Desarrollo de flujo de trabajo *(pipeline)* para análisis de mutantes sólidos de *Solanum tuberosum Gr. Phureja*

Oscar Alexis Quintero López

Universidad Distrital Francisco José de Caldas
Facultad de Ciencias y Educación
Licenciatura en Biología
Bogotá D.C.
2022

Desarrollo de flujo de trabajo *(pipeline)* para análisis de mutantes sólidos de *Solanum tuberosum Gr. Phureja*

Oscar Alexis Quintero López

Trabajo de Grado para optar al título de
Licenciado en Biología

Director
Luis Francisco Becerra Galindo
docente investigador grupo BIOMOLC

Universidad Distrital Francisco José de Caldas
Facultad de Ciencias y Educación
Licenciatura en Biología

Bogotá D.C.
2022

La Universidad Distrital Francisco José de Caldas

No se hace responsable de las ideas, ni del contenido del presente trabajo, aquí expuesto por sus autores, según el artículo 117 Acuerdo 029 que el Consejo Superior de la Universidad Distrital Francisco José de Caldas expidió en Junio de 1988.

Resumen

La bioinformática aplicada al mejoramiento de cultivos viene en auge desde el nacimiento de la misma, pues las ventajas que ofrece la digitalización de la información biológica son en muchas ocasiones, diferenciadores vitales a la hora de realización de proyectos, puesto que la posibilidad de los análisis y simulación in silico, con la posibilidad de escalarse con capacidad de cómputo, transforma la forma en que se está realizando la agrigenómica en los últimos años, el estudio de mutantes como principal herramienta para este fin.

El mejoramiento de cultivos, y actualmente la agrigenómica son campos del conocimiento que están empezando a contribuir conjuntamente puesto que las nuevas tecnologías desarrolladas en la biología molecular han permitido el análisis íntegro de variedad de organismos, por supuesto esto depende las características propias de cada especie, en el caso de las plantas al tener genomas muy extensos, la dificultad en el ensamble aumenta exponencialmente, pues los métodos usados principalmente fueron diseñados para genomas pequeños o bacterianos, con el desarrollo de nuevas técnicas y tecnologías el secuenciamiento de genomas de plantas han venido en auge, y principalmente los países más grandes con fuertes políticas públicas en la seguridad alimentaria como lo pueden ser China, Japón, Brasil, Rusia, entre otros han aplicado el conocimiento de estas técnicas moleculares e informáticas para el incremento de rendimientos de sus cultivos.

Por otro lado la papa, es un tubérculo cultivado por sus capacidad de producir grandes rendimientos por hectárea, donde los estudios moleculares no se hacen esperar, en el caso de Colombia, al poseer una gran diversidad de papas, todo esfuerzo es necesario, las variedades nativas son de especial interés, por esta razón se busca incorporar estas variedades no solamente en el ámbito comercial, si no también en el panorama académico e investigativo, razón de ser de este proyecto.

El estudio de los principales genes involucrados en la dormancia del tubérculo en las variedades de papas Colombianas nos permitirá orientar las futuras investigaciones en el campo de agrigenómica, de igual manera este trabajo de grado espera obtener en sus re-

sultados un pipeline, que facilite el manejo de información proveniente de las variedades (mutantes sólidos) de *Solanum tuberosum Gr. Phureja* producto de las investigaciones anteriores llevadas a cabo en el grupo de investigación BIOMOLC, de la Universidad Distrital Francisco José de Caldas, con el fin de fortalecer la rama de agrigenómica dentro del grupo y discernir sobre las herramientas existentes, para realizar un estudio a fondo, seguro y práctico de los datos obtenidos de manera experimental.

Índice

Índice de figuras

Índice de tablas

Glosario de Términos

- ADN (ácido desoxirribonucleico): Molécula de doble cadena formada por un conjunto lineal de nucleótidos (véase la figura 2). El ADN contiene el código genético de un organismo en la disposición de las bases. La doble cadena del ADN es el resultado de los enlaces de hidrógeno que se forman entre las bases cuando se asocian dos cadenas de polinucleótidos, idénticas, pero que van en direcciones opuestas.

- ADNc (ADN complementario): Una pieza artificial de ADN que se sintetiza a partir de una plantilla de ARNm (ARN mensajero) y se crea utilizando la transcriptasa inversa. La forma monocatenaria del ADNc se utiliza con frecuencia como sonda en la preparación de un mapa físico de un genoma. Se prefiere el ADNc para el análisis de la secuencia porque los intrones que se encuentran en el ADN se eliminan en la traducción del ADN —->ARNm —->ADNc.

- ADN polimerasa: enzima que ensambla el ADN en una doble hélice añadiendo bases complementarias a una sola cadena de ADN. Los enlaces se forman añadiendo nucleótidos en el grupo hidroxilo 5' al grupo fosfato situado en el hidroxilo.

- Alineamiento múltiple: Conjunto de biosecuencias dispuestas en una tabla de forma que cada fila de la misma está formada por una secuencia rellena de huecos. Las columnas de la tabla destacan la similitud (o conservación de residuos) entre las posiciones de cada biosecuencia. Una alineación múltiple óptima es la que tiene el mayor grado de similitud, o el menor coste.

- Amplificación: El proceso de hacer repetidamente copias del mismo trozo de ADN.

- Anotación: Campos de texto con información sobre una biosecuencia que se añaden a las bases de datos de secuencias. La anotación (la elucidación y descripción de las características biológicamente relevantes de la secuencia).

- ARNm (ARN mensajero): ARN que se utiliza como molde para la síntesis de proteínas. El primer codón de una secuencia de ARN mensajero es casi siempre AUG.

- BAM: Binary Alignment Map - Mapa de alineamiento binario.
- Base: Una de las cinco moléculas que se ensamblan, junto con una ribosa y un fosfato, para formar nucleótidos (Figura 1). La adenina (A), la guanina (G), la citosina (C) y la timina (T) se encuentran en el ADN, mientras que el ARN está formado por adenina (A), guanina (G), citosina (C) y uracilo (U).
- BP - Par de bases : Las bases complementarias en hebras opuestas del ADN que se mantienen unidas por enlaces de hidrógeno. La estructura atómica de estas bases preselecciona el emparejamiento de la adenina con la timina y el de la guanina con la citosina (o el uracilo en el ARN).
- BLAST: Basic Local Alignment Search Tool - Herramienta básica de búsqueda de alineación local
- Contenido GC: La medida de la abundancia de nucleótidos G y C en relación con los nucleótidos A y T dentro de las secuencias de ADN
- Contigs: Una serie de vectores de clonación que están ordenados de tal manera que cada secuencia se solapa con la de sus vecinos. El resultado es que el ensamblaje de la serie proporciona una parte contigua de un genoma.
- DEA: Diferencial Expression Analysis - Análisis de Expresión Diferencial
- Diploide: Célula que contiene dos conjuntos de cromosomas.
- EBI: El Instituto Europeo de Bioinformática (`http://www.ebi.ac.uk/`) forma parte del EMBL.
- EMBL: El Laboratorio Europeo de Biología Molecular (`http://www.embl-heidelberg.de/`) que se encuentra en Heidelberg, Alemania.
- Ensamblaje: El proceso de colocar los fragmentos de ADN que han sido secuenciados en su posición correcta dentro del cromosoma.
- EST (Expressed Sequence Tag): Secuencia parcial de un clon de ADNc que puede utilizarse para identificar sitios en un gen.

– FASTA: Formato basado en texto para la representación de las secuencias de nucleótidos.

– FASTQ: Formato basado en FASTA agregando información sobre la calidad de la secuenciación de cada nucleótido.

– GenBank: La base de datos de secuencias genéticas de los NIH. Una colección anotada de todas las secuencias de ADN disponibles públicamente que se encuentra en `http://www.ncbi.nlm.nih.gov/`. GenBank forma parte de la Colaboración Internacional de Bases de Datos de Secuencias de Nucleótidos, que está compuesta por el Banco de Datos de ADN de Japón (DDBJ), el Laboratorio Europeo de Biología Molecular (EMBL) y el GenBank del NCBI. Estas tres organizaciones intercambian datos a diario.

– Kilobase (kb): Longitud del ADN equivalente a 1.000 nucleótidos.

– Mapa de consenso: La ubicación de todas las secuencias de consenso en una serie de proteínas o polinucleótidos alineados de forma múltiple.

– Mapas de contig: La representación de la estructura de las regiones contiguas del genoma (contigs) especificando las relaciones de solapamiento entre un conjunto de clones.

– Mapeo: El proceso de determinar las posiciones de los genes y las distancias entre ellos en un cromosoma.

– Marco de lectura (también marco de lectura abierto): El tramo de secuencia de tripletes del ADN que codifica una proteína. El marco de lectura está designado por el codón de iniciación o de arranque y está terminado por un codón de parada.

– Método Shotgun: Método que utiliza enzimas para cortar el ADN en cientos (o miles) de bits aleatorios que luego se reensamblan por ordenador para que se parezca al genoma original. El método shotgun del Proyecto Genoma Humano se aplica a fragmentos de ADN clonados que ya han sido mapeados, de modo que se sabe exactamente dónde están ubicados en el genoma, lo que facilita el ensamblaje y es mucho

menos propenso a errores.

- NCBI: El Centro Nacional de Información Biotecnológica (`http://www.ncgi.nlm.nih.gov/`), una división de los Institutos Nacionales de Salud (NIH), es la sede de los servidores BLAST y Entrez.
- NCGR: Centro Nacional de Recursos Genómicos (`http://www.ncgr.org/`).
- SAM: Sequence Alignment Map - Sequence Alignment Map.
- Scaffold - Andamio: Una serie de contiguos que están en el orden correcto, pero que no están conectados en una longitud continua.
- Secuenciación: Determinación del orden de los nucleótidos en un gen o del orden de los aminoácidos en una proteína.
- Secuencia conservada: Una secuencia dentro del ADN o la proteína que es consistente en todas las especies o que ha permanecido sin cambios dentro de la especie durante su período evolutivo.
- Secuencia de consenso: El aminoácido o el nucleótido más frecuente en cada posición de una serie alineada de proteínas o polinucleótidos.
- SNP - Polimorfismo de un solo nucleótido: El tipo más común de variación en la secuencia del ADN. Un SNP es un cambio en un solo par de bases en una posición concreta de la cadena de ADN.
- SRA: Sequence Read Archive - Archivo de lectura de secuencia
- Splicing - Empalme: Proceso de corte, escisión y recombinación de un ARN o un ADN. En el ARN, el splicing se utiliza para eliminar los intrones de la secuencia codificante.
- Primer - Cebador: Secuencia corta de nucleótidos (normalmente ocho) que sirve para cebar el proceso de la ADN polimerasa durante la división celular. Los cebadores son producidos por la enzima primasa. Los cebadores también se pueden personalizar para .ªislar"secciones específicas de ADN para su replicación mediante la PCR.

1. Descripción del problema

La papa, *Solanum tuberosum* es un cultivo estratégico para la seguridad alimentaria, por esta razón es el tercer cultivo más importante luego del arroz y el trigo (*FAO Statistics*, 2021), aunque la siembra de este tubérculo, inició en América, luego se distribuyó en Europa y los países orientales, la tecnificación de estos cultivos se viene dando principalmente por estos últimos, dado que el rendimiento de estos cultivos en todo el mundo está muy por debajo de su potencial fisiológico de 120 toneladas/ha (Massa *et al.*, 2011), diferentes organizaciones alrededor del mundo como lo son la FAO, han producido informes como (*The State of Food Security and Nutrition in the World 2020*, 2020), evidenciando que la seguridad alimentaria es un problema de la humanidad, pues el acceso a dietas saludables y económicas, es algo que personas en muchos países no se pueden permitir, por esta razón los esfuerzos en el mejoramiento y tecnificación de cultivos, esto porque *Solanum tuberosum* es un cultivo que en muy corto tiempo acumula grandes cantidades de almidón, haciéndolo un alimento nutritivo y que posee una diversidad de preparaciones que podemos evidenciar en las diferentes culturas alrededor del mundo.

