2018-10

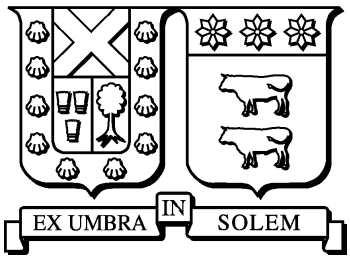
SISTEMA DE PROCESAMIENTO Y VISUALIZACIÓN DE DATOS GENÓMICOS PARA LA PRODUCCIÓN DE UVA

GALAZ VILLARROEL, CLAUDIO ESTEBAN

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INFORMÁTICA

SANTIAGO – CHILE



"SISTEMA DE PROCESAMIENTO Y VISUALIZACIÓN DE DATOS GENÓMICOS PARA LA PRODUCCIÓN DE UVA"

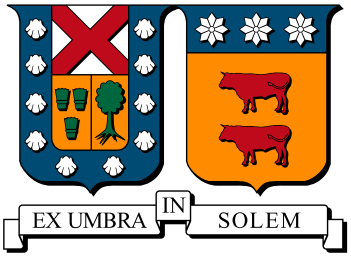
CLAUDIO ESTEBAN GALAZ VILLARROEL

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: ANDRÉS MOREIRA

OCTUBRE 2018

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE



"SISTEMA DE PROCESAMIENTO Y VISUALIZACIÓN DE DATOS GENÓMICOS PARA LA PRODUCCIÓN DE UVA"

CLAUDIO ESTEBAN GALAZ VILLARROEL

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: ANDRÉS MOREIRA

PROFESOR CORREFERENTE: FRANCISCO ALTIMIRAS

OCTUBRE 2018

Para mis padres, Angel y Viviana,

y para sus padres, mis queridos abuelos Teresa, Aida y Juan

Resumen

Chile es uno de los países con mayor producción de uva a nivel mundial, ya sea como fruta de mesa o para la elaboración de vino. Para generar un producto de calidad se hace importante supervisar todo el crecimiento de la vid, además de cosechar la fruta en el momento indicado. Actualmente, el desarrollo de la planta es controlado de manera visual, pero la disminución en costos de la secuenciación de ARN (RNA-Seq) motivan otro acercamiento a esta problemática. En este trabajo se utilizan datos de *Vitis vinifera* ya existentes, provenientes de ArrayExpress, para generar un pipeline RNA-Seq rápido que incluya el análisis de expresión génica de las etapas fenológicas de la planta. Se utiliza un flujo de trabajo que considera las herramientas FastQC, Trimmomatic, Salmon, edgeR y PANTHER DB. Además se genera el prototipo para una plataforma de visualización, utilizando Shiny, que muestra un resumen del análisis RNA-Seq realizado. El prototipo se encuentra disponible en `https://vitis-deg.shinyapps.io/shiny-app/`.

Palabras clave—Secuenciación de ARN, RNA-Seq, *Vitis vinifera*, análisis de expresión génica, visualización de datos.

Abstract

Chile is one of the top producers of grape worldwide, either as table fruit or winemaking. To elaborate a quality product it is important to monitor the whole grapevine growth process, in addition to harvesting the fruit at the exact time. Currently, the grapevine development is controlled visually, but the decrease in RNA sequencing costs motivates another approach to this problem. In this work, existing data for *Vitis vinifera* is used, provenient from ArrayExpress, to generate a fast RNA-Seq pipeline that includes gene expression analysis for phenologic stages of the grapevine. The tools used in the workflow are FastQC, Trimmomatic, Salmon, edgeR and PANTHER DB. Also, a Shiny-based prototype for a visualization platform is presented, with a summary for RNA-Seq results. The prototype is available in `https://vitis-deg.shinyapps.io/shiny-app/`.

Keywords—RNA-Seq, RNA sequencing, *Vitis vinifera*, gene expression analysis, data visualization.

Índice de Contenidos

Índice de tablas

Índice de figuras

Glosario

- ADN: molécula de cadena doble compuesta de nucleótidos, responsable de llevar el material genético necesario para el desarrollo de un organismo vivo.
- ARN: molécula compuesta de nucleótidos, responsable de la síntesis de proteínas y a veces de la transmisión de información genética.
- ARNm: molécula de cadena simple, complementaria a una de las cadenas del ADN, y que gracias al proceso de traducción se convertirá en proteína.
- Brix: valor que indica el contenido de azucar en una solución acuosa.
- Cola poli(A): trozo del ARN que consiste en solo adenina.
- Contenido GC: es el porcentaje de las bases nitrogenadas (en el ARN o ADN de un organismo) que corresponden a guanina y citosina.
- Empalme: etapa en donde los intrones son filtrados, luego eliminados, y los exones se unen entre ellos.
- Exón: parte del código genético que sirve para codificar el ARNm.
- Fenología: estudio de los efectos que tiene el clima y el entorno sobre el ciclo de vida de un ser vivo (animal o vegetal).
- Fold-change: medida que describe cuánto cambia una cantidad en comparación a otro valor. Por ejemplo para dos valores A y B, el fold-change de B con respecto a A será la razón $\frac{B}{A}$.

- Genes downregulated: aquellos genes que su expresión se ve disminuida en cantidad en comparación a otra condición biológica[1].
- Genes upregulated: son los genes que su expresión se ve incrementada en cantidad en comparación a otra condición biológica.
- Genómica funcional: campo de la biología molecular que busca utilizar los datos provenientes de proyectos de genómica/transcriptómica para caracterizar las funciones e interacciones de los genes.
- Grafo de Bruijn: es el grafo que se obtiene al tomar todos los strings que componen un alfabeto finito de largo l como vertices, y agregando aristas entre aquellos vértices que se superponen en $l - 1$.
- Indel: polimorfismo de inserción y deleción de secuencias en una sección del código genético; es la segunda mutación más común del genoma humano.
- Intrón: sección del código genético que no es codificante.
- K-mer: todos los posibles substrings de largo k que están contenidos en un string. En genómica computacional, un k-mer se refiere a todas las posibles subsecuencias que se pueden obtener a partir de una secuencia de ADN.
- Nucleótido: es uno de los componentes estructurales del ADN y ARN. Consiste en una base nitrogenada (adenina, guanina, timina o citosina), una azúcar y un ácido fosfórico.
- OIV: Organización Internacional de la Viña y el Vino
- Paired-end sequencing: secuenciación de ambos extremos de un fragmento de ARN.
- Péptido: molécula formada por la unión de varios aminoácidos. Son responsables de un gran número de funciones en un organismo.
- Pipeline: un grupo de elementos de procesamiento de datos conectados en serie, donde el output de uno sirve como input del siguiente.

[1]Para este trabajo, aquella condición será el estado fenológico.

- Polipéptido: unión de hasta 100 aminoácidos. Son responsables de un gran número de funciones en un organismo.
- Proteína: unión de más de 100 aminoácidos. Son responsables de un gran número de funciones en un organismo.
- Single-end read: secuenciación de un extremo del fragmento de ARN a otro, en una sola dirección.
- Transcriptoma: colección de todas las moléculas de ARN en una o muchas células.
- Vinificación: producción del vino, que comienza con la selección de la uva, pasando por su fermentación en alcohol, y termina con el embotellado del producto final.

Introducción

Los organismos vivos se encuentran definidos por material genético, principalmente caracterizado por el ADN, el cual es una pieza fundamental ya que contiene las instrucciones para el desarrollo del ser vivo. Además existe otra molécula llamada ARN, que es obtenida luego de copiar el ADN mediante el proceso de transcripción, y que una vez codificada servirá para convertirse en proteínas, las que cumplirán una gran cantidad de funciones en el organismo.

Por eso es que se han desarrollado diversas técnicas a lo largo de los años para obtener el ARN de un ser vivo y entender el funcionamiento del organismo en ese mismo instante de tiempo. Algo así como tomar una fotografía del material genético para luego poder observar cuáles son los genes que están participando en el desarrollo.

En el tiempo se ha trabajado para seguir mejorando aquellos métodos, llegando a la secuenciación del ARN (RNA-Seq) que en los últimos años se ha convertido en la solución más preferida por los investigadores. Esta metodología entrega gran poder en el estudio de los seres vivos, y es posible aplicarla en la caracterización de enfermedades u otras condiciones biológicas. Una aplicación de este tipo de trabajos puede ser la identificación de etapas fenológicas en plantas, distinguiendo así cuáles son los genes protagonistas en cada etapa de crecimiento.

Paralelamente, la producción de uva y la industria del vino en los últimos años se han convertido en una de las actividades más importantes para Chile, ubicando al país en octavo lugar a nivel mundial gracias a su nivel de producción de aquella bebida. Esto hace que se observe con atención la obtención de la uva en todo su proceso de crecimiento, siendo de gran importancia la etapa fenológica en la que se encuentra la vid.

La disminución en costos de secuenciación de ARN y la gran importancia económica que tienen las plantaciones de uva en nuestro país motivan este trabajo, que busca evaluar el estado del arte del análisis RNA-Seq en plantas, en base a eso generar un flujo de trabajo RNA-Seq rápido, y por último desarrollar un prototipo de plataforma web que presente un resumen del análisis al agricultor.

En el capítulo 1 de este documento se estudia el contexto de genómica y bioinformática sobre el cual se trabaja, además de presentar los objetivos generales y específicos del proyecto. En el capítulo 2 se entrega el estado del arte alrededor de las herramientas bioinformáticas disponibles para el análisis RNA-Seq. El capítulo 3 presenta la solución propuesta para abordar los objetivos planteados. El capítulo 4 habla de la ejecución y se comentan los resultados obtenidos, para finalmente dar paso a las conclusiones de esta memoria.

Capítulo 1

Definición del Problema

La vid (*Vitis vinifera*) es una de las plantas más cultivadas en el mundo, teniendo 7,5 millones de hectáreas de terreno vitícola y alcanzando 76,8 millones de toneladas de uva a nivel mundial en el año 2016, de acuerdo al último balance de la OIV [4]. La producción de uva tiene como fin su comercialización como fruta, jugo o pasas, aunque en su mayoría es utilizada para su fermentación en vino.

En las últimas décadas la industria del vino ha sido una de las más importantes para Chile, donde la superficie vitivinícola estimada en 2016 fue de 214 mil hectáreas y aquella destinada exclusivamente a la producción de vino fue de 141,9 mil hectáreas [5]. Además, según la OIV Chile se encuentra en octavo lugar a nivel mundial en cuanto a producción de vino, con 10,1 millones de hectolitros en el último año [6]. Es por esto que, con el fin de entregar resultados de calidad, la producción de uva es una actividad que requiere mucha rigurosidad y precisión en todo su desarrollo, de principio a fin. La cosecha de la uva (llamado vendimia en el caso de las uvas destinadas a la producción de vino) se realiza una vez alcanzada la maduración de la planta. Es un proceso que requiere mucha precisión, puesto que la calidad de la uva obtenida depende en gran parte del momento en el cual se elige cosechar, afectando directamente al producto final.

Con el objetivo de obtener una producción de uvas de calidad, se hace importante supervisar el crecimiento de la planta y cómo el clima y el entorno afectan el desarrollo de ésta. El

estudio de aquello es conocido como fenología. Para controlar el crecimiento de la vid, durante el último siglo se han desarrollado distintos sistemas de referencia, los cuales entregan información de los distintos estados fenológicos de la planta.

El año 1952, Baggiolini plantea la primera escala para describir las etapas de crecimiento de la vid [7], con el fin de permitir una buena sincronización en cuanto a medidas de protección de la planta. Esta escala se caracteriza por tener poca precisión y por describir solamente hasta la etapa de cuaje. Eichhorn y Lorenz en el año 1977 publican una nueva forma de describir el crecimiento de la planta, la cual mejora las deficiencias presentes en el método anterior, con un sistema más detallado que cubre 22 etapas desde el comienzo hasta la caída de las hojas [8].

En 1993, Baillod y Baggiolini [9] modificaron la escala creada por Baggiolini en 1952, añadiéndole 7 nuevas etapas que van desde cuaje hasta la caída de hojas. Lorenz *et al.* [10] publican en 1994 el sistema BBCH modificado, el cual nace como un modelo para la Unión Europea y adapta el método BBCH estándar para la vid. Finalmente, en el año 1995, Coombe [11] presenta una modificación a la escala de Eichhorn y Lorenz, la cual además de definir claramente las etapas más importantes, entrega un mayor detalle sobre las etapas intermedias.

