionados de forma positiva con el número de células (es decir, con el recuento de núcleos). Para filtrar muestras de baja calidad, se define un umbral mínimo de recuento de células, en este caso 150 células. Se eliminan 20 AOIs que no superan el umbral mínimo. Además, se puede investigar si las AOIs de baja calidad provienen de alguna muestra específica, estratificando los puntos por nombre de lámina (SlideName).

Figura 12. *Relación entre tamaño de librería y número de células.*



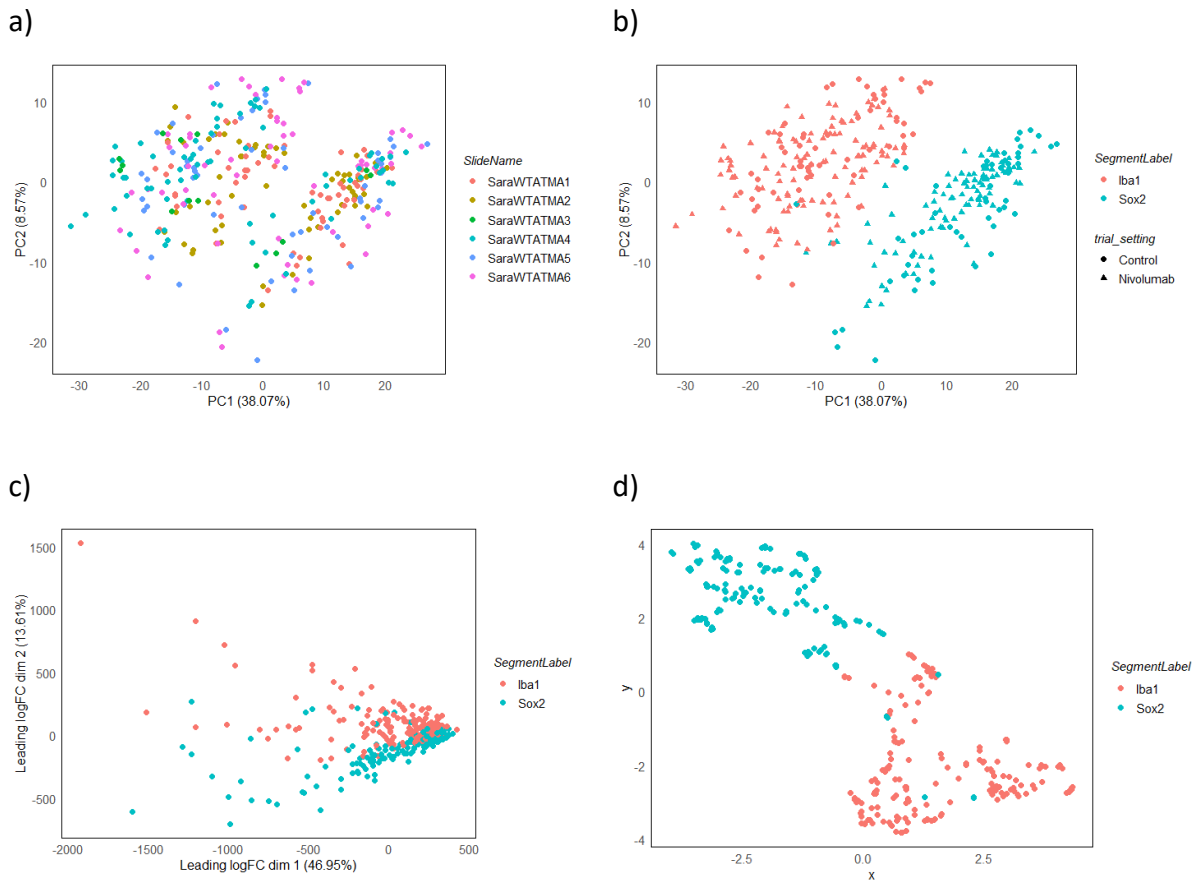
Elaboración propia.

En este *dataset*, la relación entre tamaño de biblioteca y recuento de células es relativamente estable, sin anomalías evidentes, exceptuando una pequeña desviación al inicio de la recta. Esto indica que los valores observados son más altos de lo esperado en librerías con bajo número de recuento de células. La línea roja marca el umbral de recuento mínimo de células. No se observa ninguna asociación entre AOIs de baja calidad y una muestra en específico.

4.5.3. Reducción de la dimensionalidad

Existen diferentes métodos para reducir la dimensionalidad de los datos, así como el PCA, UMAP o t-SNE. El PCA es un método de reducción de dimensionalidad utilizado para explorar conjuntos de datos de alta dimensionalidad en dos dimensiones, sin perder la información más relevante. Permite identificar agrupamientos de muestras que se asemejan biológicamente (clústeres) y visualizar posibles variaciones técnicas. En la Figura 13 se recogen diferentes gráficos en los que se han empleado algoritmos no supervisados para la reducción de la dimensionalidad.

Figura 13. *Métodos de reducción de la dimensionalidad.*



Elaboración propia. a) PCA estratificando por SlideName para evaluar si hay efecto de lote. b) PCA estratificando por el tipo de marcador (Iba1 o Sox2) y la clase de tratamiento (Nivolumab o Control*). c) Multidimensional Scaling (MDS) estratificando por tipo de marcador. d) Uniform Manifold Approximation and Projection (UMAP) estratificando por marcador.*

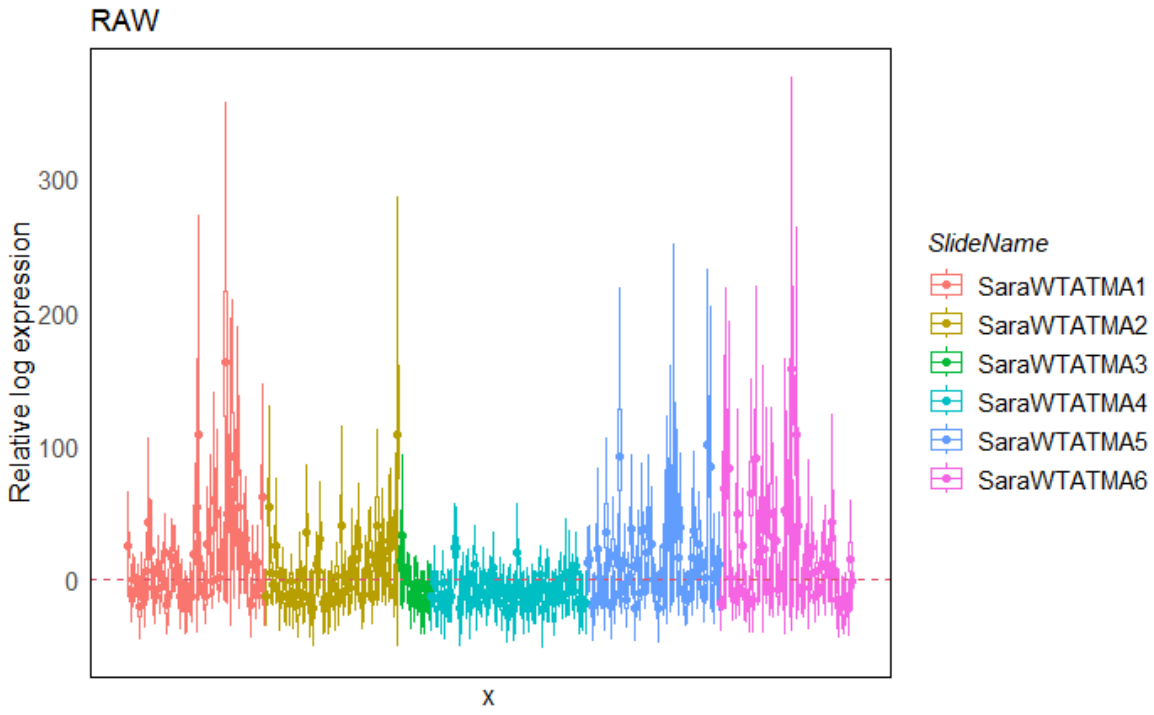
El análisis de componentes principales (PCA), así como los otros métodos de reducción de la dimensionalidad, revela que la principal fuente de variación en los datos (PC1: 38.07%) está

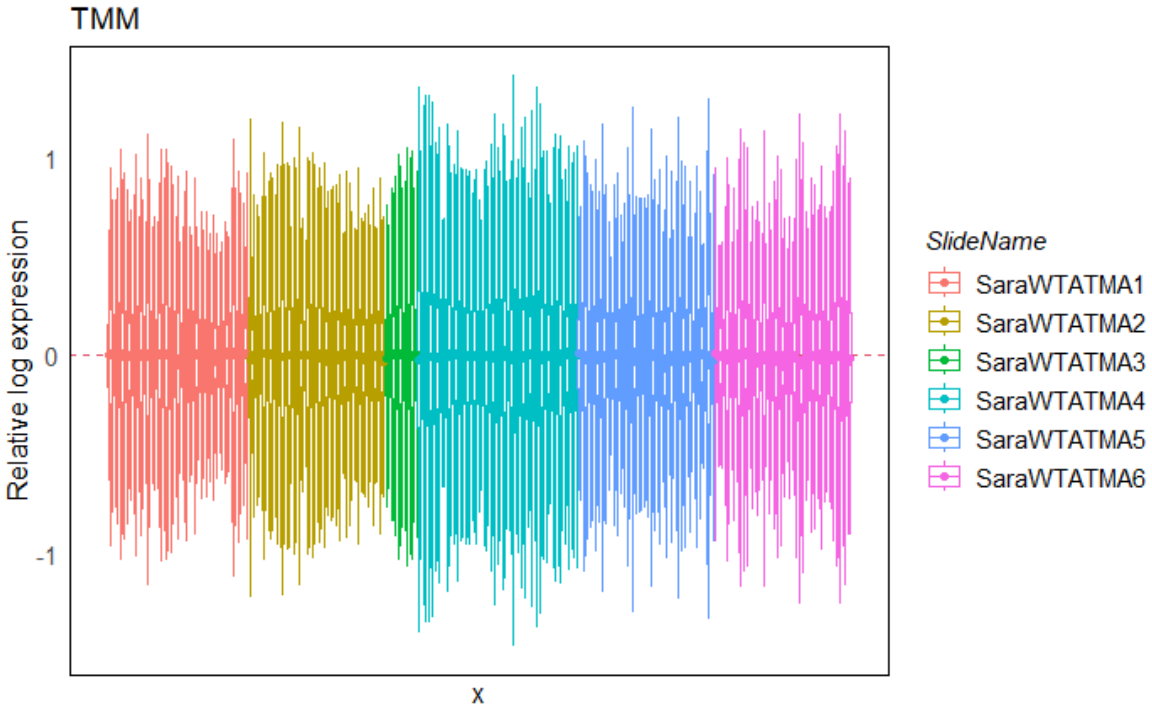
impulsada por el tipo de segmento (Iba1 vs Sox2). No se observa una separación clara basada en el tratamiento (Nivolumab vs Control), ni un efecto de lote significativo por portaobjetos (SlideName), lo cual valida la consistencia biológica de las muestras.

4.5.4. Normalización

Si se identifican variaciones técnicas en los pasos de QC, antes de realizar cualquier análisis posterior es necesario normalizar los datos para minimizar dichas variaciones. Se normaliza por el método *Trimmed Mean of M-values* (TMM) mediante la función `geomxNorm()`. Para evaluar la efectividad de la normalización, se utiliza el gráfico *Relative Log Expression*, que muestra la desviación de la expresión de cada gen en una AOI respecto a la mediana de ese gen en el conjunto de muestras (Figura 14).

Figura 14. *Normalización de los datos.*





Elaboración propia. RLE de los datos crudos i de los datos normalizados con el método TMM. La normalización produce una reducción en la variabilidad aportada por el tamaño de librería, centralizando la mediana y reduciendo la varianza de las distintas muestras.

Se observa como la normalización aporta una gran mejora acercando las medianas de las AOIs a 0, sugiriendo que las de variaciones técnicas han sido eliminadas.

4.5.5. Corrección del efecto lote

Cada lámina generalmente solo puede contener unos pocos segmentos de tejido, por lo que es común que los datos de GeoMx™ DSP estén afectados por el efecto de lote introducido por las diferentes láminas. StandR ofrece dos enfoques para corregir efectos de lote: RUV4 y limma con la función `geomBathCorrection()`. Para la corrección con RUV4, la función requiere de 3 parámetros adicionales además del objeto de entrada:

- *factors*: el factor de interés, es decir, la variación biológica que se desea conservar. En este caso es la variable: tipo de segmento (SegmentLabel).

- *NCGs*: la lista de genes de control negativo identificados con la función `findNCG()`, se seleccionan los 300 genes menos variables (ordenados por coeficiente de variación) entre las distintas láminas como NCGs.

- *k*: el número de factores no deseados a utilizar. Según la documentación de RUV, se recomienda usar la *k* más pequeña posible que elimine la variación técnica observada, tras una exploración se define *k*=2 para este experimento.

Para la corrección con limma, la función requiere 2 parámetros adicionales además del objeto de entrada.

- *batch*: un vector que indica la información de lote de todas las muestras.

- *design*: una matriz de diseño generada a partir de `model.matrix()`.

Se recomienda utilizar estadísticas resumidas para evaluar la efectividad de la corrección por lote y elegir qué método ha sido más efectivo para este *dataset*.

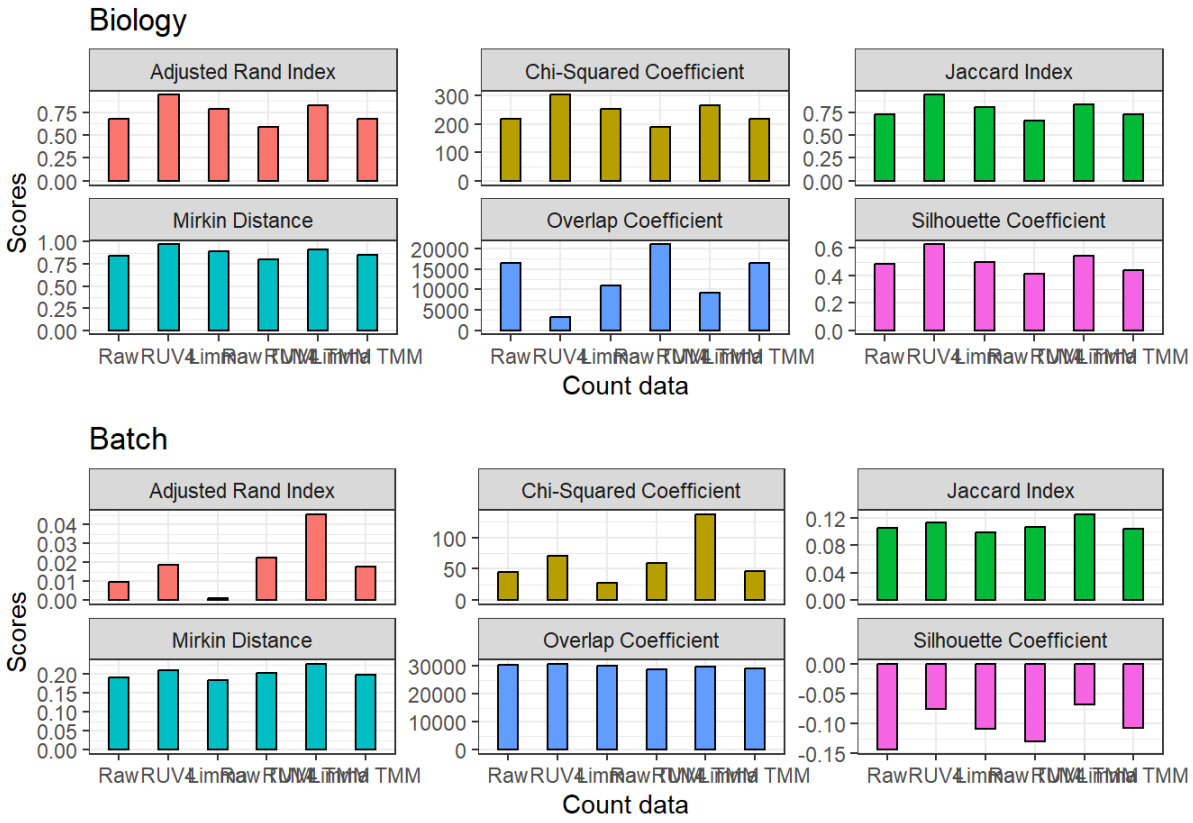
Las 6 estadísticas resumidas, incluidas en el paquete mediante la función `plotClusterEvalStats()`, son conocidas como Métricas de Evaluación de Clústeres:

- *Adjusted Rand Index* (ARI): Mide la concordancia entre los clústeres del PCA y las etiquetas reales (SegmentLabel), ajustando el valor por el azar (donde 1 es una coincidencia perfecta).
- Coeficiente de similitud de Jaccard: Evalúa qué tan similares son los conjuntos de muestras en cada clúster, calculando la proporción de coincidencias sobre el total de elementos comparados.
- Coeficiente de Silhouette: Mide qué tan cerca está cada muestra de su propio grupo en comparación con otros grupos; valores altos indican una asignación de clúster clara y robusta.
- Coeficiente Chi-cuadrado: Evalúa la independencia estadística entre los clústeres obtenidos y las variables de lote o segmento, detectando si existe un sesgo sistemático.
- Distancia de Mirkin: Cuantifica la diferencia o distancia entre dos clasificaciones; a menor valor, mayor es la similitud entre el agrupamiento del PCA y la biología esperada.

- Coeficiente de solapamiento: Mide el grado de intersección entre los grupos, siendo muy útil para comprobar si las muestras de diferentes lotes se han integrado correctamente en un mismo espacio.

Para los factores biológicos de interés, es mejor un puntaje más alto en la corrección. Para los efectos de lote, un puntaje más bajo es preferible en la corrección. Se presentan los resultados de las estadísticas resumidas para los dos métodos de corrección del efecto de lote utilizados en este pipeline (RUV4 y limma), así como para los datos crudos (tres primeras barras) y normalizados con TMM (tres últimas barras) en la Figura 15. Las barras representan el conjunto de datos y se exponen en el orden siguiente: datos_spe, datos_ruv, datos_lrb, datos_tmm, datos_ruv_norm, datos_lrb_norm.

Figura 15. *Comparativa entre métodos de eliminación del efecto lote.*



Elaboración propia.