Al ser un cultivo estratégico es importante conservar la basta diversidad de semillas que existen de este tubérculo, y la calidad de las mismas, para este fin los diferentes países líderes en la producción de papa, han optado por el desarrollado de bancos de germoplasma y la inversión en investigación, Colombia posee sus propios bancos de germoplasma, entre ellos encontramos el banco de papa y tubérculos andinos, bajo la dirección del grupo de Investigación de Papa de la Universidad Nacional de Colombia, sede Bogotá (Valencia *et al.*, 2010); con una división sobre papa criolla *Solanum tuberosum Gr. Phureja* , pues esta cuenta con una producción total anual de entre 112 787,3 y 240 122,8 toneladas, y un promedio de rendimiento para el periodo considerado de 14,7 t/ha, donde en Colombia los departamentos con mayor área sembrada de papa criolla para el periodo 2014-2018 son, en su orden: Cundinamarca, Boyacá, Nariño y Antioquia. (Ñústez-López y Rodríguez-Molano, 2020)

A través de las entidades correspondientes y universidades se desarrollan proyectos de

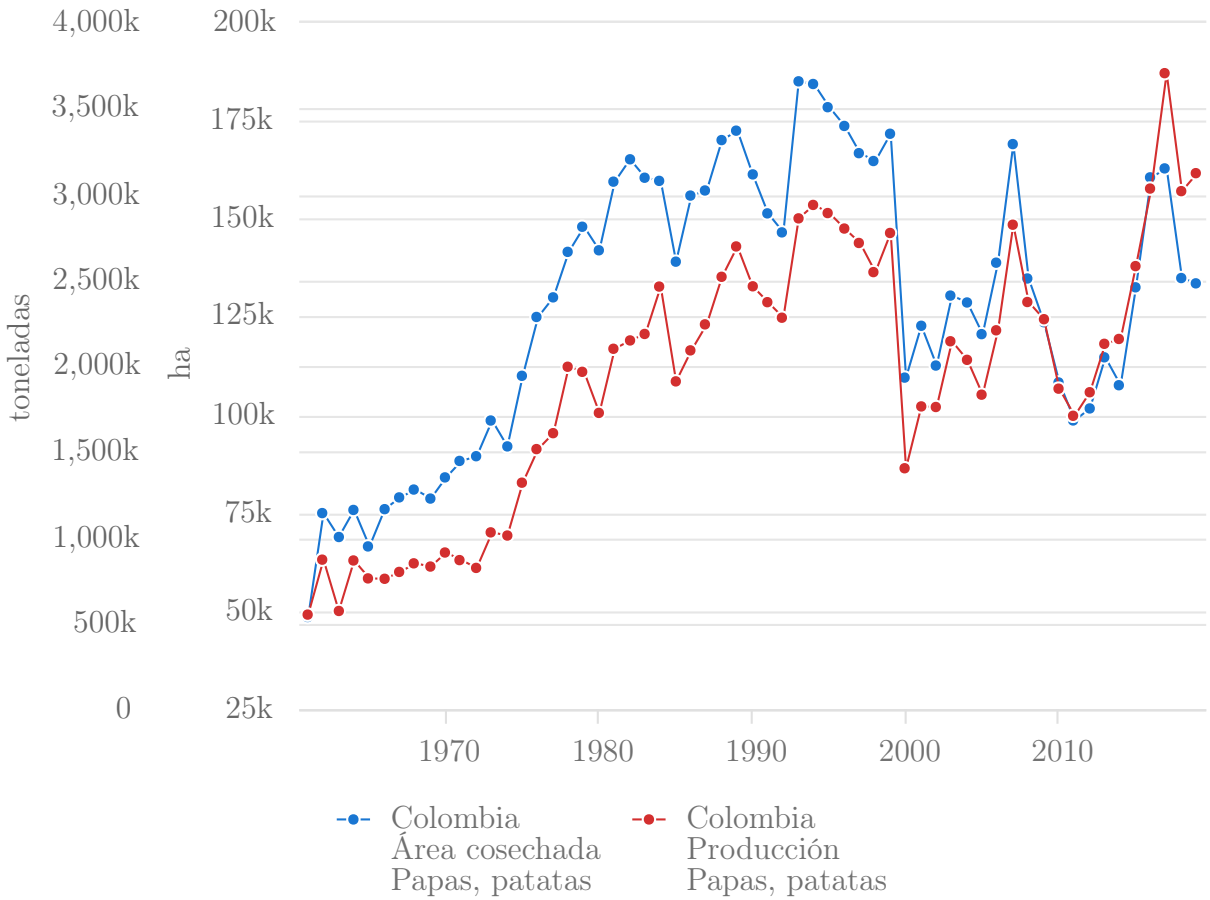

Figura 1: Área Cultivada vs Producción de papa en Colombia (*FAO Statistics*, 2021)

investigación que buscan la tecnificación y el mejoramiento de los cultivos, la producción de papa en el país (Figura 1), ha venido en aumento pero aún muy lejos de los rendimientos por hectárea de los países con cultivos en su mayoría tecnificados (*FAO Statistics*, 2021), por esta razón se busca el mejoramiento de los cultivos por hectárea, el cumplimiento de esta meta se aborda desde diferentes campos del conocimiento, como lo pueden ser la bioquímica, la fisiología vegetal o la misma biología molecular.

El fitomejoramiento, está entrando en un crecimiento exponencial gracias a la digitalización de la información biológica, lo que ha permitido abordar el mejoramiento de las semillas desde una perspectiva molecular e in silico, haciendo mucho más eficiente y acelerado el avance en la obtención de caracteres deseados, países en vía de desarrollo se ven limitados por los costos que implican la digitalización de esta data, en la actualidad las

técnicas moleculares permiten el estudio a profundidad de estos sistemas biológicos, para ello es necesario el establecimiento de protocolos tanto moleculares como bioinformáticos, que permitan la diferenciación de los varietales conseguidos con la inducción de mutaciones con agentes físicos o químicos, trabajo que se ha venido realizando en el grupo de investigación BIOMOLC de la Universidad Distrital Francisco José de Caldas.

Lo que nos deja el siguiente cuestionamiento:

¿Cuál es la manera más efectiva de analizar y organizar el flujo de trabajo para la información obtenida del transcriptoma de los mutantes sólidos logrados de Solanum Tuberosum Grupo Phureja en el grupo de investigación BIOMOLC de la Universidad Distrital Francisco José de Caldas, con la metodología de inducción de mutaciones a través de la irradiación con ^{60}Co (Cobalto 60)?

2. Estado del Arte

Los avances en el mejoramiento molecular de la papa se han visto limitados por su complejo sistema biológico, que incluye propagación vegetativa, auto-tetraploide y heterocigosidad extrema (Massa *et al.*, 2011), el estudio de los transcritos de *Solanum tuberosum* y sus variedades, han venido en crecimiento desde los 2000, tal y como nos dice (Massa *et al.*, 2011), la disponibilidad del genoma y transcriptoma de la papa, ademas de complementar con la estructura de genes correspondiente, la ubicación y la anotación funcional son recursos poderosos para comprender esta planta compleja y avanzar en los esfuerzos de mejoramiento molecular.

Estos esfuerzos se ven reflejados en trabajos como (Petek *et al.*, 2020), *Cultivar-specific transcriptome and pan-transcriptome reconstruction of tetraploid potato*, Reconstrucción de transcriptomas y pan-transcriptomas específicos de cultivares de papa tetraploide, donde se profundiza en los conocimientos sobre los SNP's y INDELs, pequeñas inserciones o deleciones, pues buscan la re-construcción de un panorama transcriptómico de *Solanum tuberosum*, teniendo en cuenta también que entre las variedades seleccionadas para el desarrollo de esta investigación tenemos plantas diploides y plantas tetraploides, estos son indicios de la pobre información que tenemos frente a la basta diversidad en las familias de genes que presentan las papas.

Razón por la cual encontramos otros trabajos como (Zhang *et al.*, 2013), *Identification and Characterization of miRNA Transcriptome in Potato by High-Throughput Sequencing*, donde usando métodos alternativos de secuenciación verificamos la información existente y caracterizamos los miARN que son de vital importancia para los procesos de expresión, es decir estos miARN de unos 21 o 25 bases regulan la expresión de los genes utilizando para ello la ruta de ribo-interferencia (Pillai, 2005), es decir se confrontan los llamados *short reads* contra los *long reads*, siendo ambos métodos validos, que al sumarse multiplican la exactitud y veracidad de los genomas y transcriptomas secuenciados, lo que permite estudiar el comportamiento de la planta bajo diversas condiciones.

Con esto me refiero a que es posible estudiar y descubrir los diferentes genes implica-

dos en los comportamientos de la planta, (Liu *et al.*, 2021), en su trabajo titulado *Gene and Metabolite Integration Analysis through Transcriptome and Metabolome Brings New Insight into Heat Stress Tolerance in Potato (Solanum tuberosum L.)* donde investiga el comportamiento de las plantas de papa bajo estrés térmico, pues estos organismos son particularmente vulnerables a estrés abiótico, defina se el estrés abiótico como los factores ambientales que alteran los procesos fisiológicos y metabólicos de las plantas (Taiz y Zeiger, 2010), este esfuerzo nos provee de un perfil metabólico y transcriptómico para *Solanum tuberosum* aunque también es posible estudiar los diferentes tejidos de la planta para determinar transcriptómica mucho mas exacta como lo hace (Tiwari *et al.*, 2020) y (Tiwari *et al.*, 2021) sometiendo la planta a diferentes tipos de estrés y evaluando la cantidad de ARN en cada tejido es decir, raíces, tallo, hojas, etc.

Otro trabajo por la misma linea (Gong *et al.*, 2015), *Transcriptome Profiling of the Potato (Solanum tuberosum L.) Plant under Drought Stress and Water-Stimulus Conditions* nos demuestra el perfilamiento transcriptómico de *Solanum tuberosum* bajo estrés hídrico y condiciones constantes de humedad, lo que nos demuestra algunas de las posibilidades que se han hecho en variedades de papa que se cultivan en China, Japón, Rusia... lo que hace posible que el estudio de moléculas estimulantes del crecimiento vegetal, es decir indagar sobre como estas moléculas inciden en los procesos de crecimiento celular, lo podemos ver en (Lemke *et al.*, 2020). *Transcriptome Analysis of Solanum Tuberosum Genotype RH 89-039-16 in Response to Chitosan*

Los trabajos anteriores nos muestran el abanico de posibilidades que se abre al tener la información genética digitalizada, y la posibilidad de evaluar con exactitud, ahora bien si este mundo parece fascinante, comprender el sistema biológico de *Solanum tuberosum* es otro universo por explorar pues tal y como lo dice (Beukema *et al.*, 1990), pues es un cultivo ampliamente distribuido en mas de 140 países y mas de 100 de ellos se encuentran ubicados en zonas tropicales y sub-tropicales, aunque al menos un $\frac{1}{3}$ de la producción se encuentra en países desarrollados, principalmente de climas templados, liderados por China.

Para el caso de Colombia se puede evidenciar más de 100 variedades nativas entre papa

de año *Solanum tuberosum* y papa criolla *Solanum tuberosum Gr. Phureja* (Ñústez-López, 2011) y (Tinjacá y Rodríguez, 2015), esto solo nos indica que Colombia como país biodiverso, tiene mucho que ofrecer en el campo de la bioinformática, se consta en artículos como *Gene regulatory networks on transfer entropy (GRNTE): a novel approach to reconstruct gene regulatory interactions applied to a case study for the plant pathogen Phytophthora infestans* , en el cual se busca aprovechar la creciente cantidad de datos genómicos, que permiten el entendimiento de las dinámicas moleculares de sistemas complejos tal como son las enfermedades en plantas y animales, esto porque la regulación transcripcional juega un papel central en los procesos celulares, y esto aun es pobremente entendido (Castro *et al.*, 2019), no obstante el artículo nos presenta un modelo matemático que permite inferir redes de genes reguladores, por ello es importante el desarrollo de nuevas técnicas que permitan igualar el mercado, en la producción de papa, y convertirse en un país capacitado para competir y cumplir con la demanda de este alimento a lo largo del mundo.

Por otro lado la secuenciación de un organismo tan complejo como una planta de *Solanum tuberosum* solo es posible a las tecnologías actuales, pero esto puede ser contraproducente el desarrollo de nuevas técnicas y tecnologías, puede ser abrumador dada la cantidad de programas y formatos que existen en la actualidad, por eso (F.-W. Li y Alex, 2018), escribe la review *A guide to sequence your favorite plant genomes* donde podemos encontrar las herramientas y conceptos básicos sobre el universo que es la secuenciación.