En la actualidad se utilizan estas metodologías como sistema de referencia para determinar de manera visual el estado fenológico, poniendo especial atención a las características físicas de la planta. Lo anterior se puede traducir en que esta tarea depende completamente de la percepción humana en ese momento, lo que lo convierte en un método inmediato pero poco preciso. En este trabajo se explora la posibilidad de ir más allá de la mera observación y propone un camino para determinar el estado fenológico a partir de las propiedades genéticas de la planta en distintas etapas de su crecimiento.

1.1 Genómica y bioinformática

Para comprender esta memoria se hace necesario introducir las bases acerca de genes, genética y genómica.

1.1.1 Genes

El término “gen” hoy posee una definición no muy precisa. Inicialmente, esta palabra era utilizada para indicar la unidad de herencia de un fenotipo, y luego este significado continuó siendo usado en un contexto no científico para referirse a la expresión de cierta característica en un individuo; por ejemplo, ”gen del cabello rubio”, o ”gen de los ojos azules”. Con el avance en investigaciones fue posible establecer que las proteínas están compuestas por muchos aminoácidos, y a su vez, aquellos aminoácidos se encontraban codificados por regiones cromosómicas que podían ser identificadas genéticamente [12] (ver Figura 1.1).

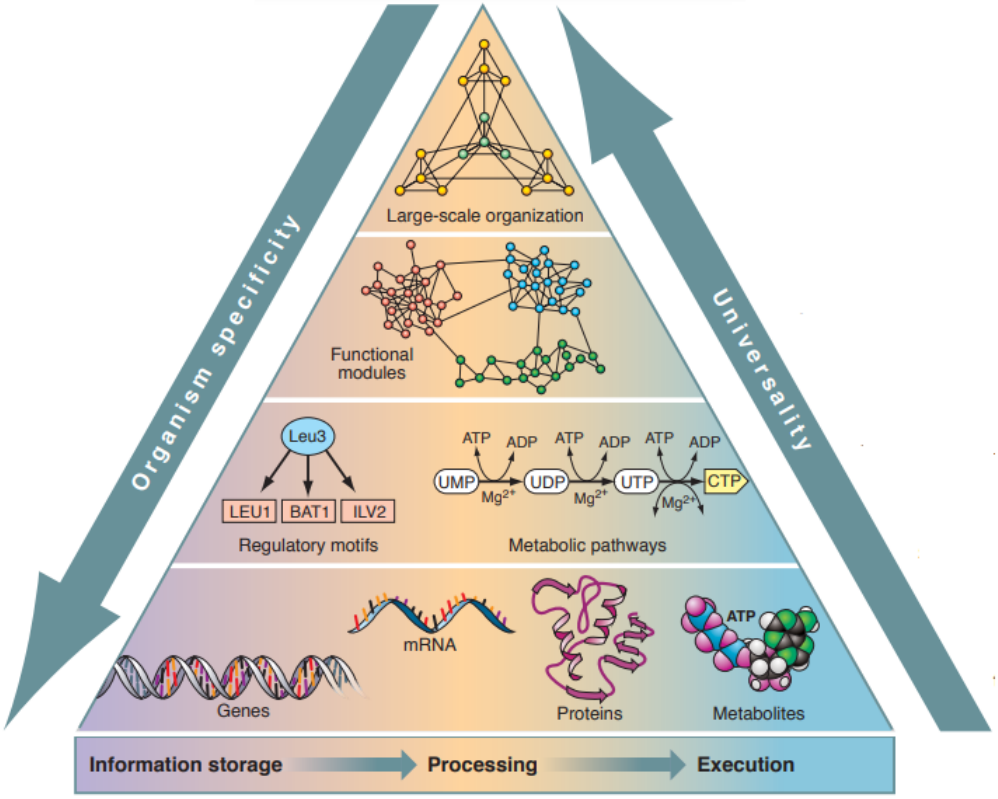


Figura 1.1: Pirámide de complejidad de la vida, según Oltvai y Barabási [1].

Finalmente, junto con la aparición de la secuenciación de ADN fue posible considerar al gen como una región molecular compuesta de nucleótidos, pudiendo así ser identificado en la región cromosómica que codifica al polipéptido. En organismos eucariontes el gen es una sección compuesta de exones e intrones, en donde solamente los exones sirven para codificar la proteína y los intrones son la parte no codificante de esta.

1.1.2 Genética

Como su nombre lo indica, genética es el nombre que recibe el estudio de los genes y el rol que ellos juegan en herencia. Durante la historia, la herencia siempre ha sido un tema de interés para las personas. Incluso desde antes que la biología existiera como una rama de la ciencia, la gente buscaba formas para mejorar su cultivo de plantas o la cruza de animales. No fue hasta mediados del siglo XIX, y reconocido de manera póstuma a comienzos del siglo XX, que la genética fue planteada formalmente con el trabajo de Gregor Mendel. Estos estudios definen un conjunto de principios y procedimientos analíticos que marcan el inicio de la genética como el estudio de herencia, genes y variación genética [13].

1.1.3 Genómica

El conjunto completo de ADN dentro de un organismo es llamado genoma. En cuanto a la especie humana, cada célula dentro del cuerpo contiene una copia completa de aproximadamente tres mil millones de pares de bases, conformando entre todas el genoma humano [14]. Con el desarrollo de técnicas de ADN recombinante, en 1977 Fred Sanger *et al.* [15] marcan la aparición de la genómica mediante el desarrollo de un nuevo método de secuenciación de ADN, pudiendo así encontrar el genoma de 5400 nucleótidos del virus $\phi X174$ [16, capítulo 21]. Con el tiempo fueron secuenciados distintos otros genomas, pero como el proceso era lento y laborioso, el estudio se limitó a organismos con genomas pequeños. A mediados de la década de 1980, investigadores interesados en el genoma humano utilizaron técnicas de ADN recombinante para mapear secuencias de ADN en ciertos cromosomas. Así se estiman, aunque incorrectamente, aproximadamente unos 100.000 genes en el genoma humano, y es posible notar que con los métodos existentes hasta esa fecha, encontrar el genoma humano sería una tarea larga y difícil.

En las tres décadas siguientes, con el aumento del poder de cómputo y el mejoramiento en las tecnologías de secuenciación de ADN, fue posible considerar el estudio de genomas eucariontes más largos y complejos, incluyendo finalmente los tres mil millones de genes que comprende el genoma humano [16, capítulo 21].

1.1.4 Bioinformática

La bioinformática es el campo de la ciencia que usa herramientas computacionales (hardware y software) y aplicaciones matemáticas para el ordenamiento y análisis de datos biológicos como estructura de genes, secuencias de genes, expresión génica y estructura y función de proteínas [16]. Según Ramsden [17], bioinformática puede definirse simplemente como ”la aplicación de ciencias de la información en la biología”. Entre algunas aplicaciones, las tecnologías bioinformáticas son usadas para identificar correlaciones entre secuencias de genes y enfermedades, para predecir estructuras de proteínas a partir de secuencias de aminoácidos, para aportar en el diseño de nuevos fármacos, o para adaptar tratamientos a pacientes a partir de su secuencia de ADN.

1.2 Proceso de transcripción

Las células procariotas son aquellos organismos unicelulares que carecen de un núcleo definido, teniendo todos sus componentes (incluyendo el material genético) repartidos en el citoplasma. El resto de los organismos vivos se conocen como eucariontes, y están compuestos por una o más células del mismo nombre. Tanto en aquellas como en las procariotas, el ADN codifica toda la información necesaria sobre las propiedades y funciones de cada célula; cierta secuencia del material genético se copia en el ARN y luego se traducirá en proteína.

A diferencia de los organismos procariotas, las células eucariontes contienen gran parte del material genético dentro de su núcleo y es aquí donde se realiza el proceso de transcripción. En palabras simples, la transcripción es el proceso en el que la secuencia de ADN es copiada para formar una molécula de ARN. Una enzima llamada ARN polimerasa tiene como tarea copiar el ADN, para lo cual separa sus dos hebras y se acopla a una de ellas, la cadena molde, con el fin de codificar el ARN. El resultado de esta etapa es una molécula (o transcrito) de ARN llamada pre-ARNm, o pre ARN mensajero, la cual está compuesta de nucleótidos en donde cada uno de ellos es el complemento de la base nucleotídica en la hebra de ADN (ver

Figura 1.2). Es decir, si en la hebra de ADN se encuentra presente la secuencia TACTAG[1], la hebra de ARN codificará el complemento de aquellas bases[2] como CUAGUA[3] (AUGAUC en la dirección opuesta).

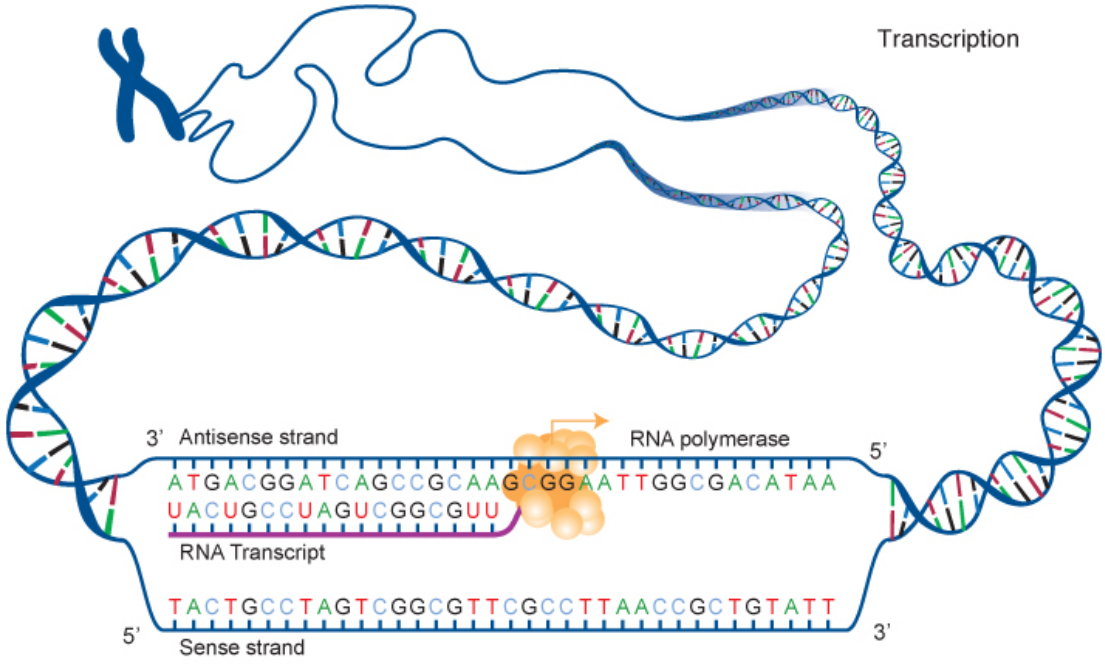


Figura 1.2: ARN polimerasa copiando la cadena de ARNm a partir de la cadena molde [2].

El siguiente paso es transformar el transcrito primario (pre-ARNm) en una molécula que más adelante pueda convertirse en proteína. Actualmente el pre-ARNm está compuesto de dos secciones importantes llamados exones e intrones, y que se encuentran presentes de forma intercalada a lo largo de la hebra. Los exones son aquellas partes codificantes y que serán útiles para que la proteína pueda cumplir la función que se espera de ella. Por el contrario, las secciones que reciben el nombre de intrones poseen material que no sirve y es inútil para el proceso de traducción (de ARNm a proteína). Es por esta razón que los intrones son filtrados, eliminados, y los exones se unen entre ellos en la etapa llamada empalme, o *splicing*. Finalmente de lo anterior resulta una cadena llamada ARN mensajero (abreviado ARNm) o

[1] A: adenina, T: timina, C: citosina, G: guanina, U: uracilo.

[2] Las parejas de nucleótidos que sirven de complemento para el otro en el caso del ADN son: A-T y C-G. Para el ARN son: A-U y C-G.

[3] El complemento de adenina en el ARN es uracilo en lugar de timina.

ARNm maduro, el que luego saldrá del núcleo celular y se dirigirá al ribosoma donde comienza la etapa de traducción, que tiene como finalidad convertir aquel ARN mensajero en péptido o proteína.

1.3 Análisis de expresión diferencial

A partir de muestras biológicas de células o tejido de un organismo es posible realizar un análisis de su transcriptoma, lo que es útil para lograr el entendimiento de ciertas variaciones fenotípicas en biología, como lo son las enfermedades [18]. A partir de dos muestras de material genético en condiciones biológicas diferentes, por ejemplo una extraída de un organismo sano y la otra obtenida de uno enfermo, es posible aplicar distintas técnicas como microarrays o RNA-Seq para encontrar cuáles son los genes expresados para cierta condición biológica; esto se encuentra evaluando cuáles genes son los que presentan una abundancia considerable comparado con la otra muestra. Este estudio es conocido como análisis de expresión diferencial.

1.3.1 Microarrays

Durante las décadas pasadas, los microarrays (chips de ADN en español) tuvieron un papel importante en el análisis de expresión diferencial, ya que hasta hace poco tiempo era una de las técnicas más usadas para encontrar genes diferencialmente expresados. Se extraen muestras de ARNm de un organismo en condiciones normales y de otro enfermo, tiñendo cada una de un color distinto. Por ejemplo, la muestra en condiciones normales será de color azul y la muestra enferma de color amarillo. Luego se combinan y se aplican al chip microarray, donde mediante análisis de imagen se pueden observar los niveles de hibridación de los genes en estudio mediante una señal de fluorescencia. Si el color observado para cierto gen es amarillo, entonces ese gen se encuentra expresado con mayor intensidad en la muestra enferma. De forma análoga, si para un gen se observa un color azul entonces ese gen estará expresado mayormente en el organismo de condiciones normales. Si el color que se obtiene para un determinado gen es el verde o alguna mezcla entre azul y amarillo, entonces aquel

gen se expresa de manera similar en ambas muestras.

1.3.2 Secuenciación de ARN (RNA-Seq)

En [19] se describe a RNA-Seq como un conjunto de métodos experimentales y computacionales utilizados para determinar la identidad y abundancia de secuencias de ARN en muestras biológicas. La metodología involucra el aislamiento del ARN de la célula o tejido, la preparación de la librería para el ARN en la muestra, la secuenciación química de la librería, y el análisis bioinformático de los datos obtenidos.

La secuenciación de ARN (RNA-Seq) tiene como objetivo identificar la secuencia, estructura y abundancia de las moléculas de ARN en una muestra biológica [19, capítulo 1]. La palabra secuencia se refiere al orden específico que tendrán las bases A, C, G y U. Estructura del ARN hace referencia a la organización que tiene la molécula en cuanto a elementos como el promotor, uniones intrón-exón, regiones sin traducir y la cola poli(A). Por último, con abundancia en el ARN se refiere a la cantidad numérica que tendrá cada secuencia (gen) en la muestra.

A diferencia de la secuenciación de ADN, la cual es útil para obtener un perfil genético de un organismo, la secuenciación de ARN entrega solo aquellas secuencias que se encuentran activamente expresadas en el organismo en el instante de tomar la muestra. Para ponerlo de una forma más cotidiana, obtener la secuencia de ARN de un organismo es como hacer una foto de su material genético en ese momento.

RNA-Seq destaca por sobre los métodos anteriores como microarrays gracias al alto rendimiento que ofrecen las plataformas de secuenciación actuales, la sensibilidad que entregan las nuevas tecnologías de secuenciación, y la habilidad de descubrir transcritos nuevos, modelos de genes y especies de ARN pequeños no codificantes [19, capítulo 1]. En cuanto al análisis de expresión diferencial, el método de RNA-Seq difiere del realizado con microarrays en que los datos observados usando el primero son conteos discretos, mientras que la medición con microarrays entrega un resultado continuo que proviene de la señal fluorescente obtenida [19, capítulo 8].

1.4 Motivación y objetivos

En las últimas décadas los costos de secuenciación del ADN han disminuido considerablemente, costando $150 mil dólares americanos en el año 2001, y llegando a costar cerca de $1.000 dólares el año pasado [3] (ver Figura 1.3). Por otra parte, el mejoramiento en los algoritmos bioinformáticos ha acortado los tiempos de ejecución del análisis RNA-Seq considerablemente. Todos estos avances tecnológicos permiten realizar mejores análisis en un menor tiempo y con un costo notablemente más bajo.

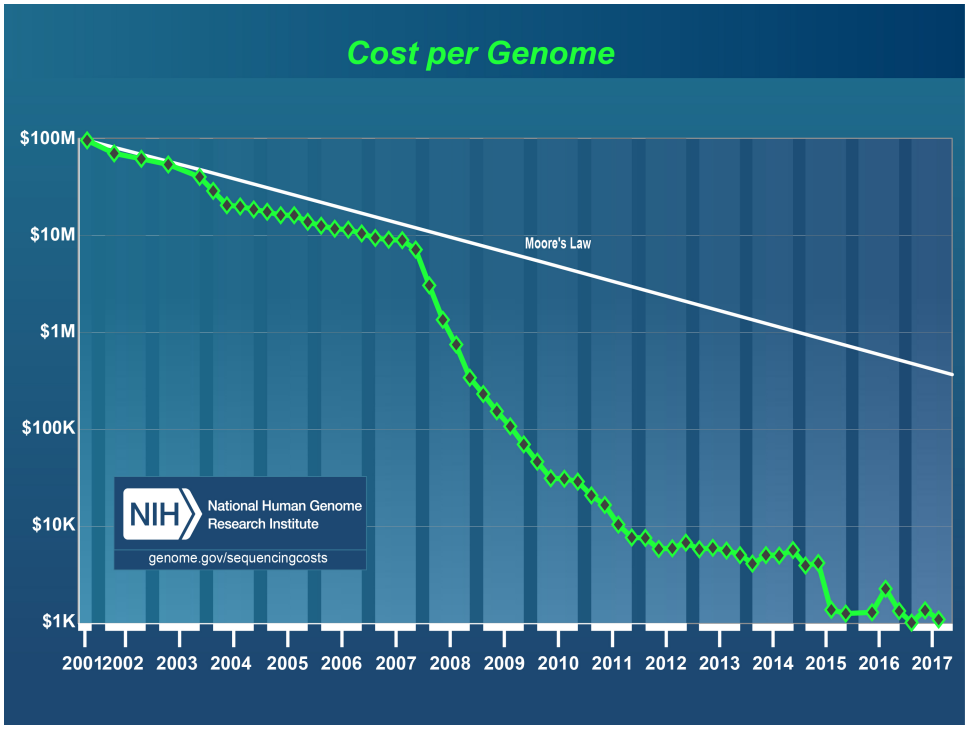


Figura 1.3: Costos de secuenciación de un genoma de tamaño humano [3].

La disminución en el costo de secuenciación de ARN, junto con el desarrollo de mejores algoritmos y el mejoramiento en las tecnologías de hardware computacional podrían permitir un análisis rápido del ARN de los organismos vivos, en específico de la *Vitis vinifera*.

Es por estos motivos que se desarrolló este estudio como trabajo en conjunto al proyecto de clasificación de fenología de la uva que se trabaja en Telefónica I+D, proyecto liderado por Francisco Altimiras, y del cual este trabajo forma parte.

1.4.1 Objetivo general

El objetivo principal de este trabajo es entregar el primer prototipo de un sistema de monitoreo aplicable a la industria agrícola mediante el análisis rápido de datos RNA-Seq, específicamente con datos provenientes de la *Vitis vinifera*.

1.4.2 Objetivos específicos

- Realizar un estudio del estado del arte en cuanto a las herramientas bioinformáticas que sirven al análisis RNA-Seq.
- Generar un flujo de trabajo para la analítica de datos RNA-Seq, específicamente para el análisis rápido de secuenciaciones de genes de la planta *Vitis vinifera*.
- Desarrollar un prototipo de plataforma de visualización basado en la web que muestre un resumen del análisis RNA-Seq de la *Vitis vinifera*.

Capítulo 2

Estado del Arte

2.1 Análisis de datos RNA-Seq

Desde el año 2008 [20] se utilizan técnicas de secuenciación de alto rendimiento o de nueva generación (NGS) para la secuenciación de ARN (RNA-Seq), con el objetivo de mostrar la presencia y cantidad de secuencias de ARN en una muestra biológica en un determinado instante de tiempo. En los últimos años una gran cantidad de procedimientos de análisis RNA-Seq han sido publicados, por lo que no existe una única forma óptima de realizar esta tarea [21].

A continuación se presentan las etapas más importantes en un procedimiento de análisis RNA-Seq, y las herramientas bioinformáticas utilizadas actualmente en cada una de ellas.

2.1.1 Obtención de datos

Una vez que la muestra de ARN ha sido aislada de la célula o tejido, el siguiente paso es secuenciar. Existen múltiples plataformas con las que se puede realizar la secuenciación del ARN, donde entre las más importantes se encuentran SOLID, Roche 454 e Illumina, siendo esta última la compañía líder en este mercado gracias a su precisión y a la secuenciación de

lecturas largas (long-reads). En la última década han aparecido otros competidores como Ion Torrent y Pacific Biosciences, pero Illumina sigue siendo el método más utilizado para la secuenciación de lecturas largas de ARN [21]. Para esta memoria los datos que serán utilizados serán de Illumina y provienen de la base de datos ArrayExpress [22], la cual es proveída por el Instituto Europeo de Bioinformática (EMBL-EBI) y que almacena datos de experimentos de genómica funcional con el fin de ponerlos a disposición de futuras investigaciones.

2.1.2 Control de calidad

Luego de la obtención de datos, estas secuenciaciones deben pasar por una etapa de control de calidad, o quality check (QC). Las *raw reads* corresponden a aquellas lecturas obtenidas directamente desde Illumina, las cuales son almacenadas como un archivo de texto de formato FASTQ. Cada secuencia dentro del texto consta de cuatro líneas: identificador, secuencia de nucleótidos, símbolo '+' seguido de una descripción (actualmente sin uso), y en último lugar valores de calidad para los nucleótidos de la línea 2. El control de calidad en las *raw reads* involucra revisar la calidad de secuencias, contenido GC, presencia de adaptadores, lecturas duplicadas o k-mers sobre representados. Los niveles aceptables de lecturas duplicadas, k-mers o contenido GC son específicos del organismo que está siendo estudiado, pero ellos deben ser homogéneos para las distintas muestras dentro del experimento.

Una herramienta que ha sido muy popular en la literatura y que sigue siendo usada en la actualidad [23] [24] es FastQC [25], la cual entrega reportes html con distintas métricas que dan cuenta de la calidad de los datos obtenidos. El reporte generado se divide en distintas secciones que informan de la calidad mediante gráficos, usando el valor Phred [26] para cuantificar la calidad. Los parámetros más comunes se presentan a continuación.

- Calidad de las bases: esta sección muestra una visión general de los rangos de las calidades de todas las bases presentes en el archivo FASTQ. Generalmente las calidades de las bases varían entre 0 y 40 [19, capítulo 3].
- Calidad de las secuencias: este reporte permite ver si un subgrupo de secuencias tiene una baja calidad entre la totalidad de ellas.

- Contenido de bases en secuencias: muestra una curva para cada base, mostrando la proporción de ella en cada secuencia. En general las curvas deben ser paralelas y no desbalanceadas comparando entre las bases.

- Contenido GC en las secuencias: mide el contenido GC a lo largo de cada secuencia y compara esta curva con otra que muestra la distribución teórica de aquel valor. Si existe mucha diferencia entre ambas distribuciones puede existir contaminación en los datos.

- Contenido de K-mer: esta sección toma en consideración que cualquier fragmento pequeño de una secuencia (k-mer) no debería tender a aparecer en una posición determinada. Con esto en cuenta, FastQC muestra una lista con aquellos k-mers que tienen algún sesgo en su posición, y además muestra un gráfico con los seis k-mers más sesgados. De forma que tenga una rápida ejecución, se analiza el 2 % de los datos y luego los resultados son extrapolados al resto.

- Secuencias duplicadas: calcula el grado de duplicación para cada secuencia, y crea un gráfico mostrando la cantidad relativa de secuencias con distinto grado de duplicación.

FastQC es una buena herramienta para realizar QC en datos obtenidos desde Illumina. En caso de que las secuencias hayan sido obtenidas a partir otra plataforma, una herramienta que se recomienda es NGSQC [27].

2.1.3 Trimming

A continuación viene la etapa llamada trimming. No existe consenso para determinar si efectivamente este proceso será útil para el análisis RNA-Seq. En un estudio del año 2016, Williams et al. [28] plantean que el uso de herramientas de trimming queda a disposición de los investigadores puesto que no existe una diferencia notoria entre el análisis con trimming o sin este. También se sugiere que el uso de esta tarea sea exclusivamente en secuenciaciones largas (como 100 o 150 bases de largo), puesto que de esta manera, luego de hacer el trimming, las lecturas seguirán siendo largas. Algunas herramientas disponibles en la literatura son SolexaQA [29], cutadapt [30] y Trimmomatic [31], que tiene el artículo más citado.

Según Del Fabbro et al. [32] no existe una herramienta que sea la mejor para realizar trimming, y recomiendan que la elección del programa se realice evaluando el tradeoff entre la pérdida de lecturas y la mejora en calidad.

2.1.4 Alineamiento (*mapping*)

La siguiente etapa del proceso de análisis RNA-Seq consiste en alinear las lecturas obtenidas a un genoma o transcriptoma de referencia con el objetivo de estimar de dónde vienen. La dificultad está en identificar correctamente las uniones de empalme (*splice junctions*), sobre todo porque pueden existir diferencias con el genoma de referencia o errores en las secuenciaciones obtenidas.

El alineador más popular es TopHat [33], o TopHat2 [34] en su versión más reciente. El alineamiento de esta herramienta consiste en hacer mapping a aquellas lecturas que no abarcan una unión, identificando así los exones. Luego, las lecturas que quedaron sin alinear son divididas y alineadas a diferentes exones. La desventaja de esta herramienta es el tiempo de ejecución, la cual es muy alta comparada con otras herramientas [35], y puede llegar hasta cinco horas [36]. La herramienta HISAT2 [37] tiene mejores tiempos de ejecución gracias a un algoritmo de búsqueda rápida, pero tiene una sensibilidad (recall) pobre, comparable con TopHat2 [35].

Distintos otros alineadores aparecen en la literatura. Las herramientas GSNAP [38], PALMapper [39] y MapSplice [40] son útiles para identificar SNP (polimorfismo de un solo nucleótido) e indels (inserción - deleción), MapSplice [40] y STAR [41] [42] servirán para detectar uniones de empalme no canónicos, las herramientas GEM [43] y STAR [41] [42] son útiles para realizar un *mapping* rápido.

En general, STAR [41] [42] se encuentra muy bien evaluada en la literatura ya que incluso utilizando la configuración de parámetros por defecto tiene un buen rendimiento en precisión y rapidez [44], lo que la convierte en una herramienta confiable para la etapa de alineamiento. La única desventaja es la cantidad de memoria RAM que requiere para ser ejecutada, que ronda los 32 GB.

2.1.5 Cuantificación

Después de haber alineado las secuencias al genoma de referencia, el siguiente paso será realizar el conteo de la cantidad de secuencias alineadas para cada gen. Para realizar esto existen herramientas como HTSeq-count [45] o featureCounts [46]. HTSeq [45] es una librería de Python para realizar análisis RNA-Seq, la cual incluye el script HTSeq-count, y que realiza la cuantificación a partir de un archivo GTF (*Gene Transfer Format*) que incluye las coordenadas de los exones del genoma. En caso de realizar el análisis RNA-Seq en R, la librería Rsubread tiene entre sus componentes el programa featureCounts [46], el cual funciona de igual forma que HTSeq-count.

Otra forma para encontrar conteos es cuantificando los niveles de expresión de los transcritos, alineando o pseudo-alineando parte de las lecturas al transcriptoma [47]. La idea es cuantificar los transcritos alternativos dentro de cada gen, para luego combinarlos y encontrar el total de conteos para cada gen. Las herramientas más citadas que utilizan esta forma de cuantificar son Cufflinks [48], eXpress [49], Flux Capacitor [50], RSEM [51], kallisto [52] y Salmon [53]. Teng [47] encuentra que todas las herramientas mencionadas, a excepción de Flux Capacitor, tienen un rendimiento muy similar entre ellas.

Pseudo-alineamiento + cuantificación

Salmon y kallisto son las herramientas más recientes en el alineamiento a un transcriptoma. En el caso de kallisto el proceso no es un alineamiento propiamente tal, sino que realiza la cuantificación mediante un proceso llamado pseudo-alineamiento en el cual forma un grafo de Bruijn a partir del transcriptoma de referencia, y luego alinea los k-mers de la lectura en los nodos del grafo [52]. Por otra parte, Salmon realiza un proceso de tres partes: un *mapping* ligero al transcriptoma de referencia; una fase online que estima los niveles de expresión iniciales; y una fase offline que afina aquellas estimaciones. Juntas, las fases online y offline optimizan los conteos de abundancia estimados, cuyos resultados son más precisos que kallisto [53].

Finalmente, hoy kallisto y Salmon son consideradas las mejores herramientas a utilizar para

una cuantificación veloz, puesto que no necesita de un proceso de alineamiento previo y la ejecución de ellos sobre una muestra de secuenciación RNA es muy rápida comparada con las otras herramientas [52] [53].

2.1.6 Análisis de expresión diferencial

Una vez realizada la cuantificación (o pseudo-alineamiento) se obtiene una tabla de conteos que tiene las abundancias para cada gen de cada muestra. Esta tabla es utilizada como input para la etapa de análisis de expresión diferencial, donde existen varias herramientas para realizar aquella tarea.

El método más popular es edgeR [54], una herramienta ampliamente utilizada a pesar de ser la más antigua de todas, y que además del análisis de expresión diferencial logra normalizar considerando posibles sesgos que puedan existir en los datos. Otra herramienta es DESeq (o la más actual DESeq2) [55] [56], la cual utiliza su propia forma de normalización y que al igual que edgeR parte de la base de que los datos se distribuyen de una forma binomial negativa. Ambas herramientas, edgeR y DESeq2, son útiles para experimentos en donde las réplicas biológicas utilizadas no superan las cinco muestras por grupo. Otras herramientas como NOISeq [57] o SAMseq [58] utilizan métodos no paramétricos; es decir, no parten de una suposición sobre el comportamiento de los datos y la distribución estadística de ellos se infiere ahí mismo.

Capítulo 3

Propuesta de Solución

Gracias a la gran abundancia de herramientas bioinformáticas hoy existen múltiples maneras de realizar un análisis RNA-Seq. Para este trabajo en específico es necesario un flujo de trabajo que considere un estudio rápido de los datos biológicos, para luego realizar un análisis de expresión diferencial y obtener un output que pueda servir en la visualización de su resultado.

Con lo recogido en el párrafo anterior, también se busca desarrollar una plataforma de visualización que preste información relevante al sector agrícola y que ayude en la toma de decisiones sobre el sujeto en estudio. Para este trabajo, aquel sujeto será la *Vitis vinifera*, con un foco en las etapas de desarrollo de la planta.

A partir de la información recabada en el estado del arte, en este capítulo se presentará el flujo de trabajo RNA-Seq adecuado para el cumplimiento de los objetivos, además de una propuesta de prototipo web para la visualización de datos provenientes del análisis de expresión diferencial de genes.

3.1 Flujo de Trabajo Bioinformático RNA-Seq

3.1.1 Datos a utilizar

Como fue mencionado en la sección 2.1.1, los datos que se utilizarán para este trabajo serán secuenciaciones de la planta *Vitis vinifera* en distintos estados fenológicos, y las cuales serán obtenidas a partir de la base de datos ArrayExpress [22], del Instituto Europeo de Bioinformática. En detalle, los datasets a utilizar son los indicados en la tabla 3.1.

ID	Código ArrayExpress	Número de muestras	Etapas de crecimiento	Single-End o Paired-End
1	E-GEOD-56844	23	4	Single-end
2	E-GEOD-71146	24	4	Single-end
3	E-GEOD-58061	6	3	Paired-end
4	E-MTAB-4220	9	3	Paired-end
5	E-GEOD-63512	6	3	Single-end

Tabla 3.1: Datasets de ArrayExpress a analizar.

3.1.2 Trimming y control de calidad

Se utilizará la herramienta FastQC [25] para realizar la etapa de control de calidad de las muestras. Su elección se debe a la facilidad de uso, su compatibilidad con secuenciaciones de Illumina y la simplicidad que otorga al momento de interpretar los índices de calidad.

Con el fin de eliminar las bases de mala calidad, aquellas lecturas que tienen largo 100 o más y cuyas bases tienen un puntaje de Phred 20 o menor [28] pasarán por el proceso de trimming (poda en español), utilizando la herramienta Trimmomatic [31].

3.1.3 Alineamiento y cuantificación

Para realizar la cuantificación de los genes es necesario que, en un primer lugar, las secuenciaciones obtenidas de los dataset se alineen al transcriptoma de referencia.

El método clásico de alinear las secuenciaciones al transcriptoma es mediante herramientas como TopHat2 o STAR. Éstas son herramientas que realizan el alineamiento de forma exhaustiva, pero requieren mucho tiempo de ejecución o muchos recursos. La ventaja de TopHat2 es la precisión por sobre las otras herramientas disponibles [34], además de requerir pocos recursos. STAR, por otra parte, tiene la ventaja de ser un alineador muy rápido, estando listo en menos de una hora comparado con los otros programas. Ambas herramientas tienen desventajas que las convierten en opciones descartables para este trabajo: en cuanto a TopHat2, si bien los recursos computacionales que requiere se encuentran bajo los 8 GB de memoria RAM, es la ejecución del programa la que toma entre 8 y 18 horas en alinear; por otra parte, STAR es muy rápido para realizar el alineamiento, pero no es tan preciso como TopHat2 u otras herramientas, y en cuanto al poder computacional se requiere un mínimo de 28 GB de RAM.

Dejando atrás el método clásico de alineamiento, como fue mencionado en el capítulo 2 en los últimos años han aparecido herramientas de cuantificación donde no se realizan alineamientos exhaustivos al transcriptoma, y que aún así logran buenos resultados en cuanto al conteo de genes. Las herramientas más relevantes en esta técnica de cuantificación son kallisto y Salmon. Salmon [53] realiza un proceso de alineamiento ligero, seguido de un algoritmo online donde estima los niveles de expresión, y termina con una fase offline que ajusta las estimaciones anteriores. La herramienta es más precisa y mucho más rápida que otros alineadores y cuantificadores, con excepción de kallisto que realiza una tarea similar a una rapidez comparable, alineando los k-mers de las secuencias al grafo de Bruijn del transcriptoma [52]. Aún no hay consenso en cuál es el método óptimo para realizar el alineamiento; sin embargo, al ser uno de los objetivos de este trabajo realizar el proceso de una forma rápida es que se opta por utilizar las herramientas kallisto y Salmon como métodos de cuantificación.

Con el fin de cumplir con el objetivo de generar un flujo de trabajo rápido, en esta memoria se comparará la ejecución del trabajo utilizando ambas herramientas y así encontrar aquella

que es la más veloz.

3.1.4 Análisis de expresión diferencial

Previo a esta etapa, con la herramienta tximport se unifican los resultados de la cuantificación para generar una matriz de conteo, la cual servirá de input para el análisis de expresión diferencial. A continuación se realiza el análisis de expresión diferencial, para el cual existen varias herramientas (sección 2.1.6) de donde se ha escogido edgeR. Este programa es parte de Bioconductor[1] y, como el nombre lo sugiere, está basado en R.

Para muestras biológicas provenientes de distintas fuentes o individuos, la variabilidad entre ellas usualmente se encuentra modelada como una distribución binomial negativa [19]. Lo anterior reduce la búsqueda de herramientas a dos: edgeR y DESeq2. Se prefiere el uso de la primera puesto que tiene una mayor antigüedad y posee una buena documentación, además de apoyo en linea por parte de la comunidad.

A partir del análisis de expresión diferencial se obtendrán dos grupos de genes: upregulated (regulación positiva) y downregulated (regulación negativa). Los genes upregulated serán aquellos que, debido al estímulo externo (enfermedad, etapa fenológica, etc), incrementan en cantidad. En cambio, los genes que disminuyen en cantidad de acuerdo a aquel estímulo son los llamados downregulated.

Estas listas de genes pasarán a la siguiente etapa de anotación funcional, la cual hace referencia a la anotación y análisis estadístico de aquellas listas mediante métodos estadísticos para identificar a qué procesos biológicos, funciones moleculares y componentes celulares aquellos genes se encuentran relacionados. En este trabajo aquella tarea es realizada con PANTHER DB, herramienta en linea que recibe como input la lista de genes, y como output entrega la clasificación correspondiente para cada uno.

En la figura 3.1 se presenta el flujo de trabajo a utilizar para el análisis RNA-Seq de los distintos datos biológicos de la *Vitis vinifera*.

[1] Proyecto bioinformático de código abierto que provee herramientas para el análisis y comprensión de datos genómicos. Sitio web: `http://bioconductor.org/`

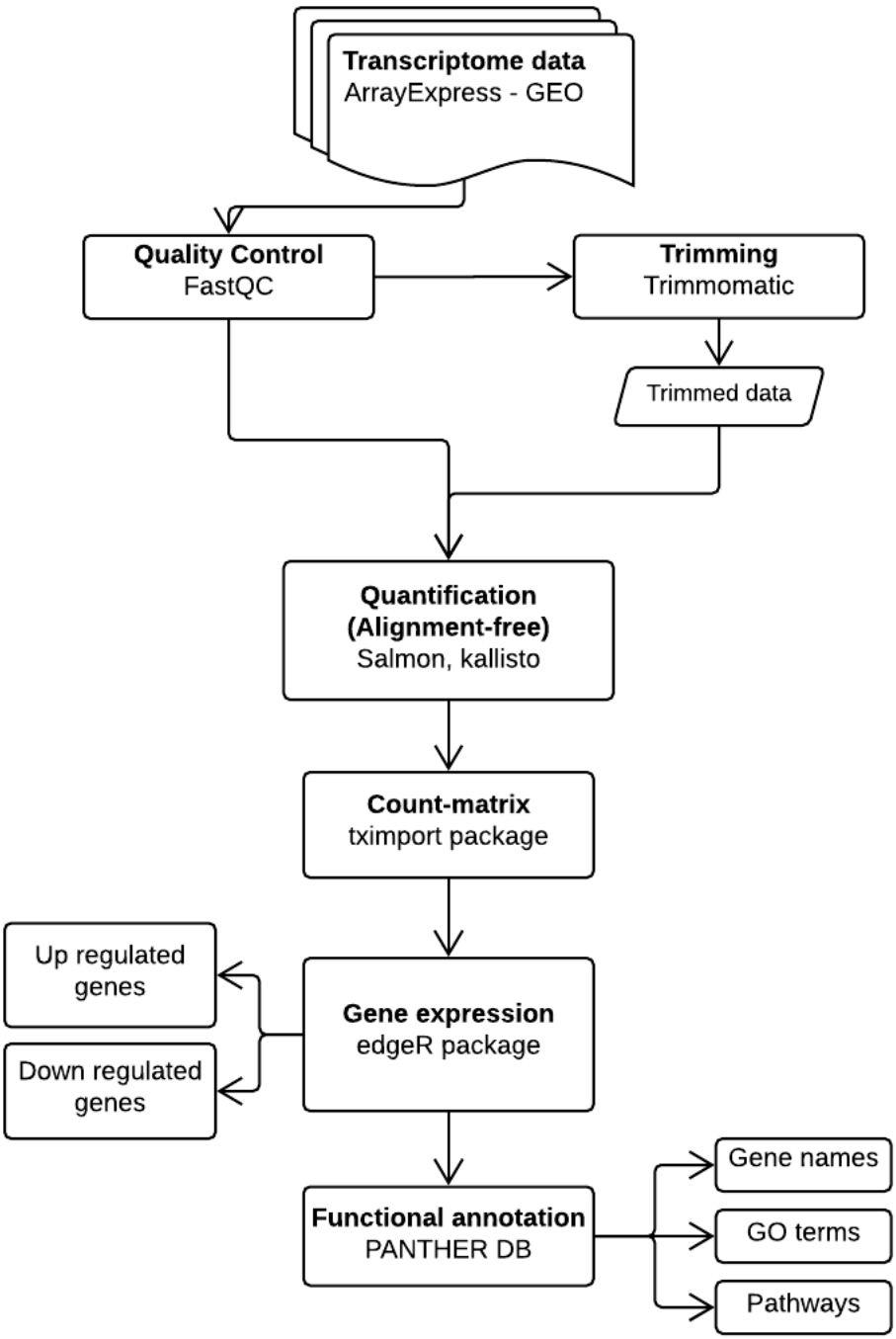


Figura 3.1: Flujo de información para el procesamiento de datos RNA-Seq.

3.2 Visualización de datos

Para hacer sentido de los resultados provenientes del análisis previo, se hace importante mostrarlos de una forma resumida y fácil de entender de forma que puedan ser comprendidos por el agricultor o el usuario final. Es por esta razón que se propone una plataforma web de visualización de los datos biológicos provenientes de diferentes plantaciones, donde el usuario pueda observar el desarrollo de estas a través del año, y además ver cuáles son los genes diferencialmente expresados y a qué procesos están asociados.

Para ello, en este trabajo se desarrolla un primer prototipo de esta plataforma, la cual estará construida utilizando Shiny, una librería de RStudio para construir sitios web interactivos basados en R. Además, se utilizará la librería de código abierto Leaflet para mostrar mediante mapas la variación de las etapas de desarrollo en las plantaciones durante un año.

En la figura 3.2 se muestra un esquema del contenido general del sitio. En la sección 1 se muestran las zonas de cultivo en un mapa, donde cada zona debe estar bien delimitada y coloreada según la etapa de desarrollo correspondiente. La sección 2 servirá para presentar mediante una tabla los genes diferencialmente expresados y sus respectivas funcionalidades. La sección estará separada en dos pestañas que distingan entre genes upregulated y down-regulated. Finalmente, en la sección 3 se encuentran los filtros que permitirán actualizar la información que muestran las secciones 1 y 2. En primer lugar deberá existir un filtro de tiempo que actualiza el desarrollo que han tenido las zonas de cultivo durante un año, y además deben haber filtros para actualizar los genes que se muestran en la sección 2 según etapa de desarrollo.

Figura 3.2: Esquema del contenido de la plataforma Shiny. Sección 1: Consiste en un mapa que muestra las zonas de cultivo de un agricultor, donde además deben estar coloreadas de acuerdo a la etapa de desarrollo en la que se encuentra. Sección 2: Presenta en una tabla los genes diferencialmente expresados en cierta etapa de desarrollo. Sección 3: Filtros que actualizan la información mostrada en el mapa y en las tablas.

Capítulo 4

Resultados

4.1 Análisis RNA-Seq

En una primera instancia se realizó un análisis RNA-Seq siguiendo las etapas mostradas en la figura 3.1. Para gran parte del análisis RNA-Seq se utilizó una máquina virtual alojada en los servidores de la universidad, la cual corre sistema operativo Ubuntu 18.04, con 31 GB de memoria RAM, 500 GB de disco duro y un procesador Intel Xeon E3-12xx de cuatro núcleos. Además fue necesario instalar R para poder trabajar con el framework de bioinformática Bioconductor (`http://bioconductor.org`).

En primer lugar se debe realizar la obtención de los datos biológicos con los que se realizará el trabajo, según fue mencionado en la sección 3.1.1. Aquello se realiza desde el sitio web de ArrayExpress (`www.ebi.ac.uk/arrayexpress/`), en donde se realizó una búsqueda de las secuenciaciones disponibles de *Vitis vinifera*, y luego se hizo un filtro para mantener solo aquellos datasets que describen la etapa fenológica de esos datos. De lo anterior se obtuvieron $89,13$ GB de datos repartidos en cinco datasets, con 68 archivos de muestras biológicas (archivos de extensión .fastq) en total; en la tabla 4.1 se muestran las etapas fenológicas presentes junto con la cantidad de muestras que se obtuvieron para cada una, además de una breve descripción de la etapa. La siguiente tarea será entonces el control de calidad de las secuencias.

Etapa fenológica	Cantidad de muestras	Descripción
EL-3	6	Brote lanudo
EL-5	6	Roseta con puntas de hojas visible
EL-15	8	8 hojas separadas, brote se alarga rápidamente
EL-17	5	12 hojas separadas
EL-27	2	Cuajado de fruta
EL-31	6	Bayas del tamaño de una arveja
EL-35	13	Bayas toman color y se agrandan
EL-36	4	Bayas con valores Brix intermedios
EL-38	15	Bayas maduras listas para cosechar
EL-41	3	Bayas post-cosecha

Tabla 4.1: Etapas fenológicas a utilizar para el análisis RNA-Seq, junto con la cantidad de muestras disponibles y una descripción de la etapa. Las etapas se encuentran nombradas por su numeración en escala Eichhorn-Lorenz.

4.1.1 Control de calidad

Se ejecutó el programa FastQC para realizar el control de calidad, donde se analizaron los 68 archivos .fastq, obteniendo reportes HTML para cada uno de ellos. Cada reporte comienza mostrando los metadatos del archivo, con información como cantidad total de secuencias, largo de una secuencia y cantidad de secuencias marcadas con mala calidad. Además, tal como se explicó en la sección 2.1.2, en el resto del reporte se presentan distintas gráficas donde cada una entrega un significado diferente en relación a la calidad de las secuencias.

La calidad de una base indica la confianza con la que se determinó aquella al momento de la secuenciación, y se expresa en escala Phred. Se determina calculando $\log_{10} P$ donde P es la probabilidad de que la base sea errónea, y luego ese valor se multiplica por -10. Por ejemplo, si existe una probabilidad de 1 en 1000 de que la base sea errónea, entonces el valor de calidad de Phred será $q = -10 * \log_{10}(0{,}001) = 30$. Valores entre 40 y 28 son considerados buena calidad, entre 28 y 20 son de calidad media y cuando cae entre 20 y 0 será de mala calidad.

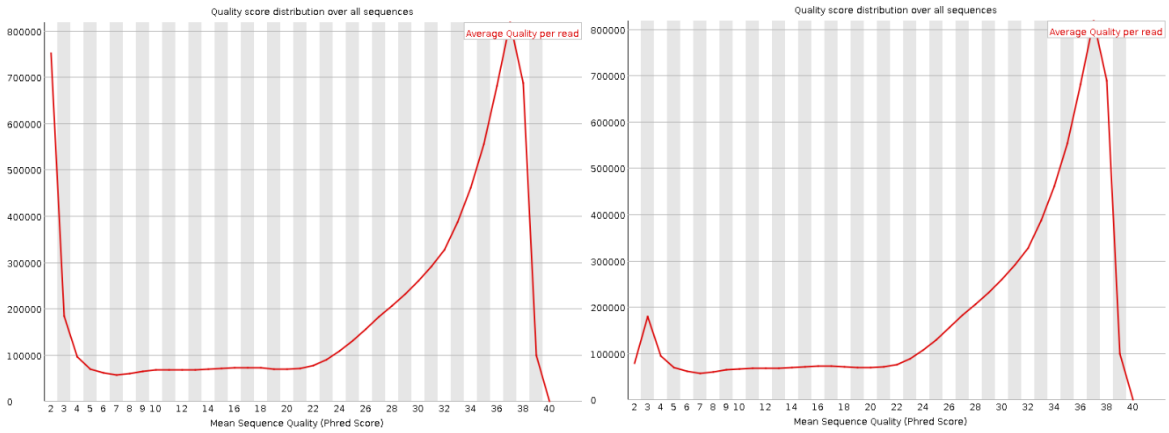
El dataset 1, según la numeración de la tabla 3.1, se evalúa de forma positiva luego del control de calidad, puesto que todas sus bases entregan un puntaje de Phred sobre 30. Además, del cuarto gráfico se puede decir con seguridad que la calidad promedio varía entre 30 y 39 para la gran mayoría de las bases. Es por estas características que el dataset 1 no tendrá la necesidad de someterse al proceso de trimming. Lo mismo ocurre con el dataset 5, donde todos los indicadores de calidad son resultados positivos, por lo que tampoco califica para trimming.

El dataset 4 varía un poco con respecto a los dos anteriores ya que según se puede ver en los reportes, el último cuartil de los box plots baja más allá de Phred 20 en las últimas bases, siendo de peor calidad que los dataset anteriores. Además se puede notar que por ser lecturas paired-end, las lecturas de dirección reversa tienen peor calidad que las lecturas de dirección directa.

Los tres dataset anteriores tienen en común que el largo de las secuenciaciones no supera las 51 bases. Lo anterior no los convierte en un buen sujeto para el proceso de trimming, ya que al ser lecturas tan cortas se corre el riesgo de podar partes que son de buena calidad.

De los datasets restantes, el 3 presenta mala calidad en las últimas cuatro bases de las lecturas, pero siempre el promedio se mantiene en buena calidad sobre el Phred 30. A pesar de ello, como era de esperar las lecturas de dirección reversa empeoran en calidad en las últimas bases. Finalmente el dataset 2 será el peor evaluado de todos. La totalidad de las bases tienen al menos su último cuartil dentro de la zona con peor calidad (bajo Phred 20), y también la mayoría de los box plot estarán bajo calidad 20, incluso llegando al peor puntaje posible. Además, como se puede ver en la figura 4.1a, una gran parte de las secuenciaciones tiene un

promedio de calidad menor a Phred 20, lo que definitivamente es muy bajo como para seguir trabajando con estos datos. Por las razones anteriores, y porque ambos datasets cuentan con un largo de 100 bp, se decide realizar trimming a estos dos grupos de datos.



(a) Antes del trimming cerca de 750.000 secuencias tienen puntaje Phred 0.

(b) Luego de hacer trimming los reads de mala calidad disminuyeron cerca de un 75 %.

Figura 4.1: Distribución de la calidad promedio de las lecturas del dataset 2.

Trimming

Utilizando la herramienta Trimmomatic mediante línea de comandos, los datasets 2 y 3 son sometidos al proceso de trimming con parámetros `LEADING:20` y `MINLEN:50`, donde la ejecución fue cercano a las cuatro horas para cada dataset. Considerando que un dataset tiene 24 muestras y el otro seis, la similitud en los tiempos de ejecución puede parecer extraña, pero la diferencia está en que la máquina toma cerca de cuarenta minutos en podar las lecturas paired-end (dataset 3), y para el dataset 2, que tiene cuatro veces más muestras, demora entre 5 y 20 minutos por cada una de ellas. Esto tiene más sentido si se observa la relación que tiene, para el dataset 2, el tamaño del archivo a podar con el tiempo que demora en ejecutarse Trimmomatic (ver Figura 4.2), donde claramente existe una relación directamente proporcional con el tamaño en GB de la muestra.

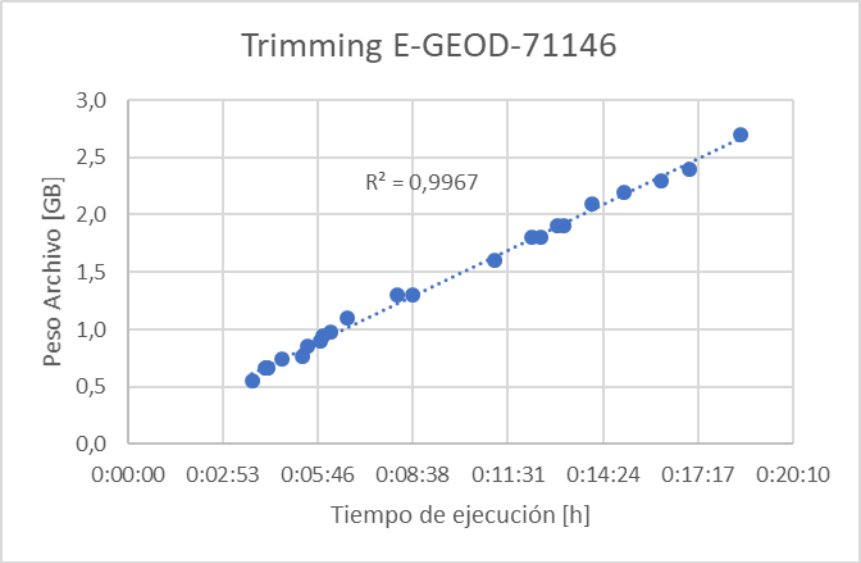


Figura 4.2: Relación entre el tamaño de la muestra biológica y el tiempo que toma Trimmomatic en podar aquel archivo.

Finalmente, como se puede ver en la figura 4.3b, se logró una mejora considerable en la calidad de las lecturas para el dataset 2 (E-GEOD-71146) en comparación con lo presentado en la figura 4.3a, eliminando una buena cantidad de secuencias con mala calidad (ver figura 4.1). En relación al dataset 3, su reporte de calidad luego de hacer trimming no muestra cambios sustanciales en la calidad de éste.

4.1.2 Cuantificación libre de alineamiento

Para la cuantificación se consideró el uso de dos herramientas de última generación: Salmon y kallisto. Antes de realizar la tarea principal, ambos programas necesitan un archivo de 'índice', el que debe ser creado a partir del transcriptoma de referencia, encontrado en el sitio web `http://plants.ensembl.org/`.

Salmon es presentado como un programa muy rápido y a la vez altamente preciso para producir cuantificaciones de datos RNA-seq, esto gracias a la técnica de *quasi-mapping* que utiliza en el proceso. El programa recibirá como input el índice del transcriptoma y las secuenciaciones producto del trimming. Como salida entrega una carpeta por cada muestra

(a) Antes del trimming al menos un 30 % tiene la peor calidad posible.

(b) Después del trimming se logró eliminar una gran cantidad de lecturas de mala calidad.

Figura 4.3: Distribución de los puntajes Phred para el dataset 2.

analizada, donde cada una de ellas contiene un archivo en formato TSV listando un gen por fila, indicando su largo (Length), largo efectivo (EffectiveLength), su abundancia relativa en transcritos por millón (TPM) y el número estimado de lecturas originados a partir del transcriptoma (NumReads). Además la opción `-p` (o `--threads`) permite elegir la cantidad de hilos con los que ejecutar la cuantificación.

Esta herramienta no se diferencia mucho de kallisto en cuanto a su ejecución. Mediante el proceso de pseudo-alineamiento, kallisto permite cuantificar secuenciaciones directamente sin la necesidad de realizar el proceso de alineamiento, razón que lo convierte en una alternativa comparable a Salmon. Al igual que esta última, kallisto recibe como input un archivo 'índice' del transcriptoma y las secuenciaciones a cuantificar, y entrega como output información sobre conteos estimados, TPM y largo efectivo.

Se ejecutaron ambas herramientas mediante la línea de comandos de Ubuntu, indicando que utilizara cuatro hilos (threads). Al comparar los programas, si bien existen algunas diferencias, ambos cuantificadores se ejecutan con buen rendimiento (ver tabla 4.2). Para los datasets 1, 4 y 5 el tiempo de ejecución nunca supera los 20 minutos, y si se compara con las herramientas de alineamiento completo [44] el rendimiento es muy bueno. En cuanto a los datasets 2 y 3, la cuantificación toma bastante más tiempo, llegando incluso a demorar 1 hora 24 minutos (dataset 3 con Salmon).

#	Dataset cuantificado	Programa	Duración	Diferencia
1	E-GEOD-56844	Salmon	0:19:44	0:02:18
		kallisto	0:17:26	
2	E-GEOD-71146	Salmon	0:46:46	0:15:00
		kallisto	0:31:46	
3	E-GEOD-58061	Salmon	1:23:29	0:37:03
		kallisto	0:46:26	
4	E-MTAB-4220	Salmon	0:13:47	0:01:25
		kallisto	0:12:22	
5	E-GEOD-63512	Salmon	0:05:23	0:01:31
		kallisto	0:03:52	

Tabla 4.2: Tiempo de ejecución utilizado en la cuantificación, por dataset y programa utilizado.

Una de las razones por las que los programas demoran más para los datasets 2 y 3 puede explicarse en el largo de sus secuencias; si bien se realizó trimming sobre estos datos, el largo que inicialmente era de 100 bp[1] no disminuyó de manera considerable, por lo que ambos datasets tienen lecturas que llegan a ser el doble de los otros.

4.1.3 Unificación de resultados

El análisis de expresión diferencial requiere como input una matriz de los conteos producidos en el paso anterior (count matrix), y es por esta razón que se hace necesario unificar los resultados de la cuantificación en matrices para su posterior uso en edgeR. Para realizar esto se utiliza la librería de R tximport, que es parte de Bioconductor. Esta librería contiene utilidades para importar y compendiar las abundancias provenientes de herramientas como Salmon o kallisto, para así entregar un resultado que pueda ser utilizado por herramientas de análisis de Bioconductor, como DESeq2 o edgeR.

Para llevar a cabo esta tarea se debió realizar un script en R donde se indica el programa

[1]bp: base pair. Cuando la muestra es single-end se refiere a los nucleótidos con la abreviación nt.

de cuantificación utilizado (si Salmon o kallisto), la ruta de los conteos de cada muestra, y la anotación de genes y transcritos, el que se puede obtener del sitio Ensembl. Una vez que se ejecuta el código se obtiene una matriz compuesta por todos los datasets cuantificados, donde cada columna corresponde a una muestra. Además, se escribió otro script para agrupar aquellas columnas por sus etapas de crecimiento. El resultado de esta tarea es una matriz de conteos, con las columnas ordenadas por etapa fenológica, para que las muestras puedan ser fácilmente agrupadas en la siguiente etapa.

4.1.4 Análisis de expresión génica diferencial

Una de las partes más importantes de este pipeline RNA-Seq se realiza en el análisis de expresión génica, donde el objetivo es encontrar los genes que se encuentran más o menos expresados entre dos condiciones biológicas distintas, observando así los niveles de ARN como la cantidad de lecturas que se superponen en la región del gen en el transcriptoma. En este trabajo aquellas condiciones biológicas serán dos etapas fenológicas consecutivas, lo que significa que se realizarán nueve análisis distintos.

Para realizar esto se escribe un script en R que hace uso de la librería edgeR de Bioconductor. En primer lugar, el script debe obtener los datos a comparar, los cuales provienen de la matriz de conteos generada en el paso anterior. Además es necesario proveer la etapa fenológica correspondiente a cada columna de la matriz[2], de esta forma edgeR puede agrupar las muestras en base a su etapa de crecimiento respectiva. Luego en el script se eliminan aquellas filas donde la gran mayoría de los conteos son cero, quitando así más de seis mil genes y dejando una matriz de 23.657 filas.

En los datos RNA-Seq pueden existir muestras en donde el tamaño de librería (library size) sea mayor al resto, lo que lleva a tener conteos muy mayores en comparación a otras muestras, y que luego se puede traducir a identificar genes erróneamente como downregulated o viceversa. Para prevenir esto se realiza la tarea de normalización mediante la función `calcNormFactors` de edgeR. Una vez hecho esto ya es posible obtener los genes diferencialmente expresados por cada etapa fenológica; para esto se realiza un proceso iterativo

[2]Cada columna de la matriz es una muestra biológica, y cada fila es un gen de la planta

donde se compara una etapa con la consecutiva, luego esta última con la siguiente, y así hasta llegar a la etapa EL-41. Se realizó la selección de genes diferencialmente expresados con la función `decideTestsDEG` y utilizando un logFC [3] de 1,5 como filtro. En la tabla 4.3 se presenta la cantidad de genes obtenidos por cada comparación. Las listas de genes son muy grandes y pueden ser observadas en el sitio web presentado en la sección 4.2 de este trabajo.

	Etapas fenológicas	**Upregulated**	**Downregulated**
1	EL-3 y EL-5	10	7
2	EL-5 y EL-15	454	261
3	EL-15 y EL-17	19	15
4	EL-17 y EL-27	942	997
5	EL-27 y EL-31	1522	1810
6	EL-31 y EL-35	282	1661
7	EL-35 y EL-36	308	1834
8	EL-36 y EL-38	263	193
9	EL-38 y EL-41	1757	3417

Tabla 4.3: Cantidad de genes upregulated y downregulated obtenidos a partir del análisis de expresión génica. Se encuentran clasificados por comparación entre etapa de desarrollo fenológico.

4.1.5 Anotación funcional de genes

Producto del análisis de expresión génica se obtienen los genes diferencialmente expresados para cada transición entre etapas. El objetivo en esta etapa es encontrar las funcionalidades a las que están asociados aquellos genes, con el fin de proveer esta información en la plataforma web que verá el agricultor. Para ello se utiliza la herramienta web PANTHER, que es parte del proyecto Gene Ontology [59], y que recibe como entrada 18 archivos de texto con las listas de genes upregulated y downregulated, y entrega como salida una tabla por cada lista, que describe cada gen de la lista con su funcionalidad en caso de tenerla registrada.

[3] logFC: log_2 fold-change.

Los datos presentados en la tabla son los siguientes:

- Species: Especie a la cual pertenece el gen.
- Gene ID: ID del gen.
- Complete Gene ID: ID del gen en su forma completa.
- Gene Name and Gene Symbol: Nombre del gen.
- PANTHER family / subname: Familia y nombre que recibe el gen dentro de la clasificación PANTHER.
- PANTHER protein class: Clase de proteína proveniente del gen.
- GO-slim Molecular Function: Subset de GO[4] de funciones moleculares del producto génico.
- GO-slim Biological Process: Subset de GO de procesos biológicos en los cuales participa el gen.
- GO-slim Cellular Component: Subset de GO de componentes celulares de los cuales el gen forma parte o es un subcomponente.
- GO database Molecular Function (complete): Funciones moleculares del producto génico.
- GO database Biological Process (complete): Procesos biológicos en los cuales participa el gen.
- GO database Cellular Component (complete): Componentes celulares de los cuales el gen forma parte o es un subcomponente.

[4]Subset de GO, o GO-slim, son formas de representar la ontología de una forma amplia y abreviada.

4.2 Prototipo de plataforma de visualización

La última etapa de este trabajo es la creación de un prototipo para una plataforma de visualización del estado de los cultivos. Más específicamente, a través de un sitio web se busca presentar el desarrollo fenológico de los cultivos, con información proveniente del análisis RNA-Seq de sus plantas. En una instancia previa, mediante la figura 3.2 se mostró una organización general del sitio web; ahora, para tener una idea más precisa de las funcionalidades requeridas, se esbozó un wireframe (ver Figura 4.4) que materializa los elementos que deberá poseer el prototipo hecho en Shiny.

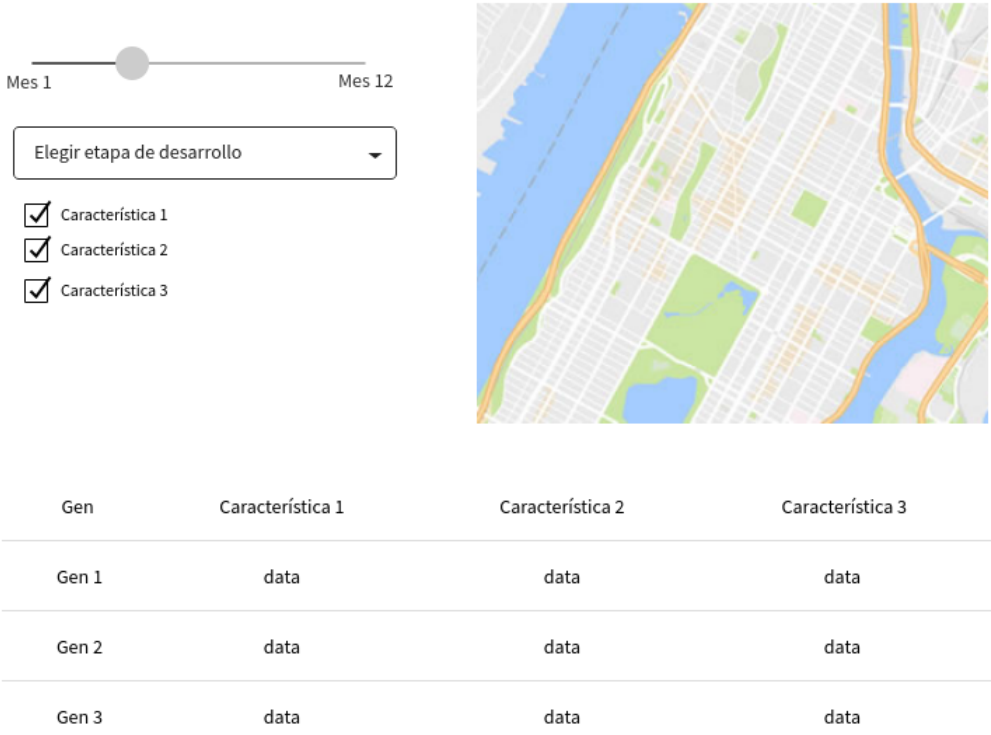


Figura 4.4: Wireframe del prototipo desarrollado.

Shiny es un paquete de R que provee un framework para el desarrollo de aplicaciones web, pudiendo generar plataformas interactivas para el análisis de datos sin tener que recurrir directamente a lenguajes como HTML, JavaScript o CSS. Una aplicación en Shiny está compuesta de dos scripts R para que pueda funcionar, los cuales se comunican entre ellos. Estos componentes son: un script de interfaz de usuario, el que tendrá información sobre el aspecto

del sitio y los elementos que lo conforman; y un script de servidor, el que tendrá instrucciones para los inputs de usuario, outputs y procesamiento de datos mediante código en lenguaje R.

A continuación se describe la composición del prototipo realizado, siguiendo la organización del layout que se muestra en la figura 3.2.

Sección 1:

Toda la sección 1 fue desarrollada gracias a la librería Leaflet para R. Aquí se presenta un mapa que muestra las zonas de cultivo analizadas, las cuales se encuentran debidamente delimitadas, y cada una de ellas lleva un color de relleno según la etapa fenológica que corresponda. Aquella etapa fenológica irá cambiando a medida que el filtro "mes del año" sea deslizado. Además, cada zona lleva una etiqueta asociada a ella, la que aparece solo cuando el cursor está moviéndose por sobre alguna de las zonas. La etiqueta muestra datos relacionados con la zona correspondiente, informando sobre: especie, etapa fenológica actual, cantidad de genes diferencialmente expresados, y cantidad de genes upregulated y downregulated.

Sección 2:

Esta sección se encontrará compuesta por dos pestañas llamadas "Upregulated" y "Downregulated". Cada pestaña mostrará una tabla con los genes diferencialmente expresados (upregulated o downregulated) en una determinada etapa fenológica. Esta etapa debe ser proveída por el usuario en la sección de filtros (sección 3). Además, las columnas de la tabla pueden mostrarse u ocultarse mediante el filtro de checkbox que pueden encontrarse en la sección 3.

Sección 3:

La sección 3 corresponde a la barra lateral izquierda de la interfaz, y que entrega distintos filtros interactivos que, al modificarlos, cambian la información presentada al usuario. En primer lugar destaca la presencia de una barra slider que entrega doce opciones (una por mes

del año), donde por cada ítem se actualiza la información del mapa, cambiando también el color de las zonas según corresponda. Otro filtro disponible es un menú desplegable (drop down menu), del cual cada opción es una etapa fenológica del organismo estudiado, y al seleccionar una opción se actualizan ambas tablas de la sección 2. El último filtro será de checkbox, los cuales habilitan/deshabilitan las columnas mostradas en las tablas de la sección 2.

La figura 4.5 muestra el resultado final del prototipo hecho en Shiny; en la imagen el slider ha sido fijado en "Abril 2018", y la tabla muestra los genes diferencialmente expresados upregulated correspondientes a la etapa EL-17. También el prototipo puede ser visitado en `https://vitis-deg.shinyapps.io/shiny-app/`.

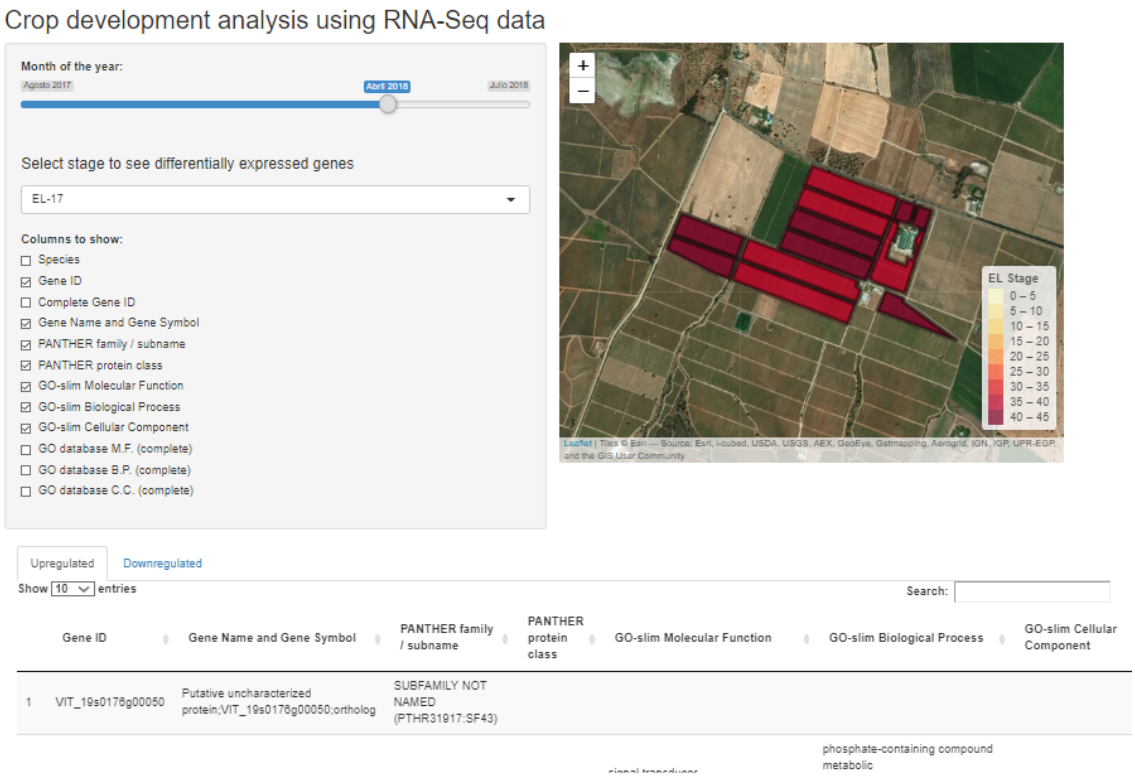


Figura 4.5: Captura de pantalla del prototipo.

Conclusiones

En este documento se planteó un marco de trabajo RNA-Seq para el análisis de datos biológicos provenientes de plantas; para ello se utilizaron secuenciaciones de ARN de *Vitis vinifera*. En primer lugar se hace uso de la herramienta de control de calidad FastQC, y luego algunos de los datasets analizados pasan por una fase de trimming para eliminar lecturas de mala calidad, gracias al programa Trimmomatic. La ejecución de FastQC se realiza sin mayores dificultades, obteniendo los reportes esperados y que entregan la información necesaria para realizar trimming.

La etapa de poda se realizó solo a dos datasets, donde para cada uno la ejecución tomó poco menos de cuatro horas. Llama la atención la demora que tiene esta etapa del análisis ya que la duración es bastante más larga que otras de las tareas, lo que convierte al trimming en el cuello de botella del análisis RNA-Seq. La poda de los datasets 2 y 3 demoran un tiempo similar, pero la gran diferencia es que el último, si bien tiene menos muestras que el otro dataset, sus datos son paired-end con cada archivo pesando hasta cinco veces más que uno proveniente del dataset 2.

En la figura 4.2 se observa una relación directamente proporcional entre el peso de la muestra y el tiempo que demora el programa, lo que explica la demora con la que se ejecuta Trimmomatic en el dataset 3, siendo de más de 38 minutos en promedio para cada muestra. Con esta evidencia y considerando lo que dice la literatura, se pone en duda la etapa de trimming puesto que, a menos que las lecturas sean todas de un largo mayor a 100 bp, la ejecución de Trimmomatic se convertirá en un obstáculo para el análisis gracias a su gran tiempo de ejecución, además de no asegurar resultados de mejor calidad.

En la etapa de cuantificación se utilizaron dos herramientas, kallisto y Salmon, y como muestra la tabla 4.2 el rendimiento de ambos programas es muy similar, a pesar de las diferencias para los datasets 2 y 3. Además del tiempo de ejecución, la literatura indica que Salmon es más preciso al detectar expresión génica en comparación con kallisto y otras herramientas [53]. Si bien kallisto tiene mejores tiempos de ejecución que Salmon por unos minutos, esa diferencia no hace mucha distinción si se compara con los otros métodos que consideran un alineamiento completo.

Posterior a la fase de cuantificación, con la herramienta tximport se unificaron los resultados de los conteos en una matriz. Se obtuvieron matrices con las cuantificaciones de kallisto y Salmon, y a partir de ellas se llevó a cabo el análisis de expresión génica con edgeR. Una vez obtenidas las listas de genes diferencialmente expresados, se realizó una validación de tal forma de verificar la cantidad de genes en común entre ambos métodos de cuantificación. De esto se encontró que la cantidad de genes en común para kallisto y Salmon es siempre mayor que el número de genes presentes solo en uno de ellos.

Por conclusión, la herramienta de cuantificación preferida será Salmon gracias a su rapidez y a la precisión que otorga en comparación con otras herramientas existentes. No obstante, se recomienda considerar kallisto como alternativa puesto que, junto con Salmon, ambos son programas en constante mejora y desarrollo.

Finalmente se desarrolló el prototipo de una plataforma web que toma los datos provenientes del análisis RNA-Seq y los presenta en una interfaz fabricada en Shiny, donde además el bioinformático o el agricultor puede observar un mapa que tiene sus zonas de plantación delimitadas y coloreadas según el estado fenológico, gracias a la librería Leaflet para R. El prototipo puede ser visitado en `https://vitis-deg.shinyapps.io/shiny-app/` y el código fuente se encuentra disponible en el repositorio de GitHub `https://github.com/claudiogalaz/vitis-deg/`.

Por último, como trabajo a futuro se propone considerar la omisión de la etapa de trimming para próximos análisis RNA-Seq, además del desarrollo de una plataforma web que, junto con los datos del análisis RNA-Seq, entregue información resumida sobre los procesos biológicos para cada etapa fenológica. Además, sería útil considerar un análisis para una

lista predeterminada de genes, de forma de evaluar cómo cambia la expresión diferencial de éstos en las distintas etapas fenológicas. También se propone la secuenciación y utilización de datos de *Vitis vinifera* provenientes de los campos de agricultores chilenos, además de la creación de un modelo clasificador que, a partir de los datos de expresión génica de una planta, indique en qué etapa de crecimiento se encuentra.

Bibliografía

[1] Z. N. Oltvai and A.-L. Barabási, "Life's complexity pyramid," *Science*, vol. 298, no. 5594, pp. 763–764, 2002.

[2] National Human Genome Research Institute (NHGRI), "Transcription," 2018. https://www.genome.gov/glossary/index.cfm?p=viewimage&id=197, Acceso: 03/10/2018.

[3] K. A. Wetterstrand, "DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP)," 2013. www.genome.gov/sequencingcostsdata, Acceso: 22/05/2018.

[4] "Balance de la OIV sobre la situación vitivinícola mundial," 2017. http://www.oiv.int/public/medias/5347/press-release-2017-bilan-es.pdf, Acceso: 10/11/2017.

[5] S. B. Piazza, "Antecedentes de los mercados del vino y de la uva vinífera," 2017. http://www.odepa.cl/wp-content/uploads/2017/07/mercadoVino2017.pdf, Acceso: 11/11/2017.

[6] "Aspectos de la coyuntura mundial," 2017. http://www.oiv.int/public/medias/5288/oiv-noteconjmars2017-es.pdf, Acceso: 11/11/2017.

[7] M. Baggiolini, "Les stades repères dans le développement de la vigne et leur utilisation pratique, Station Féd," *Essais Agric., Lausanne*, 1952.

[8] K. Eichhorn and D. Lorenz, "Phänologische Entwicklungsstadien der rebe," *Nachrichtenblatt des Deutschen Pflanzenschutzdienstes*, 1977.

[9] M. Baillod and M. Baggiolini, "Les stades repères de la vigne," *Rev. Suisse Vitic. Arboric. Hortic*, vol. 25, no. 1, pp. 10–12, 1993.

[10] D. Lorenz, K. Eichhorn, H. Bleiholder, R. Klose, U. Meier, and E. Weber, "Growth Stages of the Grapevine: Phenological growth stages of the grapevine (Vitis vinifera L. ssp. vinifera)—Codes and descriptions according to the extended BBCH scale," *Australian Journal of Grape and Wine Research*, vol. 1, no. 2, pp. 100–103, 1995.

[11] B. Coombe, "Growth stages of the grapevine: adoption of a system for identifying grapevine growth stages," *Australian Journal of Grape and Wine Research*, vol. 1, no. 2, pp. 104–110, 1995.

[12] J. W. Dale, M. Von Schantz, and N. Plant, *From Genes to Genomes: Concepts and Applications of DNA Technology*. John Wiley & Sons, 2012.

[13] A. J. Griffiths, *An Introduction to Genetic Analysis*. Macmillan, 2005.

[14] National Human Genome Research Institute (NHGRI), "A Brief Guide to Genomics," 2015. `https://www.genome.gov/18016863/a-brief-guide-to-genomics/`, Acceso: 06/11/2017.

[15] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences*, vol. 74, no. 12, pp. 5463–5467, 1977.

[16] W. S. Klug, M. R. Cummings, and C. A. Spencer, *Concepts of Genetics*. Pearson Education, 2015.

[17] J. Ramsden, *Bioinformatics: an introduction*, vol. 21. Springer, 2015.

[18] C. Soneson and M. Delorenzi, "A comparison of methods for differential expression analysis of RNA-seq data," *BMC Bioinformatics*, vol. 14, no. 1, p. 91, 2013.

[19] E. Korpelainen, J. Tuimala, P. Somervuo, M. Huss, and G. Wong, *RNA-seq data analysis: a practical approach*. Chapman and Hall/CRC, 2014.

[20] R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker, "Highly integrated single-base resolution maps of the epigenome in Arabidopsis," *Cell*, vol. 133, no. 3, pp. 523–536, 2008.

[21] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, *et al.*, "A survey of best practices for RNA-seq data analysis," *Genome Biology*, vol. 17, no. 1, p. 13, 2016.

[22] N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, *et al.*, "ArrayExpress update—simplifying data submissions," *Nucleic Acids Research*, vol. 43, no. D1, pp. D1113–D1116, 2014.

[23] K. Froussios, N. J. Schurch, K. Mackinnon, M. Gierlinski, C. Duc, G. G. Simpson, and G. J. Barton, "How well do RNA-Seq differential gene expression tools perform in higher eukaryotes?," *bioRxiv*, p. 090753, 2016.

[24] T. O'Grady, M. Baddoo, and E. K. Flemington, "Analysis of EBV Transcription Using High-Throughput RNA Sequencing," *Epstein Barr Virus: Methods and Protocols*, pp. 105–121, 2017.

[25] S. Andrews *et al.*, "FastQC: a quality control tool for high throughput sequence data," 2010.

[26] X. Li, A. Nair, S. Wang, and L. Wang, "Quality control of RNA-seq experiments," *RNA Bioinformatics*, pp. 137–146, 2015.

[27] M. Dai, R. C. Thompson, C. Maher, R. Contreras-Galindo, M. H. Kaplan, D. M. Markovitz, G. Omenn, and F. Meng, "NGSQC: cross-platform quality analysis pipeline for deep sequencing data," *BMC genomics*, vol. 11, no. 4, p. S7, 2010.

[28] C. R. Williams, A. Baccarella, J. Z. Parrish, and C. C. Kim, "Trimming of sequence reads alters RNA-Seq gene expression estimates," *BMC Bioinformatics*, vol. 17, no. 1, p. 103, 2016.

[29] M. P. Cox, D. A. Peterson, and P. J. Biggs, "SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data," *BMC bioinformatics*, vol. 11, no. 1, p. 485, 2010.

[30] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet. journal*, vol. 17, no. 1, pp. pp–10, 2011.

[31] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

[32] C. Del Fabbro, S. Scalabrin, M. Morgante, and F. M. Giorgi, "An extensive evaluation of read trimming effects on Illumina NGS data analysis," *PloS one*, vol. 8, no. 12, p. e85024, 2013.

[33] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[34] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biology*, vol. 14, no. 4, p. R36, 2013.

[35] G. Baruzzo, K. E. Hayer, E. J. Kim, B. Di Camillo, G. A. FitzGerald, and G. R. Grant, "Simulation-based comprehensive benchmarking of RNA-seq aligners," *Nature Methods*, vol. 14, no. 2, pp. 135–139, 2017.

[36] I. Medina, J. Tárraga, H. Martínez, S. Barrachina, M. I. Castillo, J. Paschall, J. Salavert-Torres, I. Blanquer-Espert, V. Hernández-García, E. S. Quintana-Ortí, and J. Dopazo, "Highly sensitive and ultrafast read mapping for RNA-seq analysis," *DNA Research*, vol. 23, no. 2, pp. 93–100, 2016.

[37] D. Kim, B. Langmead, and S. L. Salzberg, "HISAT: a fast spliced aligner with low memory requirements," *Nature Methods*, vol. 12, no. 4, pp. 357–360, 2015.

[38] T. D. Wu and S. Nacu, "Fast and SNP-tolerant detection of complex variants and splicing in short reads," *Bioinformatics*, vol. 26, no. 7, pp. 873–881, 2010.

[39] G. Jean, A. Kahles, V. T. Sreedharan, F. D. Bona, and G. Rätsch, "RNA-Seq Read Alignments with PALMapper," *Current protocols in bioinformatics*, pp. 11–6, 2010.

[40] K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, *et al.*, "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery," *Nucleic Acids Research*, p. gkq622, 2010.

[41] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[42] A. Dobin and T. R. Gingeras, "Optimizing RNA-Seq Mapping with STAR," *Data Mining Techniques for the Life Sciences*, pp. 245–262, 2016.

[43] S. Marco-Sola, M. Sammeth, R. Guigó, and P. Ribeca, "The GEM mapper: fast, accurate and versatile alignment by filtration," *Nature Methods*, vol. 9, no. 12, pp. 1185–1188, 2012.

[44] G. Baruzzo, K. E. Hayer, E. J. Kim, B. Di Camillo, G. A. FitzGerald, and G. R. Grant, "Simulation-based comprehensive benchmarking of RNA-seq aligners," *Nature Methods*, vol. 14, no. 2, pp. 135–139, 2017.

[45] S. Anders, P. T. Pyl, and W. Huber, "HTSeq–a Python framework to work with high-throughput sequencing data," *Bioinformatics*, p. btu638, 2014.

[46] Y. Liao, G. K. Smyth, and W. Shi, "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, vol. 30, no. 7, pp. 923–930, 2014.

[47] M. Teng, M. I. Love, C. A. Davis, S. Djebali, A. Dobin, B. R. Graveley, S. Li, C. E. Mason, S. Olson, D. Pervouchine, *et al.*, "A benchmark for RNA-seq quantification pipelines," *Genome Biology*, vol. 17, no. 1, p. 74, 2016.

[48] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

[49] A. Roberts and L. Pachter, "Streaming fragment assignment for real-time analysis of sequencing experiments," *Nature Methods*, vol. 10, no. 1, pp. 71–73, 2013.

[50] S. B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E. T. Dermitzakis, "Transcriptome genetics using second generation sequencing in a Caucasian population," *Nature*, vol. 464, no. 7289, pp. 773–777, 2010.

[51] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, no. 1, p. 323, 2011.

[52] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nature Biotechnology*, vol. 34, no. 5, pp. 525–527, 2016.

[53] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature Methods*, 2017.

[54] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[55] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, p. R106, 2010.

[56] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, no. 12, p. 550, 2014.

[57] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa, "Differential expression in RNA-seq: a matter of depth," *Genome Research*, vol. 21, no. 12, pp. 2213–2223, 2011.

[58] J. Li and R. Tibshirani, "Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data," *Statistical Methods in Medical Research*, vol. 22, no. 5, pp. 519–536, 2013.

[59] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, p. 25, 2000.