En este conjunto de datos, se identifica RUV4 en los datos sin normalizar, como la herramienta más potente para preservar el carácter biológico de las muestras y limma en los datos normalizados, como la mejor herramienta para minimizar el ruido técnico. En este punto,

elegir la metodología adecuada para la corrección, viene influido por el objetivo del experimento, diferenciando en si la prioridad es la precisión biológica (datos_ruv), o la integración perfecta de las muestras (datos_lrb_norm).

Se elije el método RUV4 (datos_ruv) para continuar con los análisis, priorizando la precisión biológica.

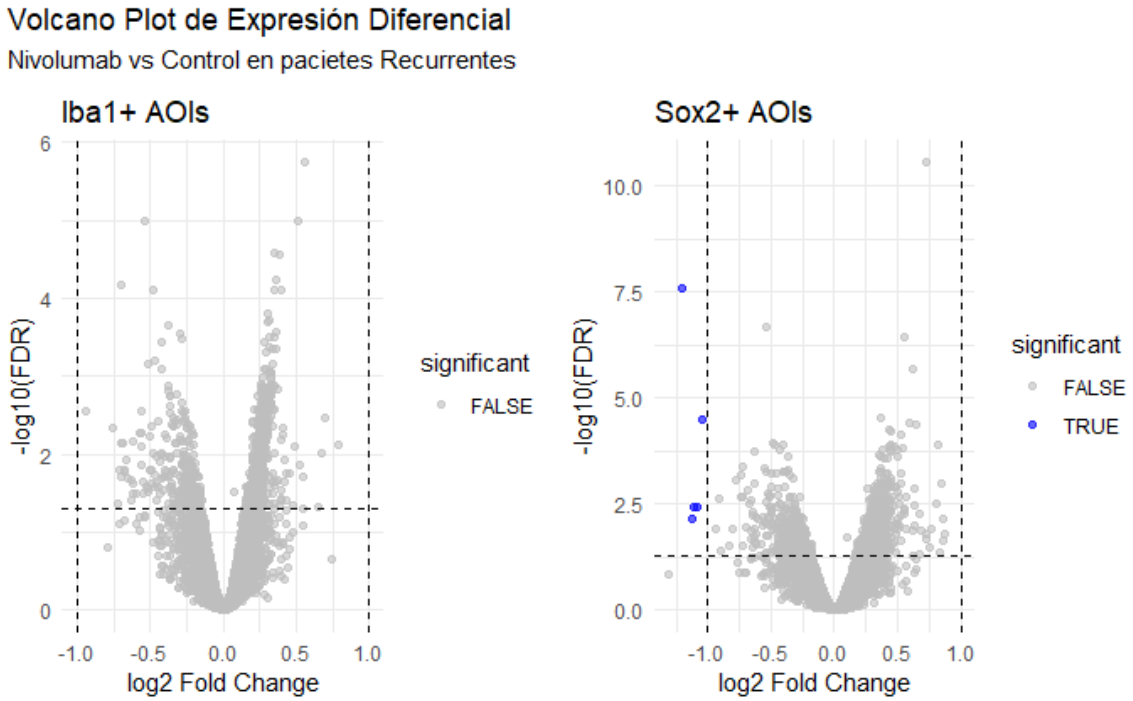
4.5.6. Análisis posteriores

4.5.6.1. Análisis de expresión diferencial (DE)

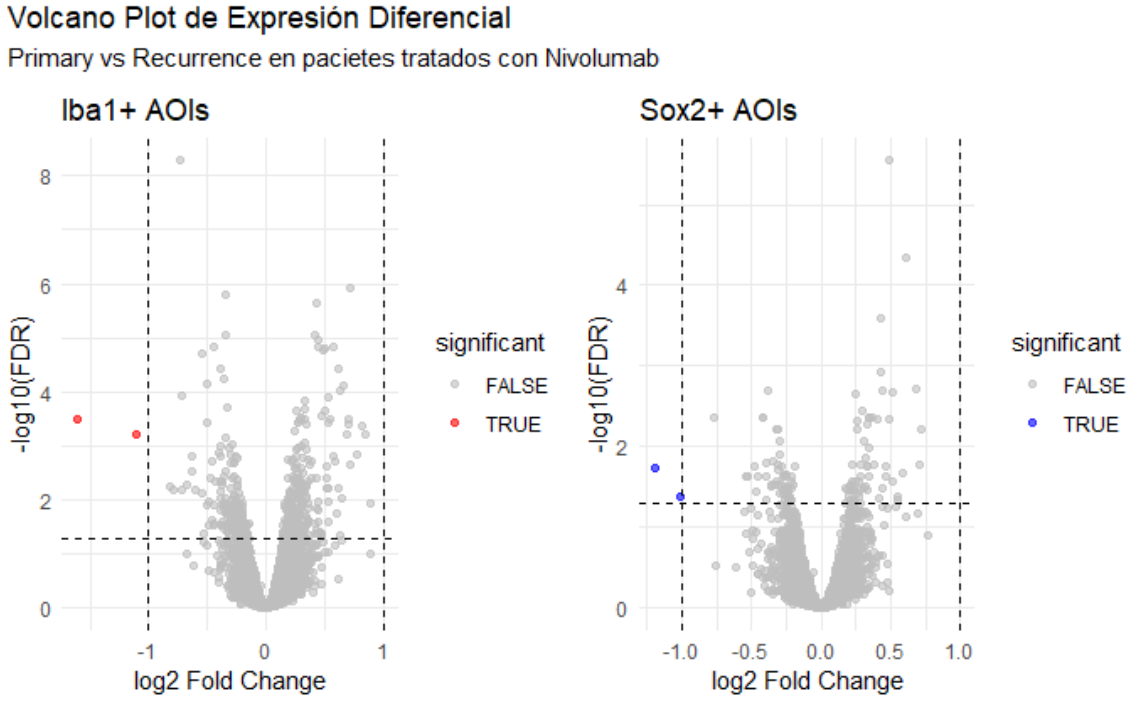
En este *pipeline*, se desarrollará el análisis DE usando el *pipeline* limma-voom, que se caracteriza por evaluar un modelado lineal realizado sobre los valores log-CPM utilizando las funciones voom, lmFit, contrasts.fit y eBayes. Voom transforma los datos y estima pesos, limma ajusta el modelo lineal estimando los coeficientes y eBayes estabiliza la varianza y calcula los estadísticos finales. Posteriormente, se clasifican los genes según si están sobreexpresados o subexpresados en las condiciones que se contrastan. Se construye una matriz de diseño con la variable "grupo" que se compone de las 3 variables agrupadas (Segment_Label, trial_setting, tummor_setting), usando los datos crudos con la corrección del efecto lote con Ruv4. De esta forma, se obtienen 8 niveles de "grupo", que corresponden a las distintas anotaciones y se evalúa la efectividad del tratamiento y su impacto en tumores primarios y recurrentes (Figura 16).

Figura 16. *Volcano Plot para el análisis de expresión diferencial.*

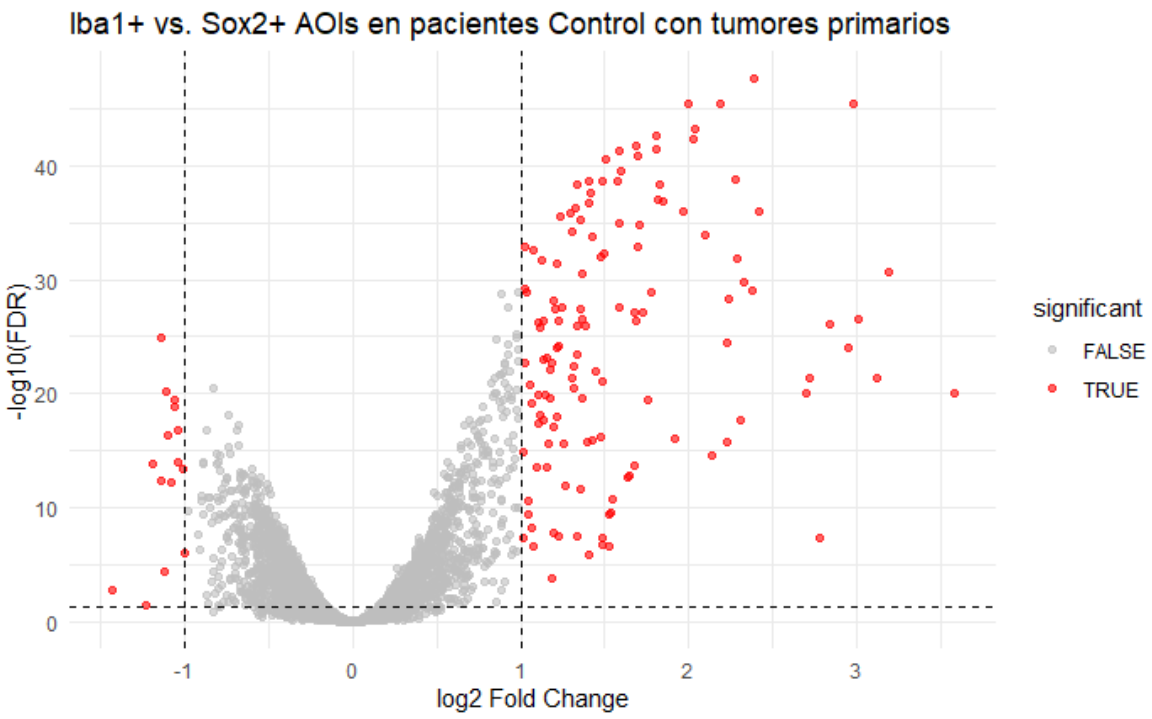
a)



b)



c)



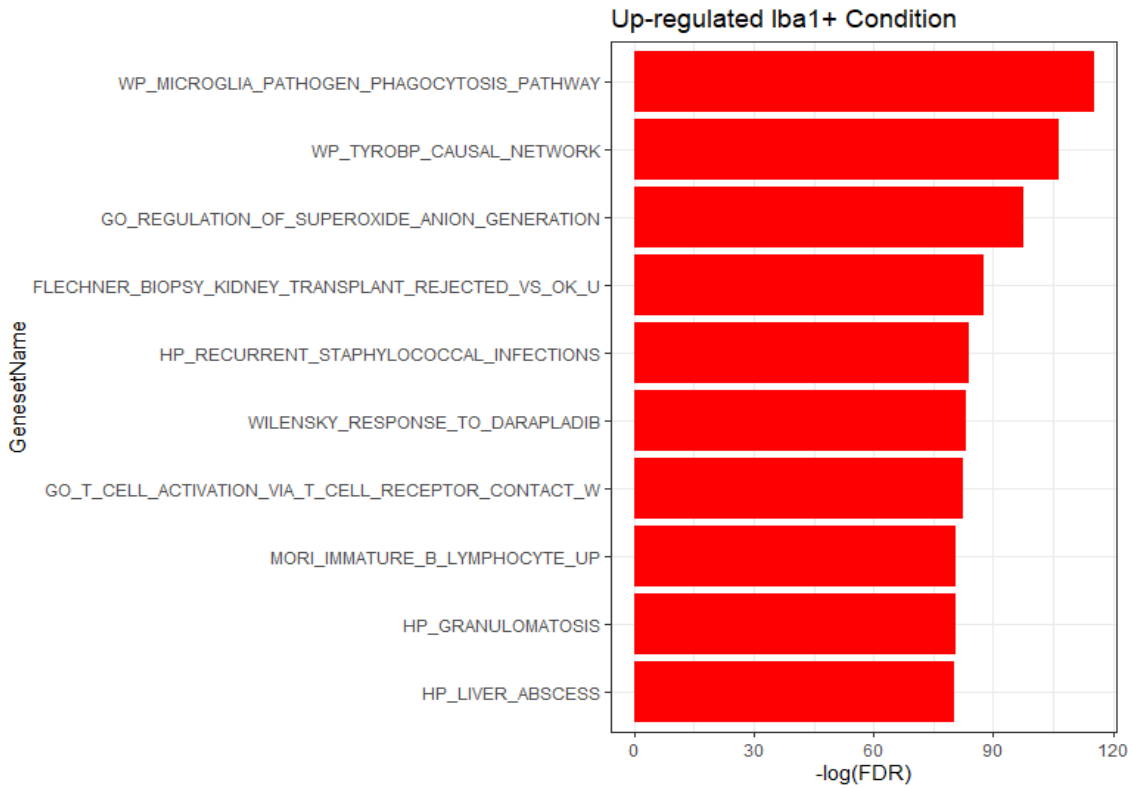
Elaboración propia. a) Nivolumab vs Control en pacientes Recurrentes. b) Primary vs Recurrence en pacientes tratados con Nivolumab. c) Iba1+ vs. Sox2+ en pacientes Control con tumores primarios.

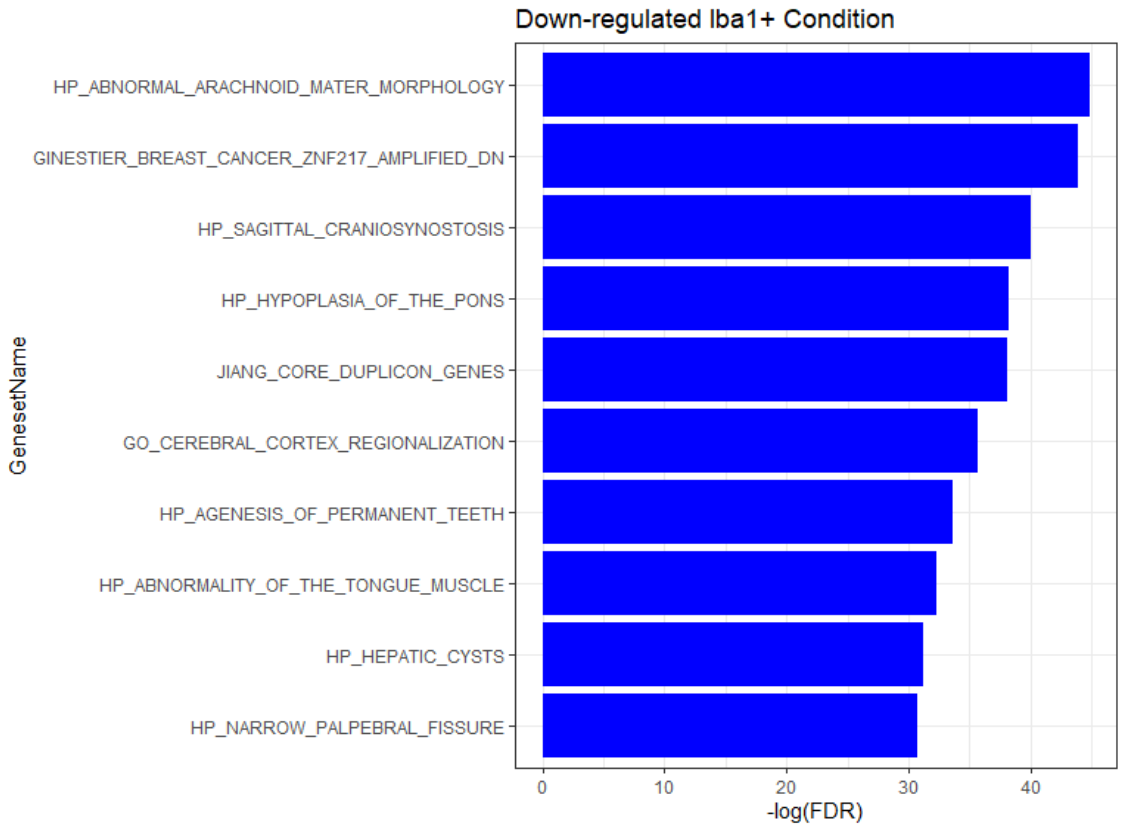
No se observan diferencias significativas entre la expresión génica de los pacientes recurrentes tratados con Nivolumab o pacientes recurrentes Control, en ninguna de las dos zonas celulares analizadas. Esto indica que el tratamiento no está siendo efectivo, ya que no se está desarrollando una activación de la respuesta inmunitaria, que cambiaría el perfil de expresión genética de ambas zonas celulares. Tampoco hay diferencias significativas entre los pacientes con tumores primarios y recurrentes tratados con Nivolumab en ninguna de las dos zonas celulares. Esto demuestra que el tratamiento no es más efectivo en pacientes con tumores primarios, hecho que desempeña de nuevo el pensamiento de que el tratamiento no es efectivo. Finalmente, se decide realizar la comparación entre las zonas celulares Iba1 y Sox2. Para ello, se construye una nueva matriz que contrasta los grupos Iba1_Control_Primary vs Sox2_Control_Primary, intentando representar condiciones “vírgenes”. El contraste Iba1+ vs. Sox2+ muestra una asimetría hacia el enriquecimiento de genes en la población mieloide (Iba1), con una significancia estadística extremadamente elevada (FDR <).

4.5.6.2. Análisis de enriquecimiento

El análisis de enriquecimiento de conjuntos de genes (GSEA) es un método bioinformático que permite determinar si un conjunto de genes de un *dataset* está relacionado y muestra cambios de forma coordinada en una condición concreta. En este *pipeline* se utiliza el método GSEA del paquete limma que utiliza la función `fry()`. Los resultados se encuentran en la Figura 17, que muestra las 10 vías metabólicas sobreexpresadas en Iba1 y Sox2 para pacientes Control con tumores primarios.

Figura 17. *Gráfico del top 10 conjuntos de genes Up y Down - Regulated en Iba1.*





Elaboración propia.

Los conjuntos sobreexpresadsos en Iba1 muestran una significancia estadística extrema (-log(FDR) > 100), lo que indica una señal biológica masiva y coherente con el linaje de macrófagos/microglía. Se destaca la fagocitosis de los patógenos, que es el conjunto de genes más robusto y que en un tumor indica que la microglía está intentando limpiar el entorno tumoral. La red Tyrobp que describe una activación profunda de las células y la regulación del superóxido (ROS), que implica actividad efectora proinflamatoria.

Los conjuntos de genes *down-regulated* son los que se sobreexpresan en Sox2 indicadores de células tumorales. Es coherente lo que se observa ya que se asocian a una anomalía clínica donde se modifican las estructuras del tejido cerebral, también a la proliferación celular maligna y a genes en estados menos diferenciados.

Los resultados son coherentes porque coinciden con los observados en el *volcano plot*. Indagando en las funciones de las vías, los genes “Up” definen funciones efectoras inmunes (Iba1), mientras que los genes “Down” definen la identidad del linaje neural y la progresión

tumoral (Sox2). Esto confirma que se logra capturar con éxito la interacción entre el tumor y la respuesta inmunitaria del paciente.

4.5.6.3. Análisis de deconvolución celular

La deconvolución celular es una técnica que permite estimar las proporciones de los diferentes tipos celulares que se encuentran en una muestra de tejido.

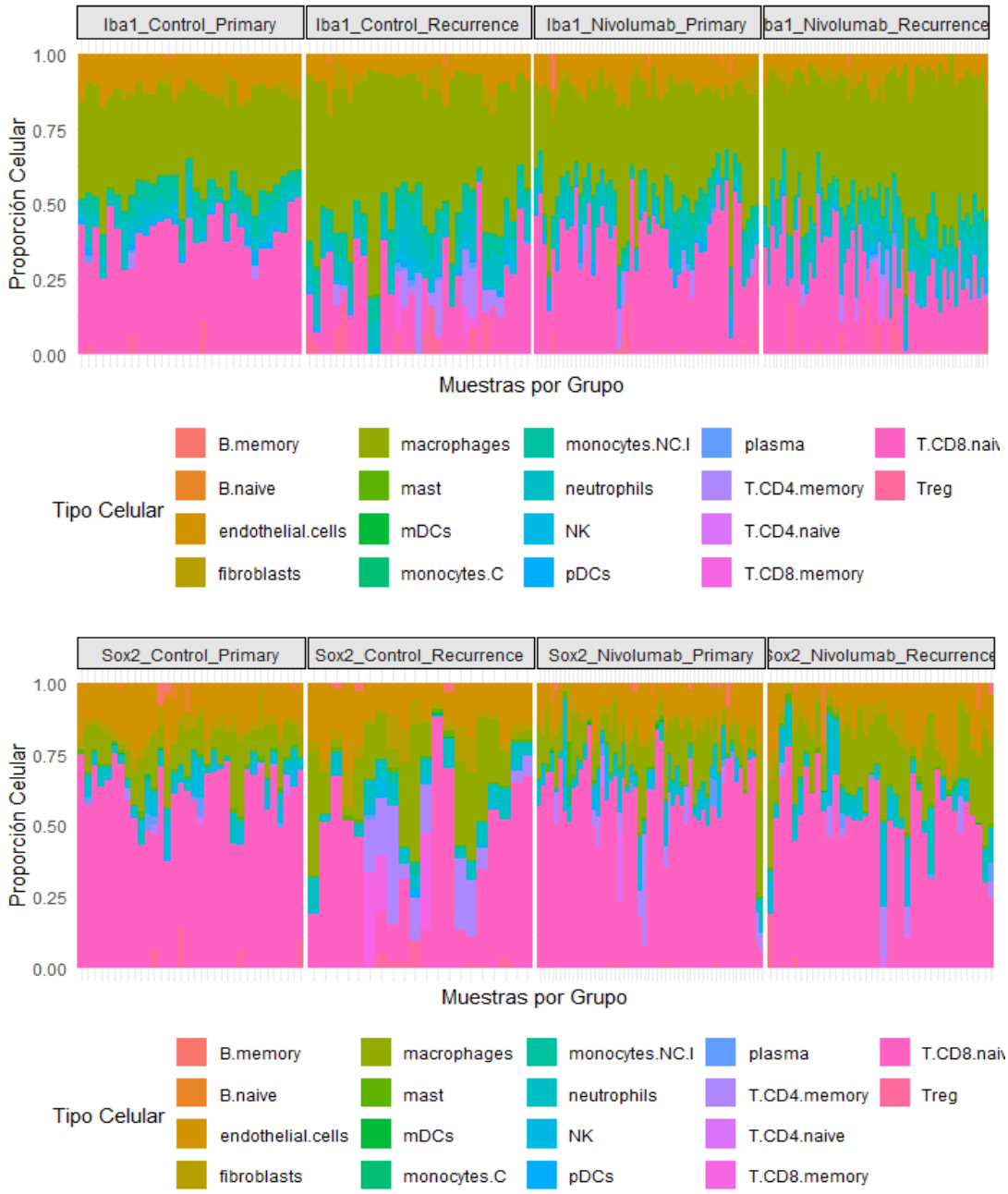
Se realiza la ejecución de este análisis mediante SpatialDecon, el paquete de R cuyo algoritmo matemático permite estimar las proporciones celulares gestionando simultáneamente el ruido de fondo y que contiene la función `spatialdecon()`. El ruido de fondo es la señal medida por el equipo que no proviene de la expresión real de los genes, sino de artefactos técnicos.

Para este análisis se utiliza la matriz llamada *Safe_TME* como base de referencia, que está diseñada para estimar células inmunes y del estroma en el microambiente tumoral. Las matrices son diccionarios que relacionan el tipo y la cantidad de cada gen que están expresados en cada linaje celular. La matriz *Safe_TME* ha sido creada específicamente para evitar genes que se expresan comúnmente en células cancerosas, de modo que la estimación de los tipos celulares no se vea sesgada.

Mediante un modelo de regresión lineal, el algoritmo utiliza esta referencia para desglosar la señal total de cada AOI, calculando qué proporción de cada tipo celular explica de forma óptima el perfil de expresión observado en la muestra.

Los gráficos obtenidos en la Figura 18, se dividen en paneles según los 8 niveles posibles de la variable grupo. Cada barra vertical corresponde a una AOI y muestra la proporción de cada tipo celular mediante una escala de colores.

Figura 18. *Gráficos de Deconvolución Celular.*



Elaboración propia. Arriba para la zona celular Iba1 en los cuatro posibles grupos y abajo para la zona celular Sox2 con la división de cuatro grupos.

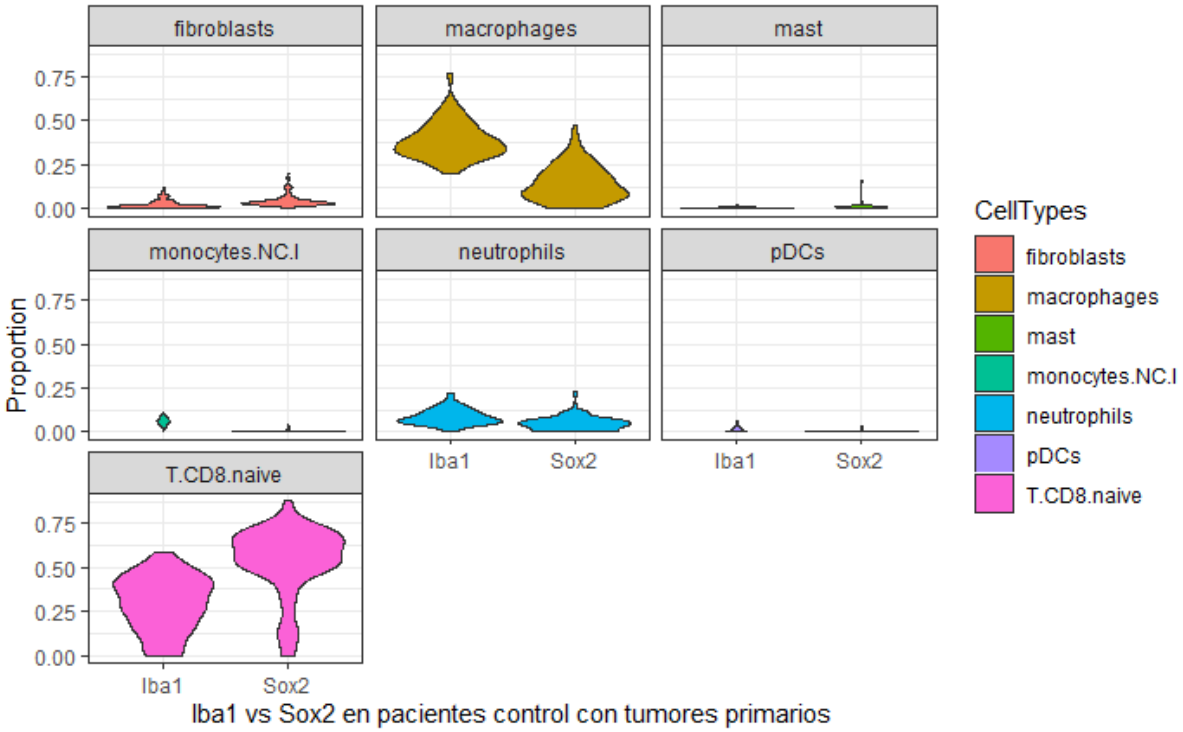
Se observa una distinción clara de las proporciones celulares entre las dos zonas estudiadas. En las AOIs de Iba1 se encuentran mayoritariamente proporciones de macrófagos, confirmando así la identidad mieloide de estos segmentos. Mientras que en las AOIs de Sox2 hay una proporción muy alta de células T (CD4+ y CD8+), este hallazgo representa que las células cancerígenas han sido infiltradas por los linfocitos. a pesar de eso, no se observan

diferencias significativas entre pacientes tratados y o tratados, ni en la deconvolución ni como se ha mostrado anteriormente en el análisis diferencial, lo que se podría clasificar como una presunta infiltración ineficaz.

4.5.6.4. Análisis de proporción diferencial

El análisis de proporción diferencial tiene como objetivo identificar estadísticamente cambios en la abundancia relativa de tipos celulares entre diferentes condiciones experimentales, tal y como se recoge en la Figura 19. Para ello, se emplea la librería speckle, que implementa la función `propeller()`. El fundamento matemático de este test reside en la transformación de las proporciones mediante la función arcoseno, la cual estabiliza la varianza y permite que el test sea más robusto. Finalmente, se aplica el contraste entre las poblaciones celulares para identificar aquellas poblaciones celulares que varían significativamente (FDR < 0.05) entre ambos nichos. En este apartado se analiza la distinción entre los nichos Iba1 y Sox2 en pacientes Control con tumores primarios y se realiza la visualización de las proporciones mediante *violin plots*

Figura 19. *Violin plots del análisis de proporciones celulares.*



Elaboración propia.

Tras el análisis, destacan dos hallazgos principales con significación estadística:

- Macrófagos: Presentan un enriquecimiento marcado en las zonas Iba1, lo que confirma la eficacia del marcador para aislar nichos mieloides.
- Células T CD8+: Se encuentran en una proporción significativamente mayor en las zonas Sox2, corroborando una vez más la infiltración de linfocitos citotóxicos dentro del compartimento tumoral.

Por el contrario, el resto de los linajes celulares (como fibroblastos o células endoteliales) mantienen proporciones similares en ambas regiones.

5. DISCUSIÓN

En la implementación del *pipeline*, se ha ejecutado un análisis completo de un *dataset* disponible en el repositorio Zenodo. Se ha realizado la descarga, la visualización y contextualización de los datos, el preprocesamiento que incluye el control de calidad, el filtrado, la reducción de la dimensionalidad mediante métodos no supervisados, la normalización y la eliminación del efecto de lote, y por último los análisis posteriores, como el análisis de expresión diferencial, el análisis de enriquecimiento, la deconvolución celular y el análisis de proporción diferencial.

Este *pipeline* está basado en el *pipeline* de StandR (45), exceptuando la fase inicial de carga y exploración de los datos, para lo que se ha optado por la viñeta del *pipeline* GeoMxTools (43). Este cambio se ha realizado para que el *pipeline* se pueda aplicar de forma directa a los datos provenientes de GeoMx™ DSP, sin tener que ajustar el tipo de formato de la matriz de conteos y de los metadatos. Otra diferencia con el *pipeline* StandR establecido, es que se ha realizado la eliminación del efecto de lote, tanto a los datos crudos como normalizados, con el fin de observar diferencias y comprender por qué en el *pipeline* StandR no se realizan los análisis posteriores con los datos normalizados. Por último, se ha creado una variable "grupo", que es la unión de tres variables: Tipo_Segmento, Tipo_Tratamiento, Tipo_Tumor, para construir el modelo y la matriz de contrastes adecuados para los análisis de expresión diferencial y así acotar los objetivos del experimento.

Se han elegido los métodos de StandR frente a GeoMxTools porque ofrecen análisis más rigurosos. El análisis de calidad está enfocado en AOI y en gen, y tiene en cuenta el tamaño de librería, el número de células por AOI y que la mayoría de los genes estén representados en la mayoría de AOIs. El método de normalización es más robusto: normaliza mediante *Trimmed Mean of M-values* (TMM), que funciona incluso con datos complejos donde los genes estén muy diferencialmente expresados. Otro factor que ha influido a la hora de elegir StandR es que cuenta con métodos de eliminación del efecto de lote, lo que permite eliminar la variación técnica y que la variación resultante, observada en los análisis posteriores, sea de carácter biológico. Para el análisis de expresión diferencial se ha utilizado el *pipeline* de limma-voom, dado que StandR no contiene funciones propias para este tipo de análisis. Es un punto

favorable que un *pipeline* se pueda mezclar con otros paquetes de Bioconductor para generar análisis. De esta forma, cada usuario puede personalizar sus metodologías entrelazando este *pipeline* con otros con los que esté familiarizado. Por esta razón, StandR ofrece un abanico de opciones mucho más abierto que GeoMxTools en cuanto a los distintos análisis posteriores.

El análisis del *dataset* permite comprobar que GeoMx™ DSP es una técnica muy rigurosa para estudiar la expresión de los genes en células tumorales y su microambiente. El uso de GeoMx™ DSP combinado con un riguroso análisis bioinformático permite diseccionar la realidad de estas poblaciones celulares que coexisten en estrecha proximidad. Este enfoque hace posible perfilar sus funciones específicas, capturando con éxito la compleja interacción entre la progresión del linaje tumoral (Sox2+) y la respuesta inmunitaria del paciente (Iba1+).

Las principales ventajas del *pipeline* definido a lo largo del trabajo, es que se puede aplicar directamente a los datos con el formato que salen del equipo GeoMx™ DSP. Cuenta además con las metodologías más robustas de la actualidad, ya que la revisión y la creación de StandR es del 2024. Es interoperable con otros *pipelines* por el tipo de objeto con el que se trabaja. Es reproducible y flexible, ya que se ha implementado con un *dataset* externo de diferente tamaño al de la demostración del *pipeline* (45).

Entre las limitaciones de este trabajo está el no haber llegado a realizar un proyecto en *workflowr*, una herramienta pensada para desarrollar código, resultados y documentos de una forma organizada para ser transparentes, reproducibles y compartibles. Otra limitación es que el *pipeline* se ha centrado en el análisis de datos GeoMx™ DSP y, actualmente, no es extrapolable a otras tecnologías. Especialmente las fases de lectura de datos, control de calidad y filtrado son altamente dependientes de la tecnología utilizada. Por tanto, en el futuro, se podrían desarrollar nuevos scripts de lectura de datos y control de calidad, adecuados para otras tecnologías, así como adaptar el resto de las etapas para que sean ejecutables a partir de cualquiera de las anteriores vías de entrada al *pipeline*.

La elaboración de este trabajo aporta un pequeño crecimiento a la comunidad científica, ya que se han enumerado y evaluado las técnicas más novedosas con las que se puede realizar el análisis de datos de GeoMx™ DSP. Se remarcan las limitaciones encontradas en GeoMx™ DSP (28) y como se establece StandR para combatirlas (29). Se reproduce el *pipeline*, con

algunos pequeños cambios, en datos de GeoMx™ DSP, para comprobar su escalabilidad y reproducibilidad. Una vez desarrollado e implementado, se acaba definiendo como una forma propia de implementar StandR para el análisis de datos de GeoMx™ DSP. Se cuelgan todos los documentos con los que se ha trabajado durante la elaboración de esta tesis en GitHub (52), para colaborar con la comunidad bioinformática y dar acceso abierto a este trabajo.

De cara al futuro, se planea continuar el desarrollo del *pipeline* incorporando nuevas funcionalidades, adaptaciones a otras tecnologías y mejoras en la automatización y reproducibilidad (*workflowr*), con el objetivo de convertirlo en una herramienta robusta y versátil para la comunidad investigadora.

6. CONCLUSIONES

Las conclusiones de este trabajo son:

- Se ha realizado una revisión de las técnicas y metodologías de la transcriptómica espacial y de la técnica de GeoMx™ DSP, lo que ha permitido comprender el estado del arte de la tecnología.
- Se ha analizado tanto la instrumentación como los métodos de análisis de datos disponibles, identificando las ventajas y limitaciones de cada enfoque.

Esta fase ha sido fundamental para contextualizar el desarrollo del *pipeline*, permitiendo seleccionar las metodologías más actuales y robustas. Se ha alcanzado un conocimiento sólido sobre las técnicas existentes y sobre cómo estas pueden aplicarse al análisis de datos de GeoMx™ DSP.

- Se ha diseñado y codificado un *pipeline* en R capaz de ejecutar de manera sistemática las etapas necesarias para analizar datos de GeoMx™ DSP, incluyendo preprocesamiento: control de calidad, filtrado, normalización, eliminación del efecto de lote, y análisis posteriores: como expresión diferencial, análisis de enriquecimiento, deconvolución celular y análisis de proporción diferencial.

La elección de utilizar StandR como base, combinada con elementos de GeoMxTools para la carga inicial de datos, ha permitido que el *pipeline* sea aplicable directamente a datos crudos del equipo GeoMx™ DSP, sin necesidad de modificaciones complejas. Con ello, se cumple el desarrollo del *pipeline*, garantizando que sea robusto, reproducible y flexible para distintos conjuntos de datos.

- Se ha implementado el *pipeline* en un *dataset* de ejemplo y se ha demostrado que es operativo y escalable, permitiendo procesar datos de manera completa desde la carga hasta los análisis avanzados. El *pipeline* mantiene un alto grado de reproducibilidad y permite combinarse con otros paquetes de Bioconductor, lo que facilita su personalización según las necesidades de cada usuario.
- Se ha validado la metodología reproducible y flexible, pero queda un margen para mejorar la automatización y la integración completa en un entorno *workflowr*.

REFERENCIAS BIBLIOGRÁFICAS

1. Babu M, Snyder M. Multi-Omics Profiling for Health. Mol Cell Proteomics [Internet]. junio de 2023 [citado 2 de febrero de 2026];22(6):100561. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/S1535947623000713

2. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet [Internet]. enero de 2009 [citado 2 de febrero de 2026];10(1):57-63. Disponible en: https://www.nature.com/articles/nrg2484

3. Transcriptome: Connecting the Genome to Gene Function | Learn Science at Scitable [Internet]. [citado 2 de febrero de 2026]. Disponible en: https://www.nature.com/scitable/topicpage/transcriptome-connecting-the-genome-to-gene-function-605/

4. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress update – from bulk to single-cell expression data. Nucleic Acids Res [Internet]. 8 de enero de 2019 [citado 2 de febrero de 2026];47(D1):D711-5. Disponible en: https://doi.org/10.1093/nar/gky964

5. Method of the Year 2013. Nat Methods [Internet]. enero de 2014 [citado 2 de febrero de 2026];11(1):1-1. Disponible en: https://www.nature.com/articles/nmeth.2801

6. Method of the Year 2019: Single-cell multimodal omics. Nat Methods [Internet]. enero de 2020 [citado 2 de febrero de 2026];17(1):1-1. Disponible en: https://www.nature.com/articles/s41592-019-0703-5

7. Moses L, Pachter L. Museum of spatial transcriptomics. Nat Methods [Internet]. mayo de 2022 [citado 2 de febrero de 2026];19(5):534-46. Disponible en: https://www.nature.com/articles/s41592-022-01409-2

8. Dong Z, Chen Y. Transcriptomics: Advances and approaches. Sci China Life Sci [Internet]. octubre de 2013 [citado 2 de febrero de 2026];56(10):960-7. Disponible en: http://link.springer.com/10.1007/s11427-013-4557-2

9. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLOS Comput Biol [Internet]. 18 de mayo de 2017 [citado 2 de febrero de 2026];13(5):e1005457. Disponible en: https://dx.plos.org/10.1371/journal.pcbi.1005457

10. Moreno V, Solé X. Uso de chips de ADN (microarrays) en medicina: fundamentos técnicos y procedimientos básicos para el análisis estadístico de resultados. Med Clínica [Internet]. febrero de 2004 [citado 2 de febrero de 2026];122(Supl.1):73-9. Disponible en: http://db.doyma.es/cgi-bin/wdbcgi.exe/doyma/mrevista.fulltext?pident=13057538

11. Raghavachari N, Garcia-Reyero N, editores. Gene Expression Analysis: Methods and Protocols [Internet]. New York, NY: Springer US; 2025 [citado 2 de febrero de 2026]. (Methods in Molecular Biology; vol. 2880). Disponible en: https://link.springer.com/10.1007/978-1-0716-4276-4

12. Rajawat J. Transcriptomics. En: Arivaradarajan P, Misra G, editores. Omics Approaches, Technologies And Applications [Internet]. Singapore: Springer Singapore; 2018 [citado 2 de febrero de 2026]. p. 39-56. Disponible en: http://link.springer.com/10.1007/978-981-13-2925-8_3

13. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. Nat Rev Genet [Internet]. agosto de 2023 [citado 2 de febrero de 2026];24(8):550-72. Disponible en: https://www.nature.com/articles/s41576-023-00586-w

14. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to spatial transcriptomics for biomedical research. Genome Med [Internet]. 27 de junio de 2022 [citado 2 de febrero de 2026];14(1):68. Disponible en: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-022-01075-1

15. Ben Raphael | Models and Methods for Spatial Transcriptomics | CGSI 2023 [Internet]. 2025 [citado 2 de febrero de 2026]. Disponible en: https://www.youtube.com/watch?v=CRuSrd8JWI0

16. Wang Y, Liu B, Zhao G, Lee Y, Buzdin A, Mu X, et al. Spatial transcriptomics: Technologies, applications and experimental considerations. Genomics [Internet]. septiembre de 2023 [citado 2 de febrero de 2026];115(5):110671. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/S0888754323001155

17. Yue L, Liu F, Hu J, Yang P, Wang Y, Dong J, et al. A guidebook of spatial transcriptomic technologies, data resources and analysis approaches. Comput Struct Biotechnol J [Internet]. 1 de enero de 2023 [citado 2 de febrero de 2026];21:940-55. Disponible en: https://www.sciencedirect.com/science/article/pii/S2001037023000156

18. Piñeiro AJ, Houser AE, Ji AL. Research Techniques Made Simple: Spatial Transcriptomics. J Invest Dermatol [Internet]. 1 de abril de 2022 [citado 2 de febrero de 2026];142(4):993-1001.e1. Disponible en: https://www.sciencedirect.com/science/article/pii/S0022202X21026312

19. Mao G, Yang Y, Luo Z, Lin C, Xie P. SpatialQC: automated quality control for spatial transcriptome data. Bioinformatics [Internet]. 25 de julio de 2024 [citado 2 de febrero de 2026];40(8):btae458. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC11333854/

20. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. mayo de 2015;33(5):495-502.

21. Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. Nat Methods [Internet]. mayo de 2018 [citado 2 de febrero de 2026];15(5):343-6. Disponible en: https://www.nature.com/articles/nmeth.4636

22. Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis *pipelines* with bioinformatics workflow managers. Nat Methods [Internet]. octubre de 2021 [citado

2 de febrero de 2026];18(10):1161-8. Disponible en: https://www.nature.com/articles/s41592-021-01254-9

23. Krull D, Haynes P, Kesarwani A, Tessier J, Chen BJ, Hunter K, et al. A best practices framework for spatial biology studies in drug discovery and development: enabling successful cohort studies using digital spatial profiling. J Histotechnol [Internet]. 2 de enero de 2025 [citado 2 de febrero de 2026];48(1):7-26. Disponible en: https://www.tandfonline.com/doi/full/10.1080/01478885.2024.2391683

24. Spatial Biology vs Bulk Analysis: An Overview of the GeoMx Digital Spatial Profiler [Internet]. 2019 [citado 2 de febrero de 2026]. Disponible en: https://www.youtube.com/watch?v=B7wztG19dkU

25. NanoString University [Internet]. [citado 2 de febrero de 2026]. GeoMx® DSP 101: Introduction to the Technology. Disponible en: https://university.nanostring.com/geomx-dsp-101-introduction-to-the-technology

26. NanoString University [Internet]. [citado 2 de febrero de 2026]. GeoMx® DSP: Data Analysis for RNA NGS Readout. Disponible en: https://university.nanostring.com/geomx-data-analysis-for-ngs-readout

27. NanoString University [Internet]. [citado 2 de febrero de 2026]. How to Run GeoMx® DSP Data Analysis Workflows. Disponible en: https://university.nanostring.com/how-to-run-geomx-dsp-data-analysis-workflows

28. van Hijfte L, Geurts M, Vallentgoed WR, Eilers PHC, Sillevis Smitt PAE, Debets R, et al. Alternative normalization and analysis *pipeline* to address systematic bias in NanoString GeoMx Digital Spatial Profiling data. iScience. 20 de enero de 2023;26(1):105760.

29. Liu N, Bhuva DD, Mohamed A, Bokelund M, Kulasinghe A, Tan CW, et al. standR: spatial transcriptomic analysis for GeoMx™ DSP data. Nucleic Acids Res. 11 de enero de 2024;52(1):e2.

30. R: The R Project for Statistical Computing [Internet]. [citado 2 de febrero de 2026]. Disponible en: https://www.r-project.org/

31. Posit [Internet]. [citado 2 de febrero de 2026]. Posit | The Open-Source Data Science Company. Disponible en: https://posit.co/

32. Blischak JD, Carbonetto P, Stephens M. Creating and sharing reproducible research code the workflowr way [Internet]. F1000Research; 2019 [citado 2 de febrero de 2026]. Disponible en: https://f1000research.com/articles/8-1749

33. Spatial Organ Atlas [Internet]. NanoString. 2026 [citado 2 de febrero de 2026]. Disponible en: https://nanostring.com/products/geomx-digital-spatial-profiler/spatial-organ-atlas/

34. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 1 de enero de 2002;30(1):207-10.

35. The European Genome-phenome Archive of human data consented for biomedical research | Nature Genetics [Internet]. [citado 2 de febrero de 2026]. Disponible en: https://www.nature.com/articles/ng.3312

36. Zenodo [Internet]. [citado 2 de febrero de 2026]. Disponible en: https://zenodo.org/

37. Bergholtz H, Carter JM, Cesano A, Cheang MCU, Church SE, Divakar P, et al. Best Practices for Spatial Profiling for Breast Cancer Research with the GeoMx® Digital Spatial Profiler. Cancers [Internet]. enero de 2021 [citado 2 de febrero de 2026];13(17):4456. Disponible en: https://www.mdpi.com/2072-6694/13/17/4456

38. Jackson KC, Pachter L. A standard for sharing spatial transcriptomics data. Cell Genomics [Internet]. 9 de agosto de 2023 [citado 2 de febrero de 2026];3(8). Disponible en: https://www.cell.com/cell-genomics/abstract/S2666-979X(23)00171-4

39. Artzi SB. Spatial transcriptomic analysis reveals lack of response to PD-1 blockade in recurrent glioblastoma [Internet]. Zenodo; 2025 [citado 2 de febrero de 2026]. Disponible en: https://zenodo.org/records/16839828

40. Artzi SB, Klausen MN, Harwood DSL, Michaelsen SR, Maarup SB, Locallo A, et al. Spatial transcriptomic analysis reveals lack of response to PD-1 blockade in recurrent glioblastoma. Acta Neuropathol (Berl) [Internet]. 17 de septiembre de 2025 [citado 2 de febrero de 2026];150(1):29. Disponible en: https://doi.org/10.1007/s00401-025-02937-9

41. NanoString University [Internet]. [citado 2 de febrero de 2026]. GeoMx® DSP Data Analysis Workflow Session for RNA NGS Readout. Disponible en: https://university.nanostring.com/geomx-academy-ngs-data-analysis-for-rna

42. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods [Internet]. febrero de 2015 [citado 2 de febrero de 2026];12(2):115-21. Disponible en: https://www.nature.com/articles/nmeth.3252

43. Analyzing GeoMx-NGS RNA Expression Data with GeomxTools [Internet]. [citado 2 de febrero de 2026]. Disponible en: https://bioconductor.org/packages/devel/workflows/vignettes/GeoMxWorkflows/inst/doc/GeomxTools_RNA-NGS_Analysis.html#8_Visualizing_DE_Genes

44. Human Kidney [Internet]. NanoString. 2026 [citado 2 de febrero de 2026]. Disponible en: https://nanostring.com/products/geomx-digital-spatial-profiler/spatial-organ-atlas/human-kidney/

45. Analysing Nanostring's GeoMx transcriptomics data using standR, limma and vissE [Internet]. [citado 2 de febrero de 2026]. Disponible en:

https://davislaboratory.github.io/GeoMXAnalysisWorkflow/articles/GeoMXAnalysisWorkflow.html

46. Human Lymph Node [Internet]. NanoString. 2026 [citado 3 de febrero de 2026]. Disponible en: https://nanostring.com/products/geomx-digital-spatial-profiler/spatial-organ-atlas/human-lymph-node/

47. limma powers differential expression analyses for RNA-sequencing and microarray studies | Nucleic Acids Research | Oxford Academic [Internet]. [citado 3 de febrero de 2026]. Disponible en: https://academic.oup.com/nar/article/43/7/e47/2414268

48. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics [Internet]. 1 de enero de 2010 [citado 3 de febrero de 2026];26(1):139-40. Disponible en: https://doi.org/10.1093/bioinformatics/btp616

49. Lucidchart [Internet]. [citado 2 de febrero de 2026]. Lucidchart | Diagramming Powered By Intelligence. Disponible en: https://www.lucidchart.com/pages

50. Delicado Mercader C. GeomxTools_Demo [Internet]. [citado 2 de febrero de 2026]. Disponible en: https://vhir-my.sharepoint.com/my?id=%2Fpersonal%2Fcarme%5Fdelicado%5Fvhir%5Forg%2FDocuments%2FFitxers%20dels%20xats%20de%20Microsoft%20Teams%2FGeoMxTools%5FDemo%2Ehtml&parent=%2Fpersonal%2Fcarme%5Fdelicado%5Fvhir%5Forg%2FDocuments%2FFitxers%20dels%20xats%20de%20Microsoft%20Teams&ga=1

51. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol [Internet]. 3 de febrero de 2014 [citado 3 de febrero de 2026];15(2):R29. Disponible en: https://doi.org/10.1186/gb-2014-15-2-r29 Snyder, Michael. Multi-Omics Profiling for Health. Mol Cell Proteomics [internet]. 2023 Jun [consultado 25 feb 2025];22(6):100561. Disponible en: https://doi.org/10.1016/j.mcpro.2023.100561

52. Delicado-Mercader, C. (2026). Desarrollo de un pipeline bioinformático para el análisis de datos GeoMx™ Digital Spatial Profiler. Disponible en GitHub: https://github.com/carmedelicado/GeoMx_Pipeline_CDM

ANEXO

1. Aplicaciones más concretas de la transcriptómica espacial.

Los análisis espaciotemporales de desarrollo del tejido, como sería la comprensión del desarrollo del feto humano a partir de las células embriogénicas; la generación de atlas con la composición espacial de los linajes celulares en distintos tejidos o regiones como el celebro, el corazón, los pulmones o el riñón; los mecanismos moleculares de la desregulación genética o celular en enfermedades, ya que se supera la incapacidad de examinar directamente las células causantes de la enfermedad y su microambiente; el estudio de la heterogeneidad celular y los microambientes en los tumores, para entender el desarrollo de la metástasis y poder buscar nuevas terapias más focalizadas; el análisis de las poblaciones celulares, para tener una mirada más profunda sobre cómo se agrupan los tipos de células, como interaccionan con sus células vecinas y cuál es su organización celular; el desarrollo de biomarcadores moleculares, celulares y microestructurales más precisos y sensibles espacialmente, lo que permitirá mejorar el diagnóstico, el pronóstico y el tratamiento de enfermedades (16).

2. Diferencias entre el preprocesamiento técnico según el método de extracción de ARN.

El preprocesamiento técnico de imagen suele implicar el registro de imágenes, la identificación de puntos (*spot*) de transcritos y la segmentación celular. Mediante diversos paquetes de software y herramientas de programación se identifica la ubicación, el tipo y la cantidad del transcrito en la imagen, y se genera la matriz de índice de ubicación y la matriz de expresión génica relacionadas (17).

El preprocesamiento técnico de secuenciación se define por la recogida de secuencias sin procesar de ARNm capturados y las imágenes microscópicas del tejido perfilado. Entonces, programas bioinformáticos realizan la alineación de lecturas con el genoma de referencia, emparejan los códigos de barras de lectura con las ubicaciones espaciales de la matriz y cuentan el número de transcripciones génicas en cada ubicación para producir una matriz de conteo génica en cada *spot* (14).

La segmentación es un posible paso en el preprocesamiento técnico y se basa en convertir la matriz de puntos génicos en una matriz de célula-gen. Esta, solo puede ser producida en las técnicas donde la resolución sea a nivel de célula única (17).

3. Descripción detallada del Control de Calidad, la Reducción de la Dimensionalidad, la Normalización y la Corrección del Efecto de Lote

El control de calidad *"Quality Control (QC)"* tiene como objetivo evaluar la fiabilidad y eliminar los elementos que puedan introducir ruido técnico o sesgar la interpretación biológica. Este proceso se aplica a distintos niveles: a nivel de *spot*, donde se examinan métricas como el número total de genes detectados, la cantidad de transcritos (UMIs) o el porcentaje de genes mitocondriales, con el fin de descartar regiones con baja calidad o sin tejido; a nivel de gen, donde se filtran aquellos con baja frecuencia de detección o escasa variabilidad; y a nivel de muestra, donde se inspecciona la integridad del tejido y la correcta alineación con la imagen histológica (14,19).

La reducción de dimensionalidad se utiliza para reducir los datos de expresión génica a dimensiones más pequeñas, para obtener análisis más informativos de cómo se distribuyen los datos. Se pueden emplear con el objetivo de sintetizar o de visualizar los datos. Para sintetizar se hace uso del Análisis del Componentes Principales (PCA) y para visualizar UMAP o t-SNE (18).

La normalización de los datos es esencial para tener en cuenta las diferencias en la tasa de captura de ARNm en todo el tejido. Es una forma de evitar interpretaciones erróneas producidas por las limitaciones de la técnica. Si normalizamos los datos, hacemos que los *spots* se puedan comparar entre ellos y que las diferencias que se encuentren sean debidas a patrones biológicos (14).

La corrección del efecto de lote, *Batch-effect*, se define como variaciones técnicas provenientes de diferentes lotes experimentales, así como, la variabilidad aportada por las distintas muestras y secciones de tejido, los distintos días en que se realizan los experimentos, los kits y reactivos de laboratorio y los operarios o equipos utilizados. Necesitamos utilizar herramientas de integración especializadas para corregir el efecto de lote y mejorar la fusión de datos, de modo que los efectos experimentales se puedan mostrar con mayor claridad (17).

4. Conversiones de funciones de tipo `NanostringGeoMxSet a SpatialExperiment:`

- Conteos: `NanostringGeoMxSet->` `exprs()`, `SpatialExperiment->` `assays()`.
- Anotaciones AOI: `NanostringGeoMxSet->` `pData()`, `SpatialExperiment->` `colData()`.
- Anotaciones genes: `NanostringGeoMxSet->` `fData()`, `SpatialExperiment->` `rowData()`.

5. Repositorio de datos donde descargarse *dataset*s, número de *Dataset*s de GeoMx™ DSP que contienen y tipo de tejido.

El mayor repositorio accesible para descargar datos es el *Spatial Organ Atlas de Nanostring,* donde se encuentran seis bases de datos de tejidos como el cerebro, los nódulos linfáticos, el páncreas, el hígado, el colon o los riñones en humano y otras dos que provienen del cerebro y el embrión en ratón. Es por excelencia el repositorio oficial con muestras accesibles para familiarizarse y empezar a analizar datos de *GeoMx™ DSP*.

GEO es un repositorio público de datos de expresión génica de alto rendimiento y de *arrays* de hibridación, chips y *microarrays*. Aparecen aproximadamente 200 series como resultado donde se pueden fácilmente descargar las matrices de conteos y la documentación adicional. La variedad de tejidos que compone este repositorio es muy alta.

BioStudies de *ArrayExpress* es una base de datos abierta de EMBL-EBI diseñada para agrupar, describir y alojar toda la información asociada a un estudio biológico. En esta herramienta se encuentran cinco proyectos que trabajan con el tejido pleural, el intestino, el colon, los gánglios limfáticos y el cáncer de mama.

EGA es un servicio para archivar y compartir datos de investigación biomédica, se encuentran ocho proyectos. Los estudios se centran en el cáncer gástrico, el cáncer de mama metastásico, en el cartílago de la rodilla y en el endotelio de las arterias intracraneales.

Zenodo es un repositorio abierto para subir y publicar *datasets*, código, figuras, artículos, software y cualquier material científico, sin coste y con un identificador único. Se encuentran aproximadamente 80 *datasets* de todo tipo de tejido.