La papa o *Solanum tuberosum* , es una herbácea anual que alcanza una altura de un metro y produce un tubérculo, la papa misma, con tan abundante contenido de almidón que ocupa el cuarto lugar mundial en importancia como alimento, después del maíz, el trigo y el arroz.(*The State of Food Security and Nutrition in the World 2020*, 2020) La papa pertenece a la familia de floríferas de las solanáceas, del género Solanum, formado por otras mil especies por lo menos, como el tomate y la berenjena. El S. tuberosum se divide en dos subespecies apenas diferentes: la andigena, adaptada a condiciones de días breves, cultivada principalmente en los Andes, y tuberosum, la variedad que hoy se cultiva en todo el mundo y se piensa que desciende de una pequeña introducción en Europa de papas andigena, posteriormente adaptadas a días más prolongados.(*La papa*, 2008)

El tubérculo, Al crecer, las hojas compuestas de la planta de la papa producen almidón, el cual se desplaza hacia la parte final de los tallos subterráneos, también llamados estolones.(Beukema *et al.*, 1990) Estos tallos sufren a consecuencia un engrosamiento y así se producen unos cuantos o hasta 20 tubérculos cerca de la superficie del suelo. El número de tubérculos que llegan a madurar depende de la disponibilidad de humedad y nutrientes del suelo. El tubérculo puede tener formas y tamaños distintos, y por lo general pesa hasta 300 g.(*La papa*, 2008)

La tecnología que ha permitido la progresión en las investigaciones in silico, ha sido el RNA-seq, esta es una tecnología perteneciente al grupo de secuenciaciones de nueva generación (Next Generation Sequencing, NGS), dicha tecnología permite la identificación y cuantificación de ARN presente en una muestra, permitiendo la inferencia sobre los genes que expresa dicho ARN, por ende es ampliamente utilizada para el análisis diferencial de genes, pero la principal dificultad de este modelo de análisis es la carencia de un workflow unificado, esto debido a la amplia oferta de herramientas y la generación de nuevas, dadas las ventajas que ofrece el open source (Hernández Ballesteros, 2016). He aquí la importancia de conocer los formatos utilizados en cada uno de los pasos de un análisis de RNA-seq, y a su vez las herramientas necesarias para manipular cada uno de estos formatos, entre ellas podemos encontrar Bowtie (Langmead y Salzberg, 2012), Samtools (Danecek *et al.*, 2021), HTSeq (Putri *et al.*, 2022), DESeq2 (Michael Love, 2017).

3. Marco Teórico

3.1. Generalidades de *Solanum tuberosum Gr. Phureja*

La papa comúnmente consumida es papa tetraploide también conocida como "papa blanca" o "papa de año" (2n=2X=48) Solanum tuberosum L. Grupo Andigenum (Ñústez-López y Rodríguez-Molano, 2020), por otro lado también existen papas de características diploides, ellas se congregan en el Grupo Phureja y poseen una amplia diversidad, además de una adaptación al día corto y ausencia de reposo, algunos de los representantes de este grupo puede ser Criolla Latina, Criolla Paisa y Criolla Colombia (Rodríguez *et al.*, 2009), cultivares desarrollados para el cultivo y comercialización en Colombia, las particularidades convierten a este conjunto de ejemplares cultivados en especiales, pues se cosecha con follaje verde y su periodo de pos-cosecha para consumo fresco es muy corto(Ñústez-López y Rodríguez-Molano, 2020).

El Grupo Phureja sitúa como el centro de su diversidad y pluralidad el sur de Colombia y el norte de Ecuador, estas se distribuyen y comercializan principalmente en el departamento de Nariño pues las múltiples variedades que existen de están papas diploides (Tornilla, Mambera y Ratona, entre otros) tienen un valor cultural en las comunidades nativas (Ñústez-López y Rodríguez-Molano, 2020).

3.1.1. Morfología

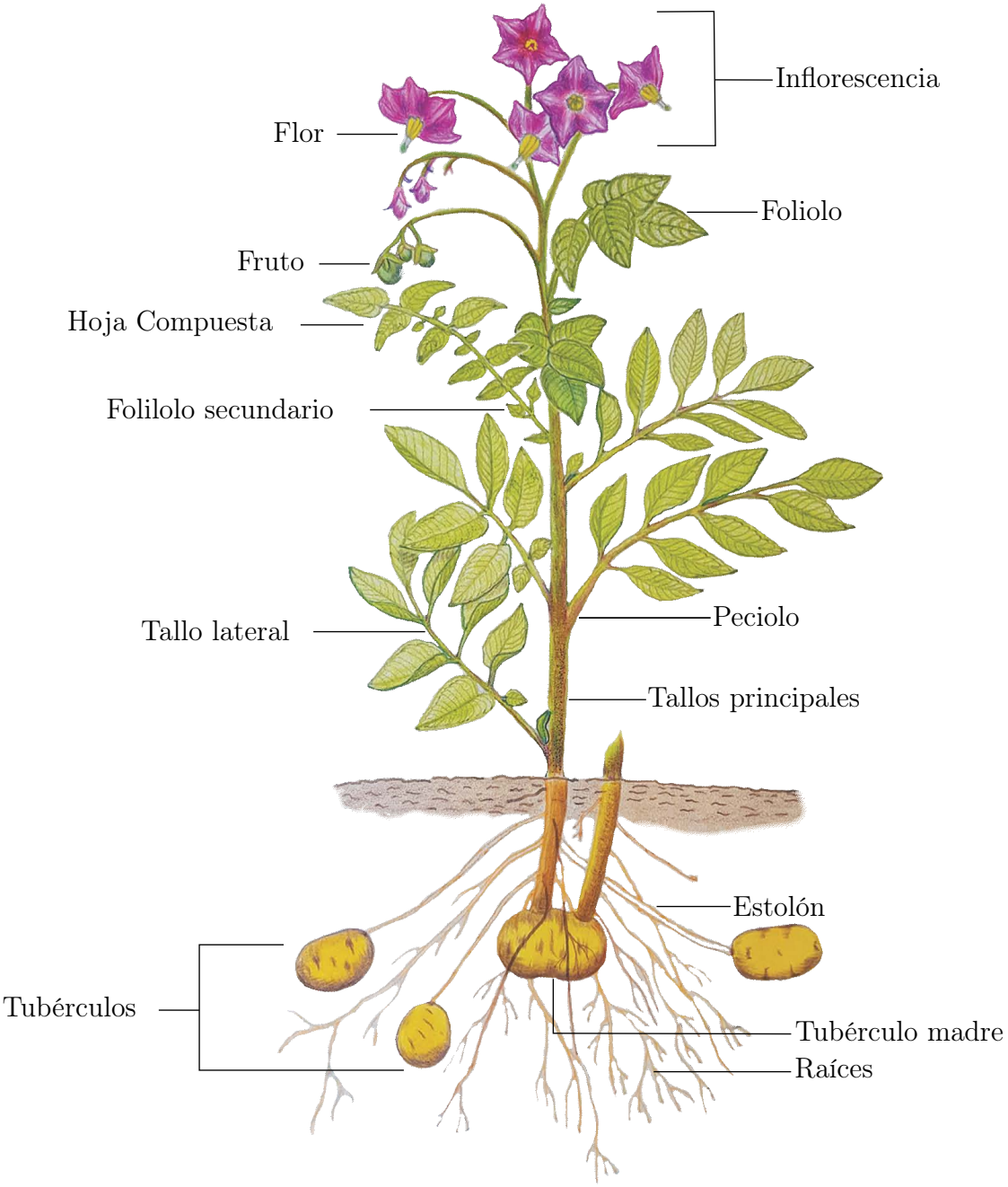

Figura 2: Morfología de una planta de *Solanum tuberosum* (Morales-Puntes, 2021)

3.1.2. Clasificación Taxonomica

Solanum tuberosum Gr. Phureja, es un grupo que se caracteriza por ser doble haploide diferenciándose del resto de especímenes de Solanum Tuberosum, que son tetraploides, esta junto con las carencia de periodo de reposo son los diferenciadores principales de este grupo.

Reino	Plantae
División	Magnoliophyta
Clase	Magnoliopsida
Orden	Solanales
Familia	Solanaceae
Género	Solanum
Especie	*Solanum tuberosum (Linneo, 1753)*

Tabla 1: Taxonomía *Solanum tuberosum*

3.1.3. Características Genéticas

Solanum tuberosum Gr. Phureja , también conocida como la papa Criolla posee peculiaridades en su genoma, pues estas Variedades Silvestres son diploides, a diferencia de la papa blanca o de año, *Solanum tuberosum Grupo Andigenum* la cual es tetraploide (Ñústez-López y Rodríguez-Molano, 2020). Colombia, ha sido identificado como un foco de diversidad de papas criollas, convirtiéndose en un recurso invaluable para el mejoramiento de la papa cultivada, el cual no ha sido completamente categorizado a nivel citogenético (Gómez Pulgarín *et al.*, 2012). esta información es de vital importancia para la tecnificación y el aumento de la producción de la papa criolla.

3.2. Dormacia

La dormancia es una fase del ciclo biológico de un organismo, donde su crecimiento, desarrollo y actividad física se anula temporalmente, como consecuencia de esta etapa del ciclo, su metabolismo se reduce drásticamente permitiendo así que el organismo conserve su energía, esta se presenta tanto en animales como plantas y otros organismos, y depende

generalmente de las condiciones ambientales. El fenómeno de la dormancia, se manifiesta en los diferentes grupos taxonómicos con diferentes estrategias, entre ellas podemos mencionar Diapausa en mamíferos, Estivación en Peces y anélidos, Brumación en Reptiles, y en plantas encontramos Semillas durmientes y Árboles durmientes. (Universidad Nacional de Colombia, 2014)

3.2.1. Dormancia en Papa

La formación del tubérculo de papa tiene un estrecho nexo biológico con la dormancia, este entra en un estado de dormancia al ser cosechado y durante un breve periodo de tiempo determinado, en el cual no se observa ningún crecimiento de brotes, esto es inducido principalmente por ácido abscísico (ABA) y etileno, no obstante solo el ácido abscísico es necesario para su mantenimiento, esta fitohormona incrementa su aparición durante la formación del tubérculo, se mantienen durante la dormancia y disminuyen con el crecimiento de los brotes. (Rodriguez y Moreno, 2010) muchos de estos genes tienen desarrollos pleitrópicos, lo que dificulta el estudio molecular de este proceso biológico.

3.3. Fitomejoramiento en *Solanum tuberosum Gr. Phureja*

La última década este producto vegetal, al igual que los otros representantes de *Solanum tuberosum* , han sido ampliamente estudiados tanto en Colombia como en otras partes del mundo, esto es por una las principales diferencias entre la papa de año o papa blanca (*Solanum tuberosum*) y la papa criolla (*Solanum tuberosum Gr. Phureja*), al ser diploide su sistema genético es mucho más sencillo de entender y modificar, pues la complejidad al aumentar el número de cromosomas, en el estudio y modificación aumenta de manera exponencial.

En Colombia la Universidad Distrital Francisco José de Caldascon el grupo de investigación BIOMOLC han liderado trabajos para el mejoramiento y establecimiento de protocolos moleculares e identificación de rutas metabólicas, importantes en el proceso de dor-

mancia (Baracaldo Huertas y Velasco Triana, 2017; Guzmán Vásquez, 2016; Sánchez Álvarez y Baracaldo Pinto, 2019; Zabala Pardo, 2015; Avila Vargas, 2018; Garay Alvarez y Espinosa Ladino, 2019; Parra Pérez y Ortíz Arévalo, 2017; García Mejía y Parra Rodríguez, 2019).

3.3.1. Variedad Criolla Colombia

La variedad Criolla Colombia es un varietal que proviene de *Solanum tuberosum Gr. Phureja* , fue registrado como nuevo cultivar ante el ICA en el 2005 en compañía de Criolla Latina y Criolla Paisa por (Rodríguez *et al.*, 2009), la características principales organolépticas de este clon, lo hicieron apto para su continuo trabajo, aunque estas dependen principalmente de las condiciones microclimáticas, altitud, radiación solar y humedad durante el ciclo de producción (Fano *et al.*, 1998).

Rodríguez *et al.* (2009), describe a Criolla Colombia con un hábito de crecimiento erecto, buen desarrollo de follaje, color verde claro, flor lila oscuro. Tubérculos de forma redonda, ojos semiprofundos, ausencia de periodo de reposo, color de piel y carne amarillo intenso.

3.4. Bioinformática en Colombia

Colombia es una sociedad que tiene unos niveles de producción de conocimiento bioinformático muy bajos, comparados con el resto de países en el mundo, ignorando las dificultades económicas que impiden el progreso científico, Colombia y la bioinformática tienen factores adicionales de fondo, estos radican esencialmente en un déficit académico en cuanto a la enseñanza de la bioinformática, siendo esta el principal motor en la investigación científica, un carencia en los procesos educativos produce un crecimiento lento, y por ende una pobre oferta de investigadores en este campo (Benítez-Páez y Cárdenas-Brito, 2010).

Aun así es un área con un creciente interés, dado que el nacimiento de la bioinformática como disciplina científica viene dada por la imperativa necesidad de analizar la enorme

cantidad de datos provenientes de las plataformas NGS, cuyo análisis e interpretación ha proporcionado nuevos conocimientos y perspectivas sobre las bases moleculares de las enfermedades humanas y vegetales y sobre la resistencia de las plantas a los estreses bióticos y abióticos (Madroñero *et al.*, 2019).

Muchos de estos trabajos de NGS realizados en Colombia aplicados a plantas, se orientan a epidemiología vegetal, es decir se usan estas herramientas moleculares para afrontar las problemáticas enfermedades que atacan los diferentes cultivos, esto lo podemos observar en trabajos como (Riascos Chica *et al.*, 2018; Gallo García *et al.*, 2019; Carreño *et al.*, 2007), donde se establecen protocoles moleculares y bioinformáticos para la identificación de virus u hongos en plantas de papa.

3.4.1. Bioinformática en *Solanum tuberosum Gr. Phureja*

En papa criolla, *Solanum tuberosum Gr. Phureja* los avances en bioinformática, han sido escasos y tienen como principal enfoque la lucha contra las principales patologías que enfrenta este cultivo, como lo pueden ser el Virus del amarillamiento, PYVV, que obstruye los tejidos conductores, dificultando el transporte de nutrientes lo que produce el amarillamiento en las venas de las hojas y la pérdida de vigor de la planta (ICA, 2013), como también en los procesos ómicos que permiten a *Phytophthora Infestans* infectar y enfermar la planta de papa (Pinzón *et al.*, 2009; Riascos Chica *et al.*, 2018).

También se están empezando usar los marcadores moleculares, en la selección asistida para el fitomejoramiento, no solo caracteres organolépticos (Veramendi y Gabriel, 2015), es por eso que recursos como Spud DB (C. D. Hirsch *et al.*, 2014), una base de datos especializada en Solanáceas, principalmente en papa (*Solanum tuberosum*), son de principal interés para los investigadores que buscan modelos predictivos y de eliminación más eficientes y rápidos para la creación de varietales.

La necesidad de detenerse a revisar las herramientas de análisis de *Solanum tuberosum Gr. Phureja* es como se mencionó anteriormente la diferencia en la ploidía del grupo, dado que la papa blanca o de año, *Solanum tuberosum* es tetraploide, y la papa criolla es diploi-

de, esto aunque existan estimados para la diferencia entre los sistemas biológicos mencionados, para el uso de la bioinformática en la agrigenómica no es suficiente, como ejemplo la realización de un DEG o un diagrama de expresión génica, comprando estos dos grupos de papa inducirá más error al diseño estadístico, usado en los programas de análisis (Weiß *et al.*, 2018).

3.5. Transcriptoma

De acuerdo con NHGRI (2019) un transcriptoma es una colección de todas las lecturas de genes presentes en una célula, definiendo las lecturas como los transcritos que se crean en la transcripción (lectura) de cada gen, es decir son ARN (ácido ribonucleico). Dicho de otro forma un transcriptoma es la recopilación de todo el ARN presente en un organismo ya sean unicelulares o pluricelulares.

Hay varias clases de ARN. La clase más importante, llamada ARN mensajero (ARNm), desempeña un papel vital en la elaboración de proteínas. En este proceso el ARNm se transcribe a partir de genes; luego, los transcritos de ARNm se entregan a los ribosomas, donde se leen o "traducen", la secuencia de las letras químicas en el ARNm y ensamblan componentes básicos llamados aminoácidos para formar proteínas. El ADN también puede transcribirse a otros tipos de ARN que no codifican proteínas. Tales transcritos pudieran servir para influir en la estructura celular y regular los genes (NHGRI, 2019).

Los experimentos de análisis del transcriptoma permiten a los investigadores caracterizar la actividad transcripcional (codificante y no codificante), centrarse en un subconjunto de genes y transcritos objetivo relevantes, o perfilar miles de genes a la vez para crear una imagen global de la función celular. Los estudios de análisis de expresión génica pueden proporcionar una instantánea de los genes y transcritos expresados activamente en diversas condiciones (illumina, 2021).

3.5.1. Experimentos RNA-Seq

El RNA-seq es un tipo especial de secuenciación cuyo objetivo es ver los patrones de expresión de los genes. En vez de leerse y secuenciarse todo el genoma u exoma, se centra en secuenciar el RNA mensajero (mRNA), que se traducirá a una secuencia de aminoácidos y dará como producto final una proteína. En este tipo de análisis siempre se trabaja con dos grupos de muestras, unas muestras "problema" y unas "control", con el fin de comparar los patrones de expresión entre unas y otras (Helix BioS, 2019).

Como lo muestra el texto publicado por illumina (2021), los experimentos de este tipo, pueden ser de gran utilidad para establecer un punto de partida y determinar el estado actual de un organismo, y poder llevar controles sobre los experimentos de estrés abiótico, biótico, o brindar información valiosa en los procesos de selección, incrementando los marcadores poblacionales que permitan acentuar una mutación a su vez que se mantiene una diversidad poblacional, con el fin de evitar procesos de declive.

3.5.2. SNPs

Un polimorfismo de nucleótido único (SNP) es una variante genómica en la posición de una base única en el ADN. Los científicos estudian si los SNP en un genoma influyen en la salud, la enfermedad, la respuesta a los fármacos y otros rasgos, y por cual mecanismo (NHGRI, 2022), los SNPs por sí mismos no proporcionan información sobre genes específicos; simplemente indican una localización cromosómica que es probable que esté estrechamente asociada con un fenotipo dado (Chial, 2008).

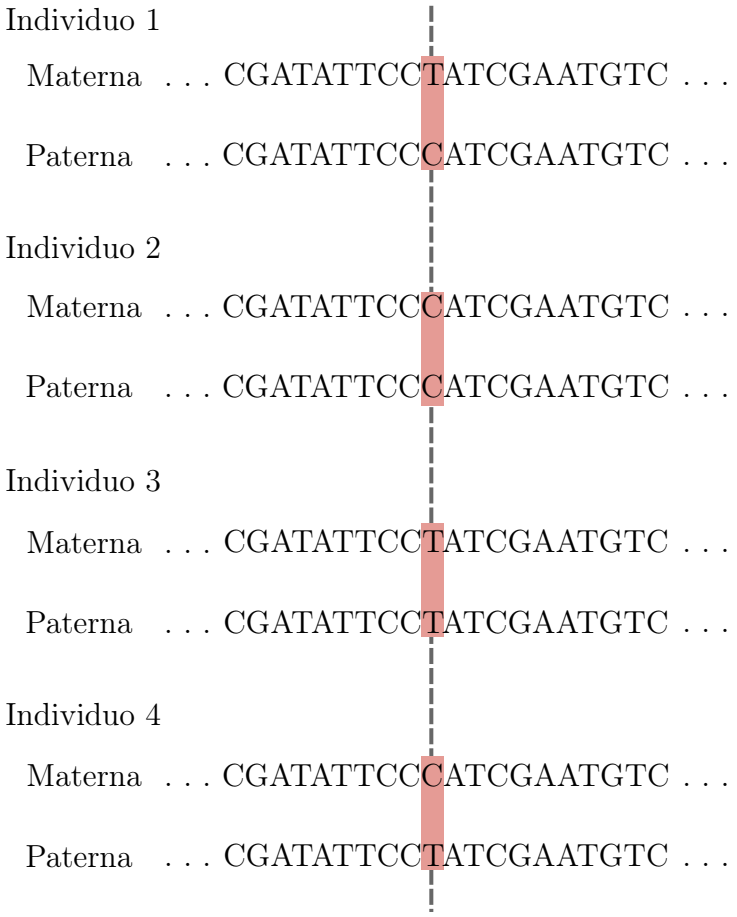

Figura 3: Diagrama SNPs tomada de (NHGRI, 2022)

Los SNP que se localizan dentro de una secuencia codificante pueden modificar o no la cadena de aminoácidos que producen; se llama SNP no-sinónimos a los primeros y SNP sinónimos (o mutaciones silenciosas) a los segundos (G. Li *et al.*, 2014; Kim y Misra, 2007).

3.5.3. Analisis de Expresión Diferencial

El análisis de expresión diferencial significa tomar los datos de conteo de lectura normalizados y realizar un análisis estadístico para descubrir cambios cuantitativos en los niveles de expresión entre grupos experimentales. Por ejemplo, usamos pruebas estadísticas para decidir si, para un gen determinado, una diferencia observada en el recuento de lecturas es significativa, es decir, si es mayor de lo que se esperaría debido a la variación aleatoria natural (EMBL, 2018).

Este método evalúa y permite el perfilamiento de un conjunto de genes o de un gen en particular, en situación relacionadas, es decir que tengan influencia sobre su expresión como lo pueden ser las enfermedades o las condiciones ambientales, esto también puede verse afectado al inducir mutaciones (EMBL, 2018; Huang *et al.*, 2015), la forma más común de expresar y visualizar esta información es con los diagramas conocidos como heat maps.

3.6. Datos Públicos en *Solanum tuberosum Gr. Phureja*

Esfuerzos conjuntos como lo son C. D. Hirsch *et al.* (2014), crearon la primera base de datos de papa, que tiene como objetivo almacenar genotipos y fenotipos, en busca de acelerar el mejoramiento de este cultivo, puesto que la papa fue la primera especie de Solanácea con un ensamble de genoma de alta calidad (PGSC, 2011), aunque también existen otros recursos que tiene como principal enfoque la papa, como pueden ser la base de datos PoMaMo que contiene mapas genéticos y secuencias (`http://www.gabipd.org/projects/Pomamo/`) y la base de datos Potato Pedigree Database que almacena información de pedigree para los cultivares de papa (`http://www.plantbreeding.wur.nl/potatopedigree/`) siguiendo el lanzamiento del genoma de papa en 2011, multiples datasets fenotípicos, genéticos y genómicos han sido generados para la papa (Felcher *et al.*, 2012; Hamilton *et al.*, 2011; C. N. Hirsch *et al.*, 2013; PGSC, 2011).

C. D. Hirsch *et al.* (2014) provee un recurso centralizado para la minería de estos conjuntos de datos o datasets con Spud DB, una base de datos centrada en papa con la información generada del consorcio para la secuenciación del genoma de papa o **PGSC** por sus siglas en inglés *Potato Genome Sequencing Consortium*, provee un acceso a la información más reciente de estos cultivares, añadiendo nuevos cultivares diferentes grupos pertenecientes a *Solanum tuberosum* , como lo pueden ser *Solanum tuberosum Gr. Phureja* .

4. Objetivos

4.1. Objetivo General

Establecer un protocolo de análisis bioinformático para la expresión genética y polimorfismos de un solo nucleótido en *Solanum tuberosum Gr. Phureja*

4.2. Objetivos Específicos

- Identificar las principales rutas de expresión en *Solanum tuberosum Gr. Phureja* implicadas en procesos de dormancia (*Latencia*).
- Reseñar los sitios polimórficos de mayor relevancia en los genes de interés para dormancia, estrés abiótico.

5. Metodología

5.1. Pipeline: BIOMOLC-PhurejaMutante

Un pipeline consiste en una cadena de eventos conectados donde la salida de cada elemento, es la entrada del siguiente, es un proceso que ha sido usado para la implementación de computadoras vectoriales eficientes y de alta velocidad, también es un método efectivo para implementar procesadores multinúcleo, esto la convierte en una arquitectura computacional con una atención considerable desde la década de 1960, cuando la necesidad de sistemas más rápidos y rentables se volvió crítica (Yizhen *et al.*, 2011; Dubey y Flynn, 1990; Veen, 1986; Jordan, 1984; Ramamoorthy y Li, 1977).

Ramamoorthy y Li (1977) comentaba que el mérito de los pipelines es que puede ayudar a igualar las velocidades de varios subsistemas sin duplicar el costo de todo el sistema involucrado. A medida que la tecnología evoluciona, se dispone de circuitos LSI más rápidos y económicos, y el futuro de la implementación de pipelines, ya sea en forma simple o compleja, se vuelve más prometedor.

También puede verse como una forma de incrustar paralelismo o concurrencia a un sistema informático, esto se refiere a la segmentación de los procesos computacionales, en múltiples sub-procesos que son ejecutados por unidades autónomas dedicadas, (cada núcleo de procesadores multinúcleo), estos procesos son secuenciales, pero interrelacionados (Yizhen *et al.*, 2011; Ramamoorthy y Li, 1977), es decir el procesamiento en pipeline es lo mismo que la producción de la línea de montaje, en la que la carga de trabajo de ejecución operativa se dividió en muchas secciones operativas equilibradas en el tiempo, y luego el punto de partida desde la entrada continua de la tubería, el período de operación ejecutar de forma superpuesta (Yizhen *et al.*, 2011).

5.1.1. Que tener en cuenta al diseñar un pipeline

Diseñar un pipeline tiene ventajas importantes como se puede leer anteriormente, pues la paralelización de eventos consecutivos acelera el proceso y facilita el manejo de los pasos intermedios, permitiendo así una escalabilidad y reproducibilidad del experimento, aunque ofrece ventajas el diseño de la misma debe ser transparente y robusto, permitiendo su reutilización, con esto en mente es importante descomponer todo el workflow en reglas (McConeghy, 2019, 9:45), que nos permitirán tener claridad sobre lo que debe ocurrir en cada paso de pipeline.

Para empezar el desglose de las fases para obtener nuestro resultado deseado 4, este trabajo busca establecer un pipeline para llamado de SNPs y para la realización de un análisis de expresión diferencial de los mutantes de *Solanum tuberosum Gr. Phureja* .

Lo primero a tener en cuenta a la hora de diseñar un pipeline es la facilidad de uso y modificación, lo que permite una actualización y mantenimiento de los scripts escritos y con esto asegurar su existencia en el tiempo, ahora también debe ser portable dado que su instalación y despliegue debe ser sencilla tanto en computadoras personales como workstation o cluster de servidores, por último debe existir documentación que permita que cualquier persona pueda usarlo, todo con fines de mantener la reproducibilidad de los experimentos.(Wratten *et al.*, 2021; P. A. Ewels *et al.*, 2020; Roy *et al.*, 2018; Leipzig, 2016; Hoon *et al.*, 2003)

Por esta razón se elige GNU make (Feldman, 1976) como herramienta intermedia entre el lanzamiento de los scripts y el usuario del pipeline, esto por las ventajas que ofrece a la hora de llevar un control sobre los resultados que provee cada paso, y evitar repeticiones de lo mismos pues GNU Make evitará relanzar un proceso que tenga un archivo de salida más reciente que el archivo de entrada. (Smith, 2016).

5.2. Diagrama General

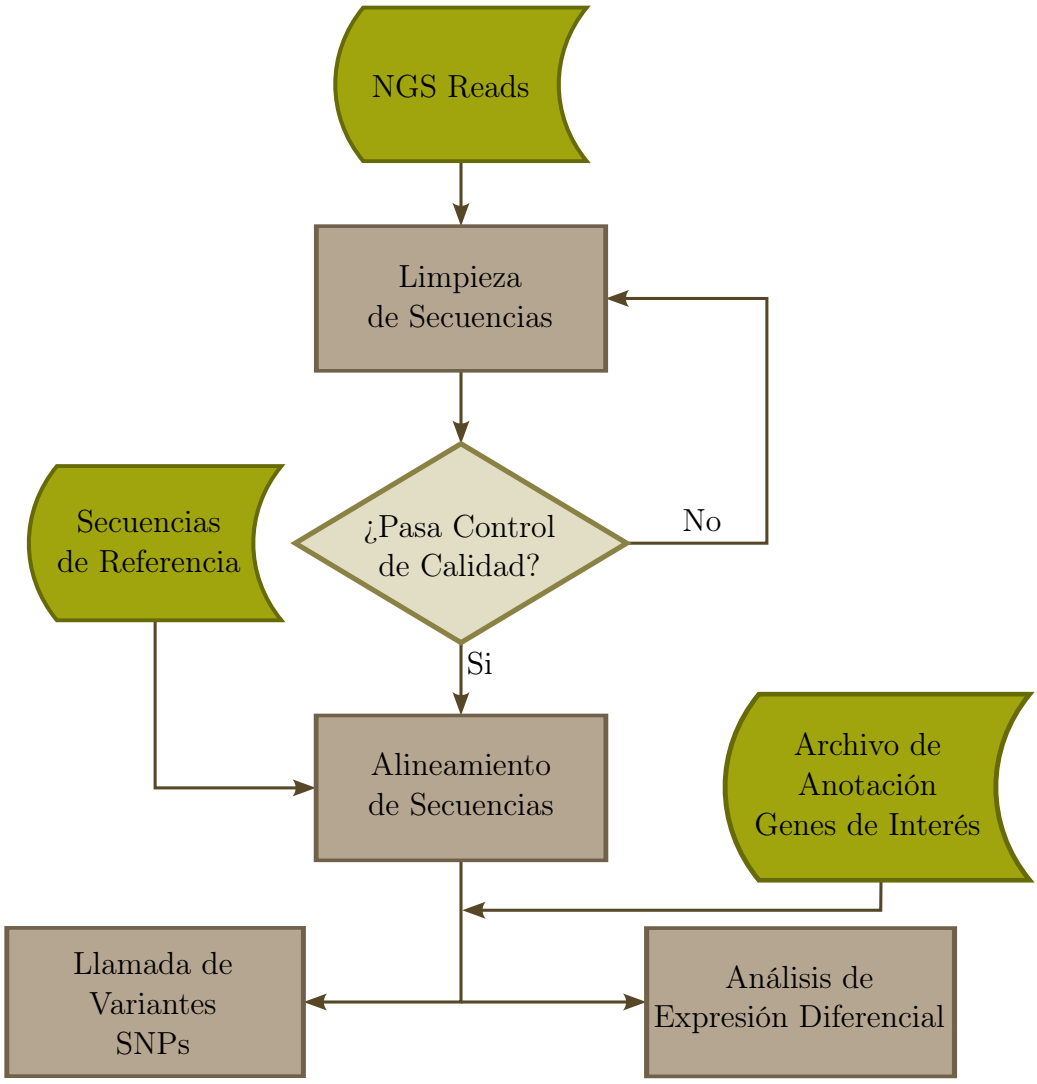

Figura 4: Diagrama representativo del pipeline

5.2.1. Datos en bruto

Los datos en bruto o raw data, son conocidos por tener un grado de curaduría bajo o nulo, existen algunas bases de datos que almacenan estos archivos como lo son el NCBI

Sequence Read Archive (SRA), mismo nombre que fue dado al archivo que almacena esta información el modelo de datos SRA se diseñó en colaboración con EBI y DDBJ bajo los auspicios de la colaboración internacional de bases de datos de secuencias de nucleótidos (INSDC) (`http://www.insdc.org`) (Shumway *et al.*, 2009).

El formato normalizado SRA se creó para admitir los principios FAIR (Findable, Accessible, Interoperable, Reusable), los datos de Sequence Read Archive (SRA), disponibles a través de múltiples proveedores en la nube y servidores NCBI, son el repositorio más grande de disponible públicamente de datos de secuenciación de alto rendimiento, un archivo SRA contiene datos sin procesar y contiene puntajes de calidad por base para todos los datos para poder extraer estos archivos en ficheros de texto plano en formato FASTQ, se hace uso del toolkit desarrollado por el mismo NCBI al diseñar en modelo de datos (Leinonen *et al.*, 2010).

Dentro de este kit de herramientas se encuentran dos herramientas actualmente fastq-dump y fasterq-dump, esta última es una herramienta desarrollada como una versión más rápida de fastq-dump esto implementando por defecto el multinúcleo de los procesadores actuales que se encuentran en máquinas personal y centros de cómputo, cualidad que será evaluada, para una posterior verificación con el dataset de prueba.

5.2.2. Filtrado y Control de Calidad

La limpieza o filtrado se realiza a las secuencias fastq extraídas de archivos SRA en busca de los adaptadores, RNA foráneos o Colas Poli A, pues la presencia de los mismos pueden alterar análisis futuros como lo pueden ser los análisis de expresión diferencial, para este fin a lo largo del desarrollo de la bioinformática como disciplina científica se han construido diferentes herramientas, entre las principales encontramos Trimmomatic desarrollada en java por Bolger, Lohse, y Usadel (2014) y cutadapt construida en python por Martin (2011).

Filtrar estas secuencias fue uno de los primeros obstáculos superados, para ello se hace uso de informes FastQC otra herramienta desarrollada en java por Andrew (2010), que nos

permiten visualizar las condiciones en las que se encuentra cada secuencia, esto lo hace a través de gráficas en HTML, que luego pueden ser compiladas en un solo informe usando MultiQC esta herramienta se desarrolló principalmente en python por P. Ewels, Magnusson, Lundin, y Käller (2016) esto con el fin de visualizar rápidamente grandes volúmenes de secuencias, pues FastQC realiza el informe por secuencia.

Teniendo en cuenta lo mencionado anteriormente, se comprobarán las dos herramientas principales para poder elegir la que se desempeñe con mejor rendimiento, teniendo en cuenta el tipo de data que se usará en este pipeline.

5.2.3. Alineamiento

Las alineaciones de secuencias son útiles en bioinformática para identificar la similitud de secuencias, producir árboles filogenéticos y desarrollar modelos de homología de estructuras de proteínas, para ello se hace uso de index que ahorran recursos de cómputo para análisis futuros, en el caso de este pipeline como lo pueden ser el llamado de variantes y el análisis de expresión diferencial.

El alineamiento necesita de una secuencia referencia en este caso se hará uso del transcriptoma alojado en SpudDB creada por C. D. Hirsch *et al.* (2014) perteneciente al genotipo de *Solanum tuberosum Gr. Phureja DM 1-3 516 R44 - v6.1*, al igual que un archivo de anotación con los genes a estudiar, del mismo modo se usará el dataset alojado en este lugar, llamado conjunto de modelos genéticos de alta confianza para *DM 1-3 516 R44 - v6.1.*

Con esta información se pretende evaluar el rendimiento de las principales herramientas disponibles, como son *Burrows-Wheeler Aligner* desarrollada por H. Li y Durbin (2009), bien conocido como BWA y Bowtie2 por Langmead y Salzberg (2012), ambos con gran trayectoria pues estos programas aún reciben mantenimiento y actualizaciones por parte de los grupos desarrolladores.

5.2.4. Llamado de Variantes

Los SNPs por sí mismos no proporcionan información sobre genes específicos; simplemente indican una localización cromosómica que es probable que esté estrechamente asociada con un fenotipo dado, por esta razón es importante localizarlos a lo largo de un genoma, pues como podemos observar en los trabajos de Visscher (2008); Lettre, Jackson, Gieger, Schumacher, y Berndt (2008); Gudbjartsson *et al.* (2008); Weedon *et al.* (2008), donde sitúan múltiples SNPs humanos que tienen inferencia directa sobre la altura en infantes y adultos.

Este es uno de los motivos por lo que realizar un levantamiento de SNP es muy importante para poder ubicar con facilidad regiones exónicas con caracteres deseados, esto con mayor detalle en áreas como la agrigenómica, para realizar este proceso se hace uso de herramientas como samtools desarrolla por Danecek *et al.* (2021) y bcftools por H. Li (2011) con estas dos herramientas se complementan y hacen parte del mismo paquete.

5.2.5. Analisis de Expresión Diferencial

Un análisis de expresión diferencial (DEA, por sus siglas en inglés) sirve para evaluar la capacidad de respuesta a un estímulo de un organismo, esto se logra mediante la comparación de un organismo en estado natural al que llamaremos control y un muestra expuesta a la condición que deseamos evaluar, un ejemplo de ello podría ser la exposición de una planta a estrés abiótico de tipo hídrico o periodos de latencia (Yalamanchili *et al.*, 2017).

Se secuencian estas dos muestras, y se compara la expresión génica de estos organismos con el fin de evaluar los genes que posee una relevancia mayor en este tipo de situaciones, y de al hacer un DEA a nivel transcriptómico, obtendremos el grupo total de los genes vitales para la supervivencia de la planta dándonos así un lista de caracteres deseados a potenciar en esta planta (EMBL, 2018), para ello es necesario tener un archivo de anotación funcional y un genoma o transcriptoma de referencia con el fin de evaluar este caso hipotético.

6.1. Datos de Prueba

Los datos aquí listados (tabla 2) fueron de utilidad para iteración y reparación de errores en la construcción del pipeline

Accesión SRA	Tejido y Organismo
SRR18888552	Tubérculo - Solanum Tuberosum
SRR18888553	Tubérculo - Solanum Tuberosum
SRR18888554	Tubérculo - Solanum Tuberosum
SRR18888555	Tubérculo - Solanum Tuberosum
SRR18272797	Plántula Hoja + Tallo - Solanum Tuberosum
SRR18272798	Plántula Hoja + Tallo - Solanum Tuberosum
SRR18272799	Plántula Hoja + Tallo - Solanum Tuberosum
SRR18272800	Plántula Hoja + Tallo - Solanum Tuberosum

Tabla 2: Datos usados que permitan la iteración de errores

6.2. Pipeline BIOMOLC-PhurejaMutante

Para facilitar el uso de variables en el interior del pipeline se crea un modulo que alojara la información de directorios, ficheros y recursos de la maquina como se puede observar en la siguiente figura 5.

```
#!/bin/sh
# Ubicaciones
DIR=$(pwd)
DATA=$DIR/data
FASTQC=$DATA/fastqc
FASTQ=$DATA/fastq
FQRAW=$FASTQ/raw
FQTRIM=$FASTQ/trim
FASTA=$DATA/fasta
BAM=$DATA/bam
SAM=$DATA/sam
SRA=$DATA/sra
VCF=$DATA/vcf
RESULT=$DIR/result
# Nombres
SNPname='HojavsTuberculo'
#Secuencias de Referencia
TREF=$FASTA/DM_1-3_516_R44_potato.v6.1.hc_gene_models.cdna.fa
GREF=$FASTA/DM_1-3_516_R44_potato_genome_assembly.v6.1.fa.gz
AREF=$FASTA/DM_1-3_516_R44_potato.v6.1.hc_gene_models.gff3.gz
#Capacidad de computo
THD=$(($(nproc --all)-1)) # Hilos disponibles en la maquina
RAM=$(($(awk '/MemTotal/ { printf "%.0f \n", $2/1024/1024 }'
/proc/meminfo)-1))G #Cantidad de RAM disponible en la maquina
```

Figura 5: Implementación variables del pipeline

6.2.1. Datos en bruto

La comparación de las herramientas nos confirma las ventajas que ofrecen las nuevas implementaciones que tienen en cuenta el auge de los procesadores multinúcleo.

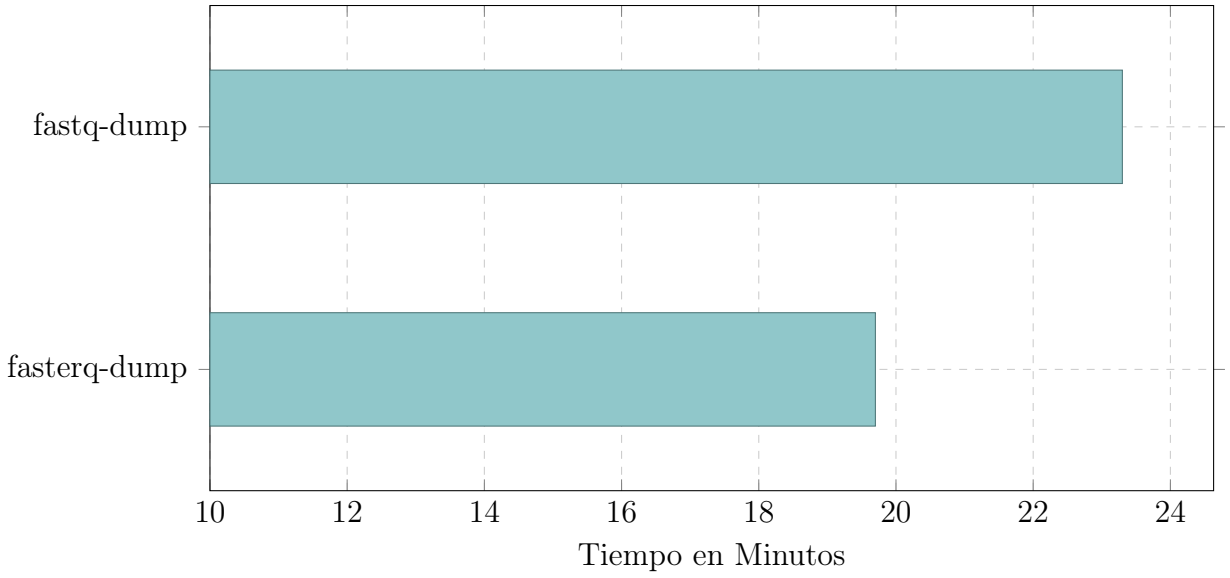

Figura 6: Herramienta *fastq-dump vs fasterq-dump*

Se hace uso del comando más rápido y se le asignan los valores más aptos de acuerdo a la máquina que lo ejecuta, como lo vemos en las variables, luego de obtenidos los archivos fastq, se procede a el filtrado.

```
#!/bin/sh
# Ejecución Faster-dump
find $SRA -iname "*.sra" -type f -execdir \
    fasterq-dump \
        --threads $THD \
        --mem $RAM \
        --outdir $FQRAW {} \;
# Renombrado de todas las secuencias quitando el sra
find $FQRAW -iname "*.fastq" -type f -execdir rename .sra_ _ {} \;
find $FQRAW -iname "*.fastq" -type f -execdir rename .fastq .fq {} \;
echo 'Comprimiendo Secuencias....'
# Compresion de todas las secuencias usando gzip
find $FQRAW -iname "*.fq" -type f -execdir pigz -k -p$THD {} \;
exit 0
```

Figura 7: Implementación fasterq-dump

6.2.2. Filtrado y Control de Calidad

Las utilidades comparadas haciendo uso de los datos tabla 2, muestran un consumo de recursos mas elevado por parte de Trimmomatic (Bolger *et al.*, 2014) esto asegurando la decision de implementar la herramienta cutadapt (Martin, 2011), por medio de TrimGalore! (Krueger *et al.*, 2021).

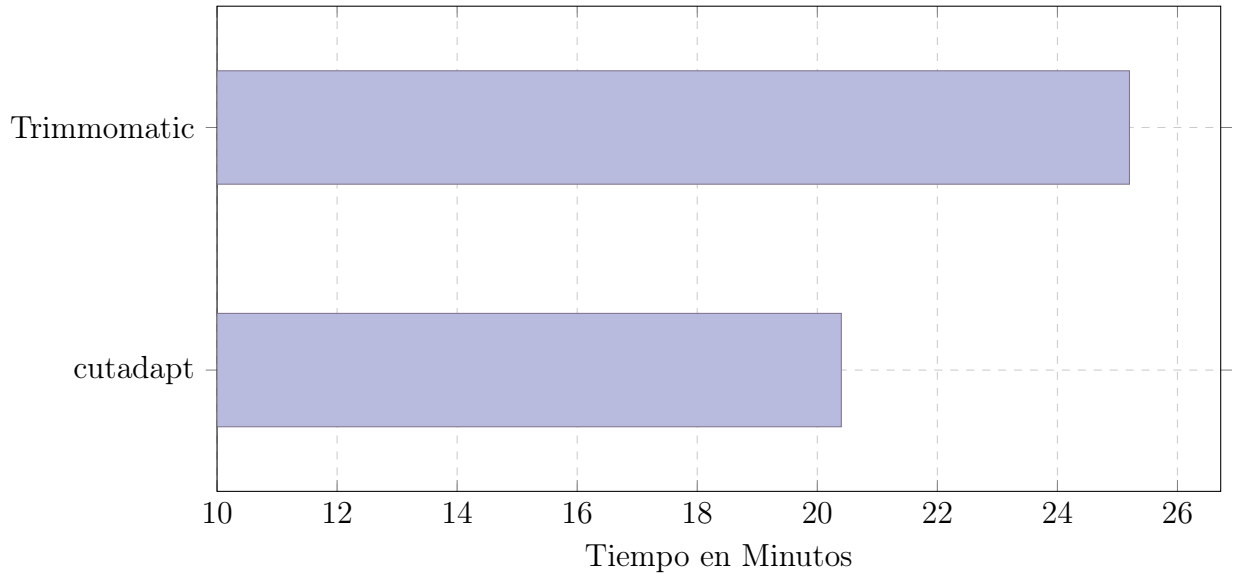

Figura 8: Herramienta *Trimmomatic vs cutadapt*

Con esto en mente podemos observar la implementación de estas herramientas en el pipeline figura 9.

```
#!/bin/bash
. ./bin/var.sh # Variables unificadas
# Lee todos los archivos fastq
for archivo in $FQRAW/*.fq.gz; do arreglo+=( "${archivo}" ) done
# Concatenar los dos archivos en un solo string para pasarlos al trimm
cont=$((${#arreglo[@]}-1))
x=0
while [ $x -le $cont ]; do
    files+=("${arreglo[$x]}"" ""${arreglo[$x+1]}")
    x=$(( $x + 2 ))
done
# Recorre cada par de secuencias y se lo pasa a trim_galore
# --gzip option quitada para prueba
for i in "${files[@]}"; do
    trim_galore \
        --paired \
        --q 30 \
        --cores $THD \
        --gzip \
        -o $FQTRIM \
        $i
done
```

Figura 9: Filtrado de secuencias FASTQ

Dado que el dataset elegido es pair end, se crea un arreglo que guarda las ubicaciones de ambas lecturas forward y reverse, eso es pasado a cutadat por medio de un wrapper llamado TrimGalore! el cual nos ayudará con banderas (flags) más legibles y autodetección de cebadores o adaptadores.

```
printf "Generando informes de secuencias.....\n"
fastqc -o $FASTQC -t $THD $FASTQ/*/*
# se compilan todos los informes usando MULTIQC
multiqc -o $RESULT $FASTQC
exit 0
```

Figura 10: Generación de Informes de Calidad

FastQC (Andrew, 2010) nos genera informes calidad de las secuencias, para esto hace uso de un solo núcleo, aunque cuenta con la opción de paralelaje permitiendo enviar todos los informes al mismo tiempo disminuyendo el tiempo de espera, creando un informe por hilo disponible, para compilar todos estos informes se hace uso de MultiQC (P. Ewels *et al.*, 2016), reuniendo todos los reportes en uno solo de fácil visualización, como lo podemos observar en la figura 10.

6.2.3. Alineamiento

El desempeño mas alto de BWA (H. Li y Durbin, 2009) puede deberse a las características de los datos, esto porque el algoritmo *Burrows-Wheeler Aligner*, es mucho mas eficiente con lecturas menores a 500pb, recordamos que estamos trabajando con secuencias de RNA.

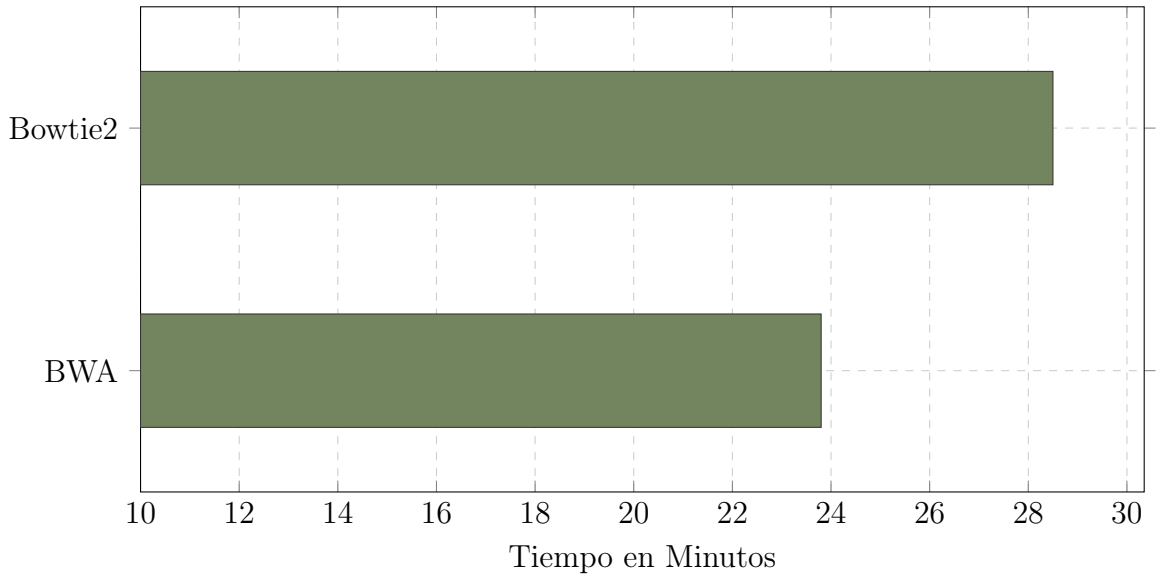

Figura 11: Herramienta *BWA vs Bowtie2*

El proceso de alineamiento se puede separar en pequeñas tareas que consisten en la indexación de la secuencia referencia y la creación de dos arreglos cada uno conteniendo los forward y los reverse, al igual que uno para recopilar los ID de las secuencias.

```
#!/bin/bash
. ./bin/var.sh # Variables unificadas
## bwa index my.fasta --> my.fasta == Secuencia de referencia
bwa index $TREF
# Lee todos los archivos el la carpeta data/fastq y los guarda en un arreglo
# Guarda los nombres de forward
for file in $FQTRIM/*1.fq.gz; do
    filesF+=( "${file##*/}" )
done
# Guarda los nombres de reverse
for file in $FQTRIM/*2.fq.gz; do
    filesR+=( "${file##*/}" )
done
# Captura del ID
for nameid in "${filesF[@]}"; do
    id+=(${nameid%%_*})
done
# Correr el bwa mem para todas las secuenias pareadas
for ((c=0; c<=$((${#id[@]}-1)); c++))
do
    bwa mem $TREF -t $THD $FQTRIM/${filesF[$c]} $FQTRIM/${filesR[$c]} > $SAM/${id[$c]}.sam
done
```

Figura 12: Implementación Alineamiento

6.2.4. Llamado de Variantes

El llamado de variantes primero necesita de la conversión del archivo sam al archivo bam, este se hace con samtools y solo se ejecuta en un núcleo, razón por la cual al hacer uso de GNU parallel (Tange, 2011) optimizamos este paso, disminuyendo notablemente el tiempo total que tomaría hacer este proceso, como se puede observar en la figura 13.

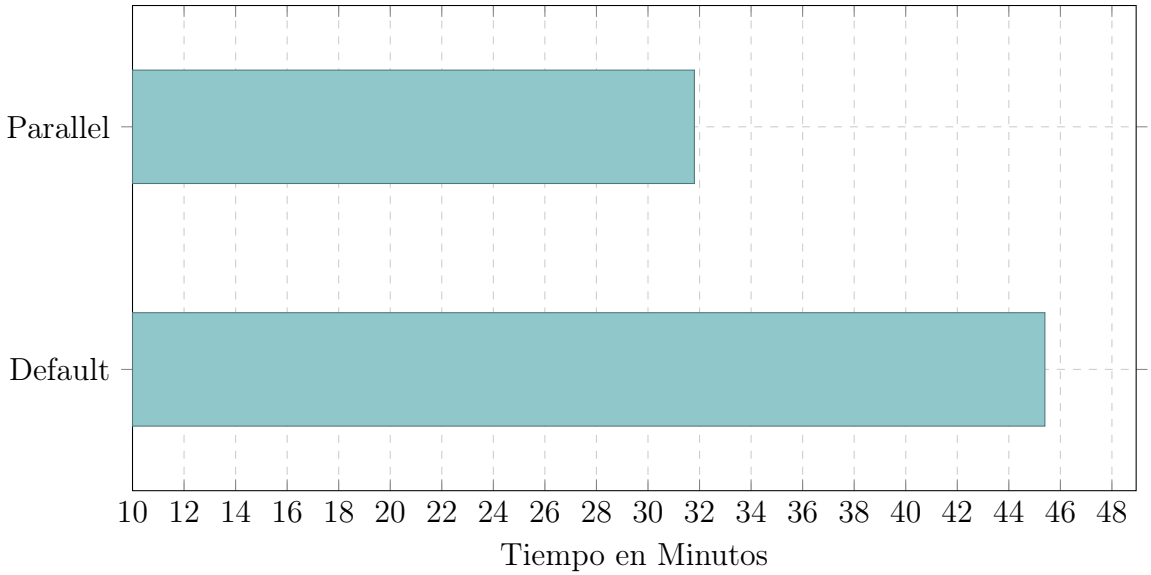

Figura 13: Herramienta *SAMtools vs SAMTools + Parallel*

Para implementar GNU parallel (Tange, 2011) es tan sencillo como listar todos los archivos que necesitemos y pasarselos por medio de una tuberia a el comando, parallel se encargara de usar todos los hilos disponibles sin sofocar la maquina, ya que es una implementación que interactua directamente con el kernel.

```
#!/bin/bash
. ./bin/var.sh # Variables unificadas
# Guarda los nombres de archivos .sam
for file in $SAM/*.sam; do
    samF+=( "${file##*/}" )
done
# Captura del ID
for nameid in "${samF[@]}"; do
    id+=(${nameid%%.*})
done
echo 'Convirtiendo secuencias .sam a .bam'
find $SAM -iname "*.sam" -print | parallel 'samtools view -S -b {} > {}.bam'
find $SAM -iname "*.bam" -execdir rename .sam.bam .bam {} \;
# Mover archivos bam a la carpeta correspondiente
find $SAM -iname "*.bam" -execdir  mv {} $BAM \;
# Guarda los nombres de archivos .bam
for file in $BAM/*.bam; do
    bamF+=( "${file##*/}" )
done
echo 'Ordenando secuencias .bam'
find $BAM -type f \( -iname "*.bam" ! -iname "*sorted*" \)
-print | parallel  'samtools sort {} -o {}_sorted.bam'
exit 0
```

Figura 14: Implementación conversión de formato SAM a BAM

Al tener los archivos bam, podemos iniciar el proceso de llamado de variantes, se necesita indexar de nuevo la secuencia referencia esto porque hace uso de otra herramienta, en este caso SAMtools (Danecek *et al.*, 2021) en el anterior usamos BWA (H. Li y Durbin, 2009), luego corremos otra utilidad de SAMTools llamada bcftools (H. Li, 2011) que nos permite extraer las variantes de las secuencias en BAM con usando como referencia el transcriptoma descargado de spud DB (C. D. Hirsch *et al.*, 2014).

```
#!/bin/bash
. ./bin/var.sh # Variables unificadas
# Indexación
samtools faidx $TREF
## Generar archivo raw de VCF Variant Call format todo los archivos ordenados sorted.bam
# Arreglo con todos los sorted.bam
for file in $BAM/*_sorted.bam; do
    bamFS+=( "${file##*/}" )
done
# Se crea el VCF file
samtools mpileup -f $TREF -s $BAM/"${bamFS[@]}" > $VCF/$SNPname-raw.bcf
for file in $VCF/*-raw.bcf; do
    vfcraw+=( "${file##*/}" )
done
# Call SNPs
bcftools view -vcg $VCF/*raw.bcf > $VCF/$SNPname-call.bcf
# Filtrado de SNPs
bcftools view $VCF/$SNPname-call.bcf | vcfutils.pl varFilter - > $VCF/$SNPname.vcf
exit 0
```

Figura 15: Implementación llamada de variantes SNP

6.2.5. Analisis de Expresión Diferencial

Un DEA consiste principalmente en dos partes, la primera usando HTSeq (Putri *et al.*, 2022) y SAMtools (Danecek *et al.*, 2021), para cuantificar las expresiones de los genes identificados en un archivo de anotación funcional, esto lo podemos evidenciar en la figura 16.

```
#!/bin/bash
. ./bin/var.sh # Variables unificadas
samtools view $SAM/SRR18272798.sam | htseq-count --idattr=Parent -s no - $AREF > $DEA/${id[0]}.txt
```

Figura 16: Implementación cuantificación de expresión

La segunda parte se realiza en el lenguaje R (*A free software project*, 2005), haciendo uso de la librería DESeq2 (Love *et al.*, 2014) como se evidencia en la figura 17, el modelo

estadístico negativo binomial del cual hace uso la librería tiene ciertos requerimientos que se Iran explicando en el código.

```
library(‘‘DESeq2’’)
sample.names <- sort(paste(c(‘‘MT’’, ‘‘WT’’), rep(1:3, each=2), sep=‘‘‘‘))
file.names <- paste(‘‘../’’, sample.names, ‘‘/’’, sample.names, ‘‘.count.txt’’, sep=‘‘‘‘)
conditions <- factor(c(rep(‘‘MT’’, 3), rep(‘‘WT’’, 3)))
sampleTable <- data.frame(sampleName=sample.names,
fileName=file.names,
condition=conditions)
# Leemos la cuantificación que hicimos en el paso anterior
ddsHTSeq<-DESeqDataSetFromHTSeqCount(sampleTable=sampleTable, directory=‘‘.’’,design=~ condition )
# Corremos el analisis
ddsHTSeq <- ddsHTSeq[rowSums(counts(ddsHTSeq)) > 10, ]
dds <-DESeq(ddsHTSeq)
# Revision de calidad
rld <- rlogTransformation(dds, blind=FALSE)
# grafica PCA
plotPCA(rld, intgroup=‘‘condition’’, ntop=nrow(counts(ddsHTSeq)))
# Grafica heatmap de correlacion
cU <-cor( as.matrix(assay(rld)))
cols <- c( ‘‘dodgerblue3’’, ‘‘firebrick3’’ )[condition]
heatmap.2(cU, symm=TRUE, col= colorRampPalette(c(‘‘darkblue’’,’’white’’))(100),
    labCol=colnames(cU), labRow=colnames(cU),
    distfun=function(c) as.dist(1 - c), trace=‘‘none’’, Colv=TRUE,
    cexRow=0.9, cexCol=0.9, key=F, font=2,
    RowSideColors=cols, ColSideColors=cols)
```

Figura 17: Implementación del DEA

Yalamanchili *et al.* (2017), en su trabajo *Data analysis pipeline for RNA-seq experiments: From differential expression to cryptic splicing*, explica con claridad los procesos necesarios para llevar acabo un DEA exitoso, teniendo en cuenta las dinámicas biológicas y moleculares a las que se enfrenta el ARN.

6.2.6. Automatización Pipeline

GNU Make, creado por Feldman (1976) fue elegido por la facilidad de uso y fácil implementación a un conjunto de scripts que funcionan en pipeline, la estructura del documento es la siguiente.

```
#Módulos de pipeline
test:
    ./bin/test.sh
sra2fq:
    ./bin/sra2fq.sh
trimm:
    ./bin/qc-trimm.sh
align:
    ./bin/align.sh
    ./bin/sam2bam.sh
snp:
    ./bin/snpcall.sh
dea:
    ./bin/dea-cont.sh
    Rscript ./bin/dea.r
clean:
    ./bin/clean.sh
```

Estos módulos pueden ser lanzados con facilidad en un emulador de terminal linux, lo que facilita su usabilidad, se observa en la figura 18.

Figura 18: Ejemplo uso GNU Make

6.3. Publicación del Pipeline: BIOMOLC-PhurejaMutante

El repositorio público de código abierto GitHub, es el host designado para el almacenamiento de millones de proyectos de código libre, además de tener versiones gratuitas, que permiten a este proyecto tener una presencia en internet a la vez que presta una facilidad al poseer un sistema de control de versiones. Se puede encontrar en el siguiente link `https://github.com/quinterol/BIOMOLC-PhurejaMutante`.

6.4. Genes implicados en la dormancia

Los genes aquí listados son de interés para el grupo de investigación BIOMOLC de la Universidad Distrital Francisco José de Caldaspara *Solanum tuberosum Gr. Phureja* , pues se busca estudiar la dormancia o latencia de las semillas-tuberculos y aunque estos genes tengan comportamientos pleitrópicos, las investigaciones realizadas en el interior del grupo han mostrado la inferencia de estos genes en este proceso biológico.

Gen ID	Trabajos de BIOMOLC
EIN2 CDPK7	Análisis de la expresión del Gen CDPK7 y evaluación del Gen EIN2 en papa criolla Solanum Tuberosum Vf. Phureja, (Variedad Criolla Colombiana) Irradiada con Cobalto 60 (Sánchez Álvarez y Baracaldo Pinto, 2019)
ABI4	Evaluación de la expresión del gen ABI4 y proteína DELLA (GAI) en mutantes candidato de Papa Criolla (Solanum tuberosum Grupo Phureja) obtenido con irradiación de cobalto 60 (Garay Alvarez y Espinosa Ladino, 2019)
CYP707A1	Identificación de CYP707A1 en la Síntesis de ABA involucrada en la respuesta por estrés hídrico en un cultivo de papa criolla (Solanum tuberosum grupo Phureja) irradiada con cobalto 60 ubicado en el municipio El Rosal, Finca El Pino Km 16 vía Subachoque, Cundinamarca (Parra Pérez y Ortíz Arévalo, 2017)
NBP35	Análisis de la expresión del gen Nbp35 en la ruta del jasmonato del cultivo de Solanum Tuberosum Vf. Phureja irradiada con Co60 (Avila Vargas, 2018)
NCED	Evaluación de la Expresión del Gen NCED, que Codifica para la Enzima 9-Cis Epoxicarotenoide Dioxigenasa, en Tubérculos de Mutantes Sólidos (Irradiados con Cobalto 60) de Solanum Tuberosum Grupo Phureja, Variedad Criolla Colombia (García Mejía y Parra Rodríguez, 2019)

Tabla 3: Investigaciones realizadas en BIOMOLC sobre *Solanum tuberosum Gr. Phureja*

7. Conclusiones

Se realizo el pipeline para el análisis de transcriptomas de *Solanum tuberosum Gr. Phureja* se realizo el proceso de automatización de los procesos de análisis de secuencias de tipo SRA , este proceso se puede utilizar en el análisis de diferentes organismos.

Se realiza un aporte pedagógico al proceso de análisis de secuencias al generar un flujo de trabajo y explicación del mismo en un lenguaje mas accesible para profesionales no informáticos que requieran esta herramienta.

El comparar varias herramientas de análisis permite el desarrollo de mejores compilaciones de las mismas par disminuir el tiempo de computo empleado en estos procesos.

Se dinamiza el proceso de análisis y expresión diferencial de genes al generar integración de herramientas robustas que permitan el análisis de fenómenos fisiológicos que tienen un comportamiento pleitrópicos como lo es la Latencia o Dormancia.

El desarrollo de pipelines son necesarios dentro de un grupo de investigación, ya que estos son el equivalente al desarrollo de protocolos de laboratorio, que permiten la estandarización de procesos como lo pueden ser la extracción de ADN, la creación de Geles para electroforesis, por este motivo es importante sumar esfuerzos para el fortalecimiento de esta rama dentro de los grupos de investigación.

8. Recomendaciones

Para próximos proyectos se recomienda:

- Se recomienda realizar secuenciación tanto en bulk como por tejidos de los mutantes sólidos y también de la variedad Criolla Colombia.
- Centrar esfuerzos en la creación de un archivo de anotación funcional, con un dataset de los genes de interés para el grupo de investigación, pues es un pilar importante para la construcción de análisis de sistemas biológicos con gran potencial de aplicación.
- Evaluar los mutantes en diferentes condiciones de estrés tanto biótico como abiótico, para la realización de DEA.

Referencias

Andrew, S. (2010). *Fastqc: A quality control tool for high throughput sequence data.* Descargado de `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

Avila Vargas, J. A. (2018, Ago). *Análisis de la expresión del gen Nbp35 en la ruta del jasmonato del cultivo de Solanum Tuberosum Vf. Phureja irradiada con Co60* (Inf. Téc.). Descargado de `http://hdl.handle.net/11349/14954`

Baracaldo Huertas, C. L., y Velasco Triana, D. (2017, Nov). *Estandarización de un medio de cultivo para la obtención a partir del cultivo in vitro de Solanum Phureja mutante flor blanca* (Inf. Téc.). Descargado de `http://hdl.handle.net/11349/14947`

Benítez-Páez, A., y Cárdenas-Brito, S. (2010). Bioinformática en Colombia: presente y futuro de la investigación biocomputacional. *Biomédica*, *30*(2), 170. doi: 10.7705/biomedica.v30i2.180

Beukema, H., van der Zaag, D. E., y cols. (1990). *Introduction to potato production* (n.º 633.491 B4). Pudoc Wageningen.

Bolger, A. M., Lohse, M., y Usadel, B. (2014, apr). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Carreño, N., Vargas, A., Bernal, A., y Restrepo, S. (2007). Problemas fitopatológicos en especies de la familia Solanaceae causados por los géneros Phytophthora, Alternaria y Ralstonia en Colombia. Una revisión. *Agronomía Colombiana*, *2*(25), 320–329. Descargado de `https://revistas.unal.edu.co/index.php/agrocol/article/view/14136`

Castro, J. C., Valdés, I., Gonzalez-García, L. N., Danies, G., Cañas, S., Winck, F. V., ... Riaño-Pachón, D. M. (2019, apr). Gene regulatory networks on transfer entropy (GRNTE): a novel approach to reconstruct gene regulatory interactions applied to a case study for the plant pathogen phytophthora infestans. *Theoretical Biology and Medical Modelling*, *16*(1). doi: 10.1186/s12976-019-0103-7

Chial, H. (2008). *Polygenic Inheritance and Gene Mapping.* Descargado de `https://www.nature.com/scitable/topicpage/polygenic-inheritance-and-gene`

-mapping-915/?error=cookies_not_supported&code=7464a5ae-4d15-41b6-88c9-da9258402d60

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... Li, H. (2021, 02). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). Descargado de https://doi.org/10.1093/gigascience/giab008 (giab008) doi: 10.1093/gigascience/giab008

Dubey, P. K., y Flynn, M. J. (1990, jan). Optimal pipelining. *Journal of Parallel and Distributed Computing*, *8*(1), 10–19. doi: 10.1016/0743-7315(90)90064-v

European Molecular Biology Laboratory. (2018). *Differential gene expression analysis.* Descargado de https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/rna-sequencing/performing-a-rna-seq-experiment/data-analysis/differential-gene-expression-analysis/

Ewels, P., Magnusson, M., Lundin, S., y Käller, M. (2016, jun). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. doi: 10.1093/bioinformatics/btw354

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., ... Nahnsen, S. (2020, feb). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, *38*(3), 276–278. doi: 10.1038/s41587-020-0439-x

Fano, H., Carmona, G., Ordinola, M., y Scott, G. (1998). Experiencias de exportacion de la papa amarilla peruana. *Centro Internacional de la Papa (CIP)*. Descargado de https://agris.fao.org/agris-search/search.do?recordID=QP1999000068

Fao statistics. (2021). FAO. Descargado de http://www.fao.org/faostat

Felcher, K. J., Coombs, J. J., Massa, A. N., Hansey, C. N., Hamilton, J. P., Veilleux, R. E., ... Douches, D. S. (2012, apr). Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS ONE*, *7*(4), e36347. doi: 10.1371/journal.pone.0036347

Feldman, S. (1976, Abr). *Make - GNU Project.* Descargado de https://www.gnu.org/software/make/

A free software project. (2005). R-Project. Descargado de https://cran.r-project.org/

doc/html/interface98-paper/paper_2.html

Gallo García, Y. M., Sierra Mejía, A., Donaire Segarra, L., Aranda, M., Gutiérrez Sánchez, P. A., y Marín Montoya, M. (2019). Coinfección natural de virus de ARN en cultivos de papa *Solanum tuberosum subsp. Andigena* en Antioquia (Colombia). *Acta Biológica Colombiana*, *24*(3), 546–560. doi: 10.15446/abc.v24n3.79277

Garay Alvarez, N. A., y Espinosa Ladino, L. G. (2019, Nov). *Evaluación de la expresión del gen ABI4 y proteína DELLA (GAI) en mutantes candidato de Papa Criolla (Solanum tuberosum Grupo Phureja) obtenido con irradiación de cobalto 60* (Inf. Téc.). Descargado de http://hdl.handle.net/11349/22267

García Mejía, E. D., y Parra Rodríguez, N. O. (2019, Nov). *Evaluación de la Expresión del Gen NCED, que Codifica para la Enzima 9-Cis Epoxicarotenoide Dioxigenasa, en Tubérculos de Mutantes Sólidos (Irradiados con Cobalto 60) de Solanum Tuberosum Grupo Phureja, Variedad Criolla Colombia* (Inf. Téc.). Descargado de http://hdl.handle.net/11349/22264

Gómez Pulgarín, T. M., López Ortiz, J. B., Pineda Tuirán, R., Galindo López, L. F., Arango Isaza, R., y Morales Osorio, J. G. (2012). Caracterización citogenética de cinco genotipos de papa criolla, solanum phureja (juz. et buk.). *Revista Facultad Nacional de Agronomía Medellín*, *65*(1), 6379–6387.

Gong, L., Zhang, H., Gan, X., Zhang, L., Chen, Y., Nie, F., ... Song, Y. (2015, may). Transcriptome profiling of the potato (solanum tuberosum l.) plant under drought stress and water-stimulus conditions. *PLOS ONE*, *10*(5), e0128041. doi: 10.1371/journal.pone.0128041

Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., ... Stefansson, K. (2008, apr). Many sequence variants affecting diversity of adult human height. *Nature Genetics*, *40*(5), 609–615. doi: 10.1038/ng.122

Guzmán Vásquez, J. D. (2016, Feb). *Caracterización fenotipica de un cultivo de papa criolla (Solanum tuberosum grupo phureja, variedad criolla colombia) irradiada con cobalto 60 ubicado en el municipio El Rosal, Finca El Pino Km 16 vía Subachoque, Cundinamarca* (Inf. Téc.). Descargado de http://hdl.handle.net/11349/23147

Hamilton, J. P., Hansey, C. N., Whitty, B. R., Stoffel, K., Massa, A. N., Deynze, A. V., ... Buell, C. R. (2011, jun). Single nucleotide polymorphism discovery in elite north american potato germplasm. *BMC Genomics*, *12*(1). doi: 10.1186/1471-2164-12-302

Helix BioS. (2019). *Análisis de RNA-seq .* Descargado de `http://www.helixbios.com/analisis-de-rna-seq`

Hernández Ballesteros, S. (2016). *Desarrollo de una gui para el anÁlisis de datos de secuenciaciÓn genÓmica* (Tesis de Master no publicada). Universidad Autónoma de Madrid.

Hirsch, C. D., Hamilton, J. P., Childs, K. L., Cepela, J., Crisovan, E., Vaillancourt, B., ... Buell, C. R. (2014). Spud DB: A Resource for Mining Sequences, Genotypes, and Phenotypes to Accelerate Potato Breeding. *The Plant Genome*, *7*(1). doi: 10.3835/plantgenome2013.12.0042

Hirsch, C. N., Hirsch, C. D., Felcher, K., Coombs, J., Zarka, D., Deynze, A. V., ... Buell, C. R. (2013, jun). Retrospective view of north american potato (*Solanum tuberosum L.*) breeding in the 20th and 21st centuries. *G3 Genes|Genomes|Genetics*, *3*(6), 1003–1013. doi: 10.1534/g3.113.005595

Hoon, S., Ratnapu, K. K., ming Chia, J., Kumarasamy, B., Juguang, X., Clamp, M., ... Stupka, E. (2003, jul). Biopipe: A flexible framework for protocol-based bioinformatics analysis. *Genome Research*, *13*(8), 1904–1915. doi: 10.1101/gr.1363103

Huang, H.-C., Niu, Y., y Qin, L.-X. (2015). Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software. *Cancer Informatics*, *14s1*, CIN.S21631. doi: 10.4137/cin.s21631

illumina. (2021). *High-impact discovery through gene expression and regulation research.* illumina. Descargado de `https://www.illumina.com/on-domain/GM-CPBU-GEX-eBook-Q42016_Landing-Page.html`

Jordan, H. (1984). Experience with pipelined multiple instruction streams. *Proceedings of the IEEE*, *72*(1), 113–123. doi: 10.1109/proc.1984.12823

Kim, S., y Misra, A. (2007). SNP Genotyping: Technologies and Biomedical Applications. *Annual Review of Biomedical Engineering*, *9*(1), 289–320. doi: 10.1146/

annurev.bioeng.9.060906.152037

Krueger, F., James, F., Ewels, P., Afyounian, E., y Schuster-Boeckler, B. (2021). *Trimgalore v0.6.7.* Zenodo. doi: 10.5281/ZENODO.5127899

Langmead, B., y Salzberg, S. L. (2012, mar). Fast gapped-read alignment with bowtie 2. *Nature Methods*, *9*(4), 357–359. doi: 10.1038/nmeth.1923

La papa. (2008). FAO - Food and Agriculture Organization of the United Nations. Descargado de `https://www.fao.org/potato-2008/es/lapapa/index.html`

Leinonen, R., Sugawara, H., y and, M. S. (2010, nov). The sequence read archive. *Nucleic Acids Research*, *39*(Database), D19–D21. doi: 10.1093/nar/gkq1019

Leipzig, J. (2016, mar). A review of bioinformatic pipeline frameworks. *Briefings in Bioinformatics*, bbw020. doi: 10.1093/bib/bbw020

Lemke, P., Moerschbacher, B. M., y Singh, R. (2020, aug). Transcriptome analysis of solanum tuberosum genotype RH89-039-16 in response to chitosan. *Frontiers in Plant Science*, *11*. doi: 10.3389/fpls.2020.01193

Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., y Berndt, S. I. (2008, apr). Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature Genetics*, *40*(5), 584–591. doi: 10.1038/ng.125

Li, F.-W., y Alex, H. (2018, mar). A guide to sequence your favorite plant genomes. *Applications in Plant Sciences*, *6*(3), e1030. doi: 10.1002/aps3.1030

Li, G., Pan, T., Guo, D., y Li, L.-C. (2014). Regulatory variants and disease: the e-cadherin -160c/a SNP as an example. *Molecular Biology International*, *2014*. doi: 10.1155/2014/967565

Li, H. (2011, sep). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., y Durbin, R. (2009, may). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Liu, B., Kong, L., Zhang, Y., y Liao, Y. (2021, jan). Gene and metabolite integration analysis through transcriptome and metabolome brings new insight into heat stress

tolerance in potato (solanum tuberosum l.). *Plants*, *10*(1), 103. doi: 10.3390/plants10010103

Love, M. I., Huber, W., y Anders, S. (2014, dec). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12). doi: 10.1186/s13059-014-0550-8

Madroñero, J., Corredor Rozo, Z. L., Escobar Pérez, J. A., y Velandia Romero, M. L. (2019). Next generation sequencing and proteomics in plant virology: how is Colombia doing? *Acta Biológica Colombiana*, *24*(3), 423–438. doi: 10.15446/abc.v24n3.79486

Martin, M. (2011, may). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10. doi: 10.14806/ej.17.1.200

Massa, A. N., Childs, K. L., Lin, H., Bryan, G. J., Giuliano, G., y Buell, C. R. (2011, oct). The transcriptome of the reference potato genome solanum tuberosum group phureja clone DM1-3 516r44. *PLoS ONE*, *6*(10), e26801. doi: 10.1371/journal.pone.0026801

McConeghy, B. W. (2019, Nov). *Setting up a bioinformatics QC pipeline.* Descargado de `https://youtu.be/lrs8WkVdIVg`

Michael Love, S. A. (2017). *Deseq2.* Bioