**RESOLUCIÓN PARA LA MEJORA Y APLICACIONES DE LA INTERFAZ OMICS INTEGRATOR**

TESIS DE MAESTRÍA

PRESENTADO POR:

**ESTEBAN VARGAS PARRA**

DIRECTOR:

**ALEX SANCHEZ PLA**

**ÍNDICE GENERAL**

[INTRODUCCIÓN 5](#_Toc100520573)

[Contexto y justificación del trabajo 5](#_Toc100520574)

[Descripción general 5](#_Toc100520575)

[Justificación del Trabajo Final de Máster 5](#_Toc100520576)

[Objetivos 6](#_Toc100520577)

[Objetivos generales 6](#_Toc100520578)

[Objetivos específicos 6](#_Toc100520579)

[Enfoque y método a seguir 6](#_Toc100520580)

[Planificación con hitos y temporización 7](#_Toc100520581)

[Tareas 7](#_Toc100520582)

[Calendario 7](#_Toc100520583)

[Hitos 11](#_Toc100520584)

[Análisis de riesgo 11](#_Toc100520585)

[Resultados esperados 11](#_Toc100520586)

[Plan de trabajo 11](#_Toc100520587)

[Memoria 12](#_Toc100520588)

[Producto 12](#_Toc100520589)

[Presentación virtual 12](#_Toc100520590)

[Autoevaluación del proyecto 12](#_Toc100520591)

[ESTADO DEL ARTE 13](#_Toc100520592)

[1. Ómicas 13](#_Toc100520593)

[1.1. Transcriptómica 13](#_Toc100520594)

[1.1.1. EST 14](#_Toc100520595)

[1.1.2. Análisis Serial y Cap de expresión génica (SAGE/CAGE) 14](#_Toc100520596)

[1.1.3. Microarrays 14](#_Toc100520597)

[1.1.4. RNA-Seq 14](#_Toc100520598)

[1.2. Análisis de datos 15](#_Toc100520599)

[1.2.1. Análisis de datos de ARN-Seq 15](#_Toc100520600)

[1.2.1.1. Control de calidad 15](#_Toc100520601)

[1.2.1.2. Alineación 16](#_Toc100520602)

[1.2.1.3. Cuantificación 16](#_Toc100520603)

[1.2.1.4. Expresión diferencial 16](#_Toc100520604)

[2. Revisión Bibliográfica 16](#_Toc100520605)

[2.1. Data Analysis Tool Extension (DAnTE) 16](#_Toc100520606)

[2.2. DanteR 17](#_Toc100520607)

[2.3. R-ODAF 17](#_Toc100520608)

[2.4. iODA 18](#_Toc100520609)

[2.5. MOVIS 19](#_Toc100520610)

[2.6. ReactomeGSA 19](#_Toc100520611)

[2.7. AutoGGN 19](#_Toc100520612)

[3. Justificación 20](#_Toc100520613)

[3.1. Análisis de enriquecimiento 20](#_Toc100520614)

[METODOLOGÍA 22](#_Toc100520615)

[Bibliografía 23](#_Toc100520616)

**ÍNDICE DE FIGURAS**

[**Figura 1.** Primera fase. 8](#_Toc100520617)

[**Figura 2.** Segunda fase. 8](#_Toc100520618)

[**Figura 3.** Tercera fase. 9](#_Toc100520619)

[**Figura 4.** Cuarta fase. 9](#_Toc100520620)

[**Figura 5.** Fase final. 10](#_Toc100520621)

[**Figura 6.** Imágenes representativas de DAnTE. (A) Cuadrícula de datos y panel de navegación; (B) Diagrama de cajas; (C) Heatmap de correlación de un conjunto de datos; (D) Resumen de péptidos a proteína del método RRollup [30]. 17](#_Toc100520622)

[**Figura 7.** Concepto general de la aplicación de R-ODAF. 18](#_Toc100520623)

[**Figura 8.** Interface principal de iODA. (1) Paneles funcionales. (2) Botones funcionales. 18](#_Toc100520624)

[**Figura 9.** Vista dividida de múltiples ómicas. 19](#_Toc100520625)

[**Figura 10.** Vista dividida de múltiples funciones de ReactomeGSA. 19](#_Toc100520626)

[**Figura 11.** Estructura del modelo y aplicaciones de AutoGGN. 20](#_Toc100520627)

# INTRODUCCIÓN

# Contexto y justificación del trabajo

## Descripción general

El presente trabajo Final de Máster (TFM) se basará en un proyecto aplicativo donde se tomará una interfaz denominada Omics Integrator, la cual fue creada en un trabajo previo, con el fin de procesar datos ómicos; a esta aplicación se le añadirán nuevas herramientas y técnicas que sean suficientemente eficientes, con el fin de manejar este tipo de datos de una manera más óptima en un tiempo razonable.

# Justificación del Trabajo Final de Máster

La genómica, proteómica y metabolómicas se han visto envueltas durante los últimos años, en el área de las ciencias biológicas en el desarrollo de nuevas técnicas, con el fin de estudiar conjuntamente procesos de expresión de los genes y sus productos [1].

Antiguamente se creía que llegar a estudiar un sistema biológico sería algo inalcanzable, sin embargo, con el auge de la bioinformática, la mejora de nuevas técnicas y el avance de la tecnología, se ha logrado interpretar procesos o sistemas enteros pertenecientes a las células, generando cantidades robustas de datos [2].

La aplicación Omics Integrator, es una aplicación que se encarga de integrar diferentes aplicaciones de análisis de datos ómicos, para llegar a resultados como lo son el VolcanoPlot, Análisis de Componentes Principales (PCA), Filtraciones y Mapas de Calor.

A pesar de ser una aplicación con las funcionalidades mencionadas previamente, ésta requiere de algunas mejoras, donde se permita relacionar de manera más sencilla al usuario que la pondrá en práctica con la interfaz, así como se incluyan nuevas funcionalidades con ejemplos y vídeos interactivos con el fin de ser más práctica al momento de emplearse.

Por ende, se empleará el paquete Shiny perteneciente a R, con el objetivo de desarrollar nuevas técnicas de análisis y funciones interactivas en la aplicación Omics Integrator. Con este TFM se buscará dejar un aplicativo que no sólo sea más potente, sino que además tenga la opción de tener mejoras concordes a las futuras nuevas técnicas de análisis vayan surgiendo en esta área de estudio.

# Objetivos

A continuación, serán enunciados los objetivos que se desean alcanzar con el Trabajo Final de Máster.

## Objetivos generales

1. Integrar a la aplicación Omics Integrator, nuevas funciones y mejoras de análisis ómicos.
2. Incluir en la aplicación Omics Integrator, ejemplos y documentación interactiva para el usuario

## Objetivos específicos

1. Funciones y mejoras:
2. Realizar una búsqueda rigurosa de métodos de análisis de datos ómicos
3. Indicar las ventajas y desventajas de los distintos métodos de análisis
4. Seleccionar el o los más completos e incluirlos en el aplicativo.
5. Validar los datos con una plataforma de análisis
6. Ejemplos y documentación:
7. Realizar una documentación en con el instructivo del uso del aplicativo.
8. Incluir vídeos interactivos con ejemplos de cada función del aplicativo.
9. Disponer el software con funciones actualizada a los usuarios de interés.

# Enfoque y método a seguir

Como primer paso para llevar a cabo respectivo proyecto, es buscar las diferentes fuentes bibliográficas investigativas a cerca de análisis ómicos, con el fin de saber y comprender el tratamiento de los datos y la funcionalidad de ello. A pesar de que esta fase no es muy detallada, se realiza la el primer contacto con el tema de interés.

Basados en ello, se van planteando las formas de trabajo:

* Primero se desarrolla la parte teórica del proyecto (búsqueda de bibliografía, antecedentes, comparaciones, etc…) y después de ello se procede a la parte práctica (funciones disponibles en R, paquetes, ejemplos), se hace reconocimiento y análisis del Omics Integrator, se evalúan las ventajas y desventajas para identificar mejoras del aplicativo.
* Empezar con las dos partes del proyecto a la vez, lo práctico y lo teórico.

Se procede a emplear la primera opción, ya que es un tema robusto y se requiere una búsqueda minuciosa por parte del investigador, para definir cuáles son las aplicaciones que el Omics Integrator no posee y además aquellas que están disponibles cómo se podrían mejorar. Además, que, con esto realizado, se puede proceder a crear un paquete en R y poder implementarlo directamente en la aplicación web.

Así mismo, aquellas aplicaciones que se desean incluir o mejorar en el software, al contar con una base teórica definida, permitirá tener en cuenta el tiempo en el cual se procede a desarrollar el código en R ya sea que se cuente con unas bases o que deba empezarse desde cero. Además, una vez se empiece con la creación del código, es importante que el adiestramiento del paquete Shiny para la modificación del aplicativo y la inclusión de nuevas funciones.

# Planificación con hitos y temporización

En esta parte será especificada las tareas que se llevarán a cabo en el Trabajo Fin de Máster, asimismo, se empleará un calendario que detallará las distintas labores y lapsos de entrega. Finalmente, se evaluarán y analizarán los riesgos que puedan alterar el proyecto.

## Tareas

A partir de lo mencionado previamente se permite desglosar en las siguientes actividades:

* Realizar una búsqueda rigurosa de métodos de análisis de datos ómicos (OE 1, con una duración de 30 horas).
* Para cada método de análisis se asignarán los siguientes trabajos:
* Identificar la funcionalidad y uso de los métodos (O.E. 1.I)
* A partir de literatura y trabajos, establecer las ventajas y desventajas de los análisis a emplear (O.E. 1.II)
* Realizar un cuadro comparativo que permita determinar cuáles son los análisis más completos y establecer cuál o cuáles serán incluidos a partir de las ventajas y desventajas (O.E. 1.III)
* Buscar paquetes estadísticos en R que permitan el cálculo de los análisis deseados, si no hay uno programarlo (O.E. 1.III)
* Comparar resultados con plataformas previas, como Galaxy (O.E. 1.IV)
* Adiestramiento en el manejo de Shiny y Git.hub (9 horas O.E. 1 y 2)
* Crear el código en R con el respectivo análisis (60 horas dependiendo de la complejidad O.E. 1)
* Crear una documentación para el uso del aplicativo con todas sus funciones con un paso a paso (10 horas O.E. 2.I)
* Realizar vídeos interactivos con ejemplos para el uso de la aplicación, una opción viable para la creación del vídeo es OBS Studio (20 horas O.E. II)
* Subir los vídeos al código con la ayuda del paquete *vembedr* (2 horas O.E. II)
* Interactuar con el software que quede a disposición del usuario de interés (Tiempo:N.A. O.E. III)

# Calendario

En la siguiente gráfica se presentará el calendario por meses del Trabajo Final de Máster. Para ello se tomó en cuenta la disponibilidad del estudiante la cual está basada principalmente los fines de semana y festivos, mientras que los días entre semana la dedicación será máxima de unas 3 horas por día.



**Figura 1.** Primera fase.

Como se puede ver en el mes de febrero, el día 23 empieza la definición de los puntos tenidos en cuenta para el plan de trabajo. Como primera fase se definieron los objetivos, cómo se llegarán a cumplir esos objetivos y, además, cuánto tiempo tomará la realización de cada tarea.



**Figura 2.** Segunda fase.



**Figura 3.** Tercera fase.

Para los meses de marzo y abril se trabajará en toda la parte robusta del presente proyecto, trabajando directamente sobre cada objetivo y tarea, además durante estos meses se hará el trabajo directo en la búsqueda completa de la bibliografía y el trabajo que llevará más tiempo, que es la creación de los códigos y la validación de los resultados.



**Figura 4.** Cuarta fase.

Para el mes de mayo en sus días finales, se tiene prevista la entrega del documento con finalizado.



**Figura 5.** Fase final.

Finalmente, en el mes de junio se llevará a cabo la realización de la presentación y la defensa pública.

# Hitos

Analizando las actividades planteadas en el proyecto, se enmarcan a continuación los hitos, que son aquellos percances donde si se presenta algún retraso, influyen negativamente en los plazos de las demás tareas.

**Tabla 1.**

|  |  |  |
| --- | --- | --- |
| **Hito** | **PEC** | **Fecha** |
| Plan de trabajo | PEC 1 | 23/02/2022 |
| Desfase de trabajo | PEC 1 | 08/03/2022 |
| Búsqueda de bibliografía | PEC2 | 07/03/2022 |
| Creación del Paquete en R y aplicación Shiny | PEC2 y PEC3 | 11/04/2022 |
| Creación de documento y vídeo | PEC3 | 16/05/2022 |
| Entrega documento | PEC4 | 02/06/2022 |
| Elaboración presentación | PEC5a | 06/06/2022 |
| Defensa pública | PEC5b | 24/06/2022 |

# Análisis de riesgo

En todos los proyectos se presentan externalidades que afectan negativamente el avance del mismo, provocando que la TFM no se lleve a cabo como se esperaba.

Enseguida, se señalan algunos de los factores que logre afectar lo que se planeó con aterioridad:

1. El tiempo en el que se realiza el proyecto es muy corto.
2. El trabajo proyecta varios temas abiertos, por lo cual definir uno específico es riguroso
3. No llevar a cabo el plan de trabajo como se había estructurado
4. Que el plan de trabajo no haya sido el correcto
5. Poco conocimiento en la creación de paquetes en R
6. No poder crear las modificaciones y aplicaciones del sitio web, debido a que el autor no ha tenido contacto con el paquete Shiny.
7. Hubo un desfase de trabajo en la entrega de la PEC 1, lo cual provocó un retraso en el avance del trabajo final como se esperaba.

## Resultados esperados

Finalizado el proyecto se obtendrán las siguientes salidas que se especifican a continuación:

## Plan de trabajo

El del documento presente, en el que se indica y propone el desarrollo del trabajo, detallando como se va a proceder en la elaboración del TMF. Además, se señalan los objetivos, los hitos, el tiempo estimado de cada actividad. Finalmente, se definen los posibles riesgos que conlleva el proyecto.

## Memoria

Como objetivo de este apartado es evidenciar el trabajo que se ha llevado a cabo, resaltando el contexto del porqué se realiza y la causa de interés para el área de las ciencias biológicas. Presentando los objetivos que se esperan obtener, describiendo la metodología empleada y exhibiendo los resultados obtenidos a lo largo del TFM.

## Producto

Se presentarán los resultados mencionados en la memoria, así como se creará una extensión de la aplicación Omics Integrator que permita realizar análisis adicionales más los que están incluidos, esto se presente realizar con el programa R y el paquete de documentos interactivos Shiny.

## Presentación virtual

Se realizará una presentación que sintetizará la estructura y los resultados obtenido en el trabajo final de máster. Además de realizará una exposición visual y oral con la información más relevante del proyecto.

## Autoevaluación del proyecto

En este aparto se estima si se alcanzaron los objetivos propuestos inicialmente. Detallando las maneras en las que se logró cumplir los objetivos o el porqué no se logaron cumplir.

# ESTADO DEL ARTE

Con el avance de las plataformas modernas para mediciones ómicas de alto rendimiento, éstas se han vuelto necesarias, con el fin obtener avances de estudios biológicos, biomédicos, entre otros; buscando adoptar un enfoque que otorgue información sobre los sistemas biológicos. Datos de distintas fuentes ómicas, como la genómica, proteómica y metabolómica, se han integrado para descubrir el complejo funcionamiento de la biología de sistemas, empleando algoritmos predictivos basados en un lenguaje de aprendizaje automático. A partir de estos métodos de aprendizajes estadísticos, se brindan nuevas técnicas para integrar y analizar datos ómicos, permitiendo el hallazgo de biomarcadores, con el potencial de predecir diferenciación de expresión génica [3].

## Ómicas

Esta palabra proveniente del inglés, la cual se refiere al estudio de un conjunto en total [2], en el área de las ciencias biológicas se emplea a delimitar el análisis de los distintos sistemas biológicos que contiene el funcionamiento de la célula [4]. La primera disciplina ómica en aparecer fue la genómica, enfocándose en el estudio de genomas completos, como oposición a lo que se realizaba tradicionalmente en la genética que examinaba variantes de manera individual o genes únicos; estos estudios genómicos otorgaron un cerco útil para mapear y estudiar variantes genéticas específicas que favorecen tanto a enfermedades mendelianas como complejas [5].

Las tecnologías ómicas que aporta la biología de sistemas, son herramientas importantes para análisis integrales, debido a que, aquellas tecnologías tradicionales que se basan desde un enfoque reduccionista, han sido insuficientes para resolver problemas en un sistema biológico. En lugar de esto, la biología de sistemas posee un enfoque holístico, con el fin de comprender mejor todo el proceso tanto a nivel cuantitativo como cualitativo, proporcionando información sobre enfermedades, terapias, toxicidades, entre otros [6].

Antiguamente el estudio en su totalidad de un sistema biológico era una labor improbable, además la bioinformática era una herramienta de acceso limitado, donde los estudios se restringían a un único elemento o interacción específica, como saber si un gen estaba expresándose o no; pero en la actualidad, las técnicas modernas son capaces de examinar procesos o sistemas completos de las células, creando cantidades robustas de datos [7].

### Transcriptómica

Un transcriptoma es el conjunto de transcritos en una célula, y su cantidad, para un momento de desarrollo o condición fisiológica determinada; entender el transcriptoma es fundamental para descifrar los elementos funcionales del genoma y dejar ver los constituyentes moleculares de células, tejidos, así como comprender el desarrollo y dado el caso una enfermedad [8].

Una vez el genoma ha sido secuenciado, el análisis del transcriptoma otorga la capacidad de comprender la expresión del genoma a nivel de transcripción, proporcionando información a cerca de la estructura del gen, la regulación de la expresión del gen, función del producto del gen y la dinámica del genoma [9]. El análisis del transcriptoma indicará la red de regulación de los procesos biológicos y en paralelo brindará orientación en el diagnóstico de enfermedades, terapia clínica, mejora de cultivos, entre otros; por ende, los objetivos principales de la transcriptómica son catalogar todas las especies de transcritos, incluyendo ARNm, ARN no codificantes y ARN pequeños [8].

La generación de datos en los transcritos de ARN se logra por medio de dos principios principales: secuenciación de transcritos individuales (EST o RNA-Seq) o hibridación de transcritos con una matriz ordenada de sonda de nucleótidos (Microarrays).

#### EST

Una EST es una corta secuencia de nucleótidos creada a través de una sola transcripción de ARN, éste se copia como ADNc por medio de una enzima transcriptasa inversa previa a secuenciar el ADNc resultante [10]. El método de secuenciación Sanger era el más común hasta la aparición de métodos con mayor rendimiento como la secuenciación por síntesis [10]. No obstante, en la actualidad se emplean métodos de mayor rendimiento, sin embargo, las bibliotecas EST normalmente otorgan información de secuencia para los primeros diseños de Microarrays [11].

#### Análisis Serial y Cap de expresión génica (SAGE/CAGE)

SAGE fue un desarrollo a partir de la metodología EST con el fin de aumentar el rendimiento de las etiquetas generadas y permitir cierta cuantificación de la abundancia de transcritos [12]. En ADNc se construye a partir del ARN, posteriormente se digiere en fragmentos de etiqueta de 11 pb empleando enzimas de restricción que cortan una secuencia específica de 11 pares de bases a lo largo de esa secuencia. Estas etiquetas de ADNc se conectan de cabeza a cola en largas cadenas (mayores a 500 pb) y se secuencian usando métodos de bajo rendimiento, pero de longitud de lectura larga, semejante a la secuenciación Sanger. Una vez se encuentren en este paso, las secuencias son restauradas en sus etiquetas originales de 11 pb [12]. Si se cuenta con una genoma de referencia, se alinean estas etiquetas y se pueden alinear para identificar un gen; si no se dispone de un genoma de referencia, las etiquetas se usan directamente como marcadores de diagnóstico si se expresan diferencialmente [13].

El método CAP de análisis de expresión génica (CAGE) es un cambio de SAGE donde se secuencia etiquetas desde el extremo 5’ de un transcripción ARNm solamente [14]. Así que, el lugar de inicio de la transcripción de los genes se identifica cuando las etiquetas son alineadas con un genoma de referencia. Identificar sitios de inicio de genes es conveniente para el análisis de promotores y la clonación de ADNc de longitud completa [13].

#### Microarrays

Consiste en oligómeros de nucleótidos cortos, denominados “sondas”, los cuales se instalan en un sustrato sólido [15]. La abundancia de transcritos es determinada a través de la hibridación de transcritos marcados con fluorescencia. La intensidad de fluorescencia en cada ubicación de la sonda sobre la matriz señala la abundancia de transcritos para esa secuencia de la sonda [16]. Los Microarrays necesitan de cierto conocimiento sobre el organismo de interés, por así explicarlo, en forma de una secuencia genómica anotada o en una biblioteca de tecnología ecológicamente racionales que puedan usarse para generar las sondas para el arreglo [16].

#### RNA-Seq

Se refiere a la combinación de una metodología de secuenciación de alto rendimiento con métodos computacionales para tomar y cuantificar las transcripciones presente en un extracto de ARN [17]. Las secuencias de nucleótidos formadas llegan a tener una longitud de 100 pb, sin embargo, pueden encontrarse entre 30 pb y más de 10000 pb, de acuerdo al método de secuenciación utilizado. Esta técnica se aprovecha del muestreo profundo del transcriptoma con muchos fragmentos cortos de un transcriptoma para permitir la reconstrucción computacional del transcrito de ARN original al alinear lecturas con un genoma de referencia entre sí [8]. Como ventaja, las cantidades de ARN de entrada son mucho menores para RNA-Seq (cantidad en nanogramos) respecto a los Microarrays (cantidad en microgramos), permitiendo un examen más detallado de las estructuras celulares [18]. En teoría, no hay un límite superior de cuantificación en RNA-Seq y su señal de fondo es muy baja para lecturas de 100 pb en regiones no repetitivas [17].

### Análisis de datos

Los métodos de transcriptómica necesitan de cálculos significativos para obtener datos significativos en los experimentos de Microarrays y RNA-Seq. Los datos de Microarrays son registrados como imágenes de alta resolución, lo que solicita detección de características y análisis espectral. Los archivos de imagen sin procesar tienen un tamaño considerablemente mayor a los que son procesados; múltiples sondas cortas concuerdan con una única transcripción que puede revelar detalles a cerca de la estructura de intrón-exón, lo que requiere de modelos estadísticos para determinar la autenticidad de la señal resultante [19].

Por otra parte, los estudios de RNA-Seq logran la producción de más de 109 secuencias de ADN cortas, las cuales se alinean con genomas de referencia formados por millones o miles de millones de pares de bases. El ensamble *De novo* de lecturas en un conjunto de datos necesita la construcción de gráficos de secuencia altamente complejos. Las operaciones RNA-Seq son bastante repetitivas y se benefician de la computación en paralelo, pero los algoritmos modernos admiten que el hardware informático de consumo sea suficiente para ensayos transcriptómicos sencillos que no requieren un ensamblaje de lectura *De novo*. Permitiendo capturas con precisión un transcriptoma humano por medio de RNA-Seq con 30 millones de secuencias de 100 pb por muestra [20][21].

#### Análisis de datos de ARN-Seq

Los ensayos de ARN-Seq producen un volumen alto de lecturas de secuencias no procesada, las cuales deben procesarse para obtener información útil; para ello es necesario la combinación de herramientas de software bioinformáticos que cambia según el diseño y los objeticos experimentales. El proceso se compone de cuatros etapas: control de calidad, alineación, cuantificación y expresión diferencial [22]. Los programas más empleados se ejecutan por medio de una interfaz de línea de comandos, como lo es un entorno Unix o dentro del entorno estadístico R/Bioconductor [23].

##### Control de calidad

Las lecturas de secuencias no son perfectas, por lo tanto, se estima la precisión de cada base en la secuencia para los siguientes análisis. Los datos no procesados se inspeccionan en busca de puntajes de alta calidad para cada base, el contenido de Guanina-Citosina debe coincidir con la distribución esperada, la sobrerrepresentación de motivos de secuencias particularmente cortos y una tasa de duplicación de lectura repentinamente alta [20]. Se encuentran múltiples opciones para el análisis de aclidad de la secuencia, incluyendo paquetes de software FastQC y FaQCs [24]. Las rarezas reconocidas pueden ser excluidas recortándose o etiquetándose para un tratamiento especial en procesos posteriores [13].

##### Alineación

Las secuencias de transcripción se alinean con un genoma de referencia, para vincular la abundancia de lectura de secuencias con la expresión de un gen en específico, o se alinean *De novo* entre sí, si no hay una referencia disponible. En avance de los softwares han solucionado estos problemas en gran medida y los aumentos en la longitud de lectura de secuenciación reducen aún más las lecturas de mapeo múltiple [25].

##### Cuantificación

Esta se puede realizar a nivel de gen, exón o transcrito. Como resultados se obtiene una tabla de recuentos para cada función suministrada al software; los recuentos pueden ser calculado por el paquete de software HTSeq [26]. La cuantificación a nivel de transcripción es más compleja y necesita de métodos probabilísticos para estimar la abundancia de isoformas de transcripción partiendo de información de lectura corta [27]. La lecturas que están alineadas con varias ubicaciones deben identificarse y eliminarse, alinearse con una de las ubicaciones posibles o alinearse con la ubicación más probable [13].

##### Expresión diferencial

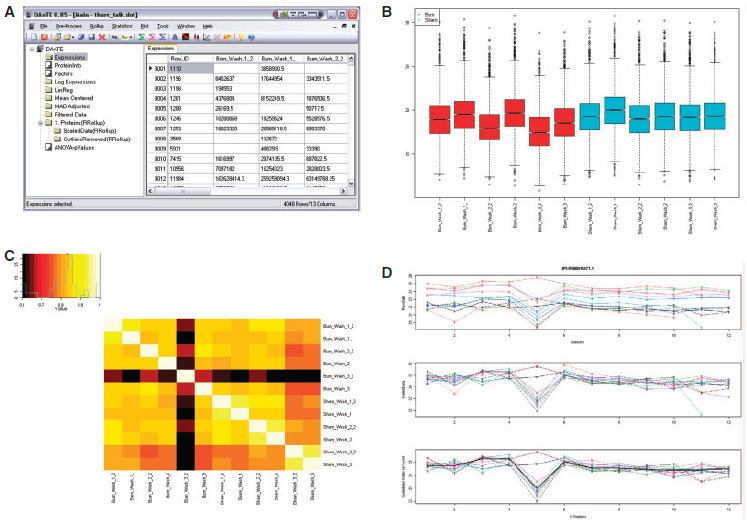
Esta expresión se mide normalizando, modelando y analizando estadísticamente los datos. La mayoría lee una tabla de genes y recuento de entrada, en otros caso, se aceptan alineaciones de lectura en forma de mapa de alineación binaria como entrada [28]. Los resultados finales son listas de genes con pruebas asociada por pares para la expresión diferencial entre tratamientos y las estimaciones de probabilidad de esas diferencias [29].

## Revisión Bibliográfica

En esta sección se mencionarán los proyectos y las principales aportaciones que se han realizado en este campo de análisis de datos ómicos empleando la bioinformática como herramienta para sus cálculos estadísticos para la obtención de datos significativos.

### Data Analysis Tool Extension (DAnTE)

Es una herramienta encargada en abordar los retos relacionados con datos cuantitativos de proteómicas. Esta herramienta se e empleado en el tratamiento para datos de Microarrays y logra ampliar a otros tipos de datos de alto rendimiento. Además, esta herramienta consta de métodos de normalización seleccionados, algorimos de imputación de valores faltantes, métodos de resumen de péptido a proteína, funciones variables de representación gráfica y un esquema integral de pruebas de hipótesis que permiten abordad datos desequilibrados y efectos aleatorio. Finalmente, contas de una interfaz gráfica (GUI) diseñada de forma intuitiva y fácil de manejar [30].



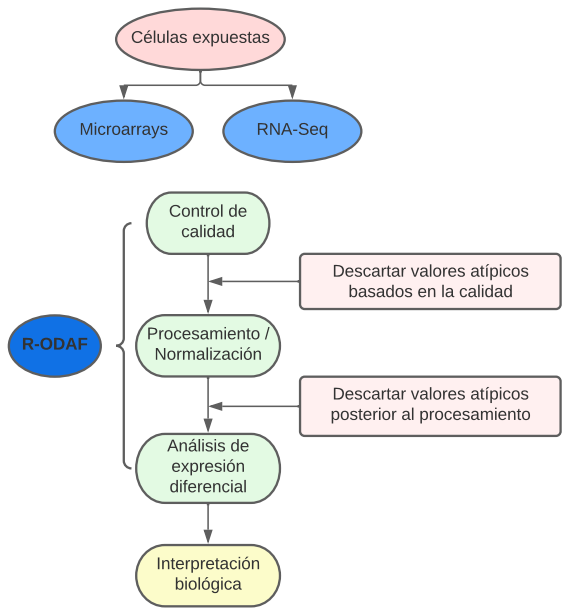
**Figura 6.** Imágenes representativas de DAnTE. (A) Cuadrícula de datos y panel de navegación; (B) Diagrama de cajas; (C) Heatmap de correlación de un conjunto de datos; (D) Resumen de péptidos a proteína del método RRollup [30].

### DanteR

Es una herramienta extensible basad en R para el análisis cuantitativo de datos ómicos, incorporando algoritmos que procesan y normalizan la abundancia de péptidos, análisis estadísticos, visualización multidimensional y crea resúmenes de análisis de publicaciones. Este software está basado en el mencionado anteriormente (DAnTE), per con funciones de más como normalización mediante la descomposción de valores singulares, filtrado de acuerdo a modelos de péptidos de baja calidad, imputación a partir de modelos de probabilidad de valores faltantes y análisis de significancia mejorado a niveles de péptidos y proteínas [31].

### R-ODAF

Es un marco de análisis de datos ómicos para aplicación regulatoria, siendo un protocolo sencillo y explícito con el fin de realizar análisis de datos transcriptómicos de las dos principales plataformas empleadas en toxicogenómica (Microarrays y RNA-Seq). El procesamiento de datos se logra a través de una instalación local de herramientas Bioconda; el código a partir del paquete DESeq2 R que contiene los pasos del filtrado estadístico (está disponible en el siguiente enlace: <https://github.com/R-ODAF>) [32]. El objetivo de este proyecto, fue llegar a un consenso a cerca de un marco de análisis de datos ómicos, para la aplicación regulatoria basada en un análisis de datos rigurosos, donde se detalle los elementos necesarios que debe conformar un experimento ómico e incluye elementos como lo es el diseño experimental, plataformas y transformaciones de datos (normalización, identificación de valores atípicos y análisis estadísticos) [33].



**Figura 7.** Concepto general de la aplicación de R-ODAF.

### iODA

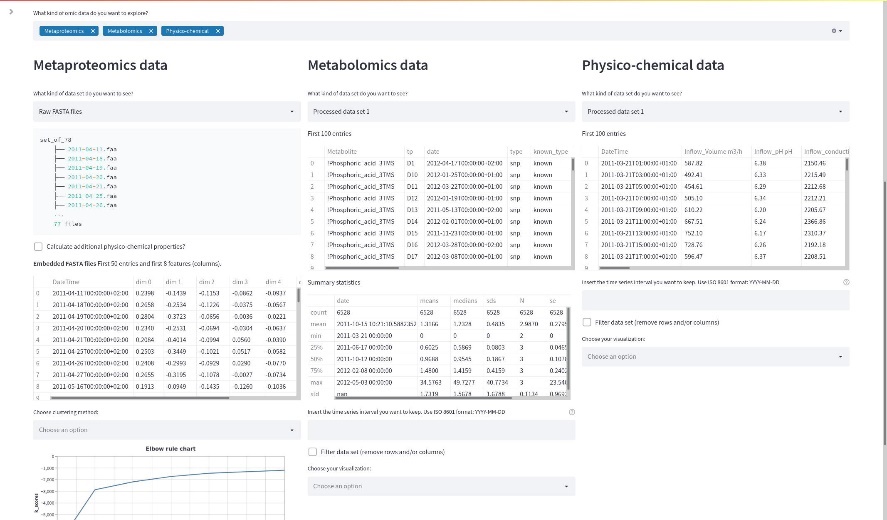
Es una herramienta integrada para analizar la consistencia de la vía del cáncer, partiendo de datos heterogéneos multiómicos para la interpretación a nivel de sistemas de datos multiómicos, en otras palabras, perfiles transcriptómicos e interacciones de proteína-ADN. Esta herramienta demostró un poder analítico a través de estudios de datos ómicos de cáncer de próstata de nivel único y cruzado. El código se encuentra abienrto bajo GNY GPL y puede descargarse desde el enlace <http://www.sysbio.org.cn/iODA> [34].



**Figura 8.** Interface principal de iODA. (1) Paneles funcionales. (2) Botones funcionales.

### MOVIS

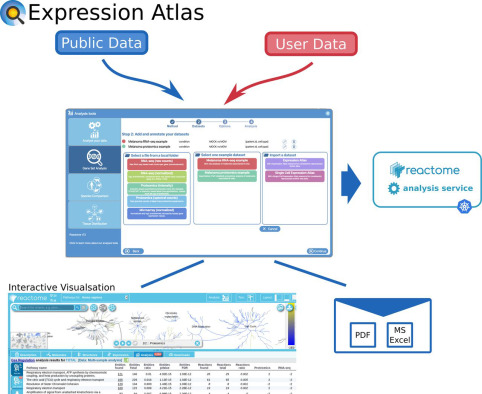
Es un software multiómico para tareas de agrupación, incrustación y visualización de series temporales multimodales, posee una interfaz web fácil de emplear, ayudando en la exploración de dichos datos. Su base es un software de código abierto que se logra ampliar fácilmente adaptándose a diferentes tareas analíticas; su versión en línea está disponible en <https://movis.mathematik.uni-marburg.de/> y en Docker Hub ( <https://hub.docker.com/r/aanzel/movis>) [35].



**Figura 9.** Vista dividida de múltiples ómicas.

### ReactomeGSA

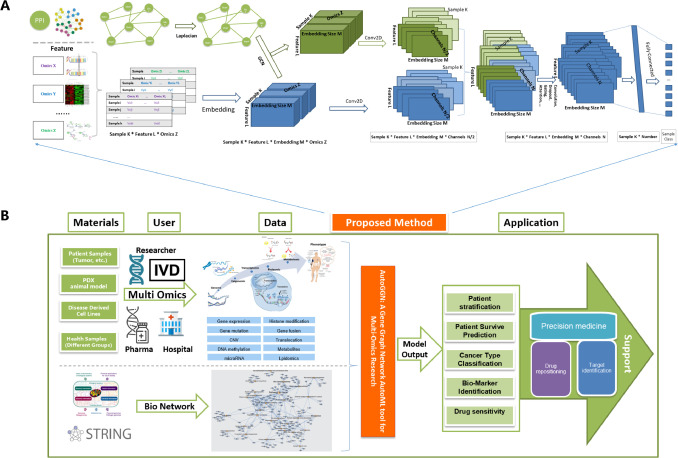
Es una herramienta para análisis comparativo eficiente de rutas multiómicas, se emplea por medio de la interfaz web Reactome y paquete ReactomeGSA R Bioconductor con soporte explícito para datos scRNA-seq. Este sortware reduce en gran medida la barrera técnica para análisis multiómicos, de especies cruzadas y de rutas comparativas [36].



**Figura 10.** Vista dividida de múltiples funciones de ReactomeGSA.

### AutoGGN

Una herramienta que sirve para integrar datos multiómicos con redes de interacción molecular basadas en redes neuronales convolucionales gráficas. AutoGGN tiene el potencial de extraer información de manera más eficaz a través de redes de interacción molecular con datos multiómicos. Además, hace predicciones e identifica genes clave que aportan a la clasificación, otorgando información para el diseño de experimentos biológicos posteriores [37].



**Figura 11.** Estructura del modelo y aplicaciones de AutoGGN.

## Justificación

La genómica, proteómica y metabolómicas se han visto envueltas durante los últimos años, en el área de las ciencias biológicas en el desarrollo de nuevas técnicas, con el fin de estudiar conjuntamente procesos de expresión de los genes y sus productos [1].

Antiguamente se creía que llegar a estudiar un sistema biológico sería algo inalcanzable, sin embargo, con el auge de la bioinformática, la mejora de nuevas técnicas y el avance de la tecnología, se ha logrado interpretar procesos o sistemas enteros pertenecientes a las células, generando cantidades robustas de datos [2].

La aplicación Omics Integrator, es una aplicación que se encarga de integrar diferentes aplicaciones de análisis de datos ómicos, para llegar a resultados como lo son el VolcanoPlot, Análisis de Componentes Principales (PCA), Filtraciones y Mapas de Calor.

A pesar de ser una aplicación con las funcionalidades mencionadas previamente, ésta requiere de algunas mejoras, donde se permita relacionar de manera más sencilla al usuario que la pondrá en práctica con la interfaz, así como se incluyan nuevas funcionalidades con ejemplos y vídeos interactivos con el fin de ser más práctica al momento de emplearse.

Por ende, se empleará el paquete Shiny perteneciente a R, con el objetivo de desarrollar nuevas técnicas de análisis y funciones interactivas en la aplicación Omics Integrator. Con este TFM se buscará dejar un aplicativo que no sólo sea más potente, sino que además tenga la opción de tener mejoras concordes a las futuras nuevas técnicas de análisis vayan surgiendo en esta área de estudio.

### Análisis de enriquecimiento

Para el presente trabajo se hará la instalación de extensión del análisis de enriquecimiento en el software Omics Integrator, esto debido a que permite la extracción de información biológica a partir de una lista de genes procedentes de un experimento, esta información se encuentra como anotaciones derivadas de bases de datos y repositorios públicos, permitiendo calcular un análisis estadístico para explicar qué anotaciones poseen mayor relevancia al momento de exponer esa lista [38]. En la metodología se explicará detalladamente los pasos a seguir para la obtención de este análisis.

# METODOLOGÍA

# Bibliografía

[1] P. Hieter and M. Boguski, “Functional genomics: It’s all how you read it,” *Science (80-. ).*, vol. 278, no. 5338, pp. 601–602, 1997, doi: 10.1126/science.278.5338.601.

[2] D. Tabas Madrid, “Herramientas eficientes para el análisis masivo de datos ómicos,” p. 167, 2018, [Online]. Available: https://eprints.ucm.es/id/eprint/49798/1/T40488.pdf.

[3] P. S. Reel, S. Reel, E. Pearson, E. Trucco, and E. Jefferson, “Using machine learning approaches for multi-omics data analysis: A review,” *Biotechnol. Adv.*, vol. 49, 2021, doi: 10.1016/j.biotechadv.2021.107739.

[4] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, “Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing,” *Nat. Genet.*, vol. 40, no. 12, pp. 1413–1415, 2008, doi: 10.1038/ng.259.

[5] Y. Hasin, M. Seldin, and A. Lusis, “Multi-omics approaches to disease,” *Genome Biol.*, vol. 18, no. 1, 2017, doi: 10.1186/s13059-017-1215-1.

[6] B. Karahalil, “Overview of Systems Biology and Omics Technologies,” *Curr. Med. Chem.*, vol. 23, no. 37, pp. 4221–4230, 2016, doi: 10.2174/0929867323666160926150617.

[7] V. Marx, “Biology: The big challenges of big data TL - 498,” *Nature*, vol. 498 VN-, no. 7453, pp. 255–260, 2013, [Online]. Available: http://dx.doi.org/10.1038/498255a.

[8] Wang Z., Gerstein M., and Snyder M., “RNA-Seq a revolutionary tool for transcriptomics,” *Nat. Rev. Genet.*, 2009.

[9] Z. C. Dong and Y. Chen, “Transcriptomics: Advances and approaches,” *Sci. China Life Sci.*, vol. 56, no. 10, pp. 960–967, 2013, doi: 10.1007/s11427-013-4557-2.

[10] M. A. Marra, L. Hillier, and R. H. Waterston, “Expressed sequence tags--ESTablishing bridges between genomes.,” *Trends Genet.*, vol. 14, no. 1, pp. 4–7, 1998, doi: 10.1016/S0168-9525(97)01355-3.

[11] T. J. Close *et al.*, “A new resource for cereal genomics: 22K Barley GeneChip comes of age,” *Plant Physiol.*, vol. 134, no. 3, pp. 960–968, 2004, doi: 10.1104/pp.103.034462.

[12] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler, “Serial analysis of gene expression,” *Science (80-. ).*, vol. 270, no. 5235, pp. 484–487, 1995, doi: 10.1126/science.270.5235.484.

[13] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee, “Transcriptomics technologies,” *PLoS Comput. Biol.*, vol. 13, no. 5, 2017, doi: 10.1371/journal.pcbi.1005457.

[14] S. Toshiyuki, “Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage,” 2017, [Online]. Available: https://search.proquest.com/docview/1891697797.

[15] V. Romanov, S. N. Davidoff, A. R. Miles, D. W. Grainger, B. K. Gale, and B. D. Brooks, “A critical comparison of protein microarray fabrication technologies,” *Analyst*, vol. 139, no. 6, pp. 1303–1326, 2014, doi: 10.1039/c3an01577g.

[16] Irena Barbulovic-Nad, Michael Lucente, Yu Sun, Mingjun Zhang, Aaron R. Wheeler, and Markus Bussmann, “Bio-Microarray Fabrication Techniques—A Review,” 2008, [Online]. Available: http://informahealthcare.com/doi/abs/10.1080/07388550600978358.

[17] Z. Wang, M. Gerstein, and M. Snyder, “RNA sequencing: advances, challenges and opportunities,” *Nat. Rev. Genet.*, vol. 12, no. 2, pp. 87–98, 2009.

[18] R. L. Stears, R. C. Getts, and S. R. Gullans, “A novel, sensitive detection system for high-density microarrays using dendrimer technology,” *Physiol. Genomics*, vol. 2000, no. 3, pp. 93–99, 2000, doi: 10.1152/physiolgenomics.2000.3.2.93.

[19] M. B. Black *et al.*, “Comparison of microarrays and RNA-Seq for gene expression analyses of dose-response experiments,” *Toxicol. Sci.*, vol. 137, no. 2, pp. 385–403, 2014, doi: 10.1093/toxsci/kft249.

[20] A. Conesa *et al.*, “A survey of best practices for RNA-seq data analysis,” *Genome Biol.*, vol. 17, no. 1, 2016, doi: 10.1186/s13059-016-0881-8.

[21] Y. Kodama, M. Shumway, and R. Leinonen, “The sequence read archive: Explosive growth of sequencing data,” *Nucleic Acids Res.*, vol. 40, no. D1, 2012, doi: 10.1093/nar/gkr854.

[22] M. C. Van Verk, R. Hickman, C. M. J. Pieterse, and S. C. M. Van Wees, “RNA-Seq: Revelation of the messengers,” *Trends Plant Sci.*, vol. 18, no. 4, pp. 175–179, 2013, doi: 10.1016/j.tplants.2013.02.001.

[23] R Core Team *et al.*, “Orchestrating high-throughput genomic analysis with Bioconductor,” *Nat. Methods*, vol. 12, no. 2, pp. 115–121, 2015, [Online]. Available: https://www.r-project.org/.

[24] C. C. Lo and P. S. G. Chain, “Rapid evaluation and quality control of next generation sequencing data with FaQCs,” *BMC Bioinformatics*, vol. 15, no. 1, 2014, doi: 10.1186/s12859-014-0366-2.

[25] N. A. Fonseca, J. Rung, A. Brazma, and J. C. Marioni, “Tools for mapping high-throughput sequencing data,” *Bioinformatics*, vol. 28, no. 24, pp. 3169–3177, 2012, doi: 10.1093/bioinformatics/bts605.

[26] S. Anders, P. T. Pyl, and W. Huber, “HTSeq-A Python framework to work with high-throughput sequencing data,” *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015, doi: 10.1093/bioinformatics/btu638.

[27] Trapnell C *et al.*, “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.,” *Nat. Biotechnol.*, vol. 28, no. 5, pp. 511-515., 2010.

[28] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, “RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays,” *Genome Res.*, vol. 18, no. 9, pp. 1509–1517, 2008, doi: 10.1101/gr.079558.108.

[29] N. J. Nelson, “Microarrays have arrived: Gene expression tool matures,” *J. Natl. Cancer Inst.*, vol. 93, no. 7, pp. 492–493, 2001, doi: 10.1093/jnci/93.7.492.

[30] A. D. Polpitiya *et al.*, “DAnTE: A statistical tool for quantitative analysis of -omics data,” *Bioinformatics*, vol. 24, no. 13, pp. 1556–1558, 2008, doi: 10.1093/bioinformatics/btn217.

[31] T. Taverner *et al.*, “DanteR: An extensible R-based tool for quantitative analysis of -omics data,” *Bioinformatics*, vol. 28, no. 18, pp. 2404–2406, 2012, doi: 10.1093/bioinformatics/bts449.

[32] M. C. Verheijen, T. W. Gant, W. Tong, and F. Caiment, “R-ODAF: Omics data analysis framework for regulatory application,” *Toxicol. Lett.*, vol. 350, p. S124, 2021, doi: 10.1016/s0378-4274(21)00539-7.

[33] M. Verheijen, W. Tong, L. Shi, T. W. Gant, B. Seligman, and F. Caiment, “Towards the development of an omics data analysis framework,” *Regul. Toxicol. Pharmacol.*, vol. 112, 2020, doi: 10.1016/j.yrtph.2020.104621.

[34] C. Yu, X. Qi, Y. Lin, Y. Li, and B. Shen, “iODA: An integrated tool for analysis of cancer pathway consistency from heterogeneous multi-omics data,” *J. Biomed. Inform.*, vol. 112, 2020, doi: 10.1016/j.jbi.2020.103605.

[35] A. Anžel, D. Heider, and G. Hattab, “MOVIS: A multi-omics software solution for multi-modal time-series clustering, embedding, and visualizing tasks,” *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 1044–1055, 2022, doi: 10.1016/j.csbj.2022.02.012.

[36] J. Griss, G. Viteri, K. Sidiropoulos, V. Nguyen, A. Fabregat, and H. Hermjakob, “ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis,” *Mol. Cell. Proteomics*, vol. 19, no. 12, pp. 2115–2124, 2020, doi: 10.1074/mcp.TIR120.002155.

[37] L. Zhang *et al.*, “AutoGGN: A gene graph network AutoML tool for multi-omics research,” *Artif. Intell. Life Sci.*, vol. 1, p. 100019, 2021, doi: 10.1016/j.ailsci.2021.100019.

[38] S. Bauer, S. Grossmann, M. Vingron, and P. N. Robinson, “Ontologizer 2.0 - A multifunctional tool for GO term enrichment analysis and data exploration,” *Bioinformatics*, vol. 24, no. 14, pp. 1650–1651, 2008, doi: 10.1093/bioinformatics/btn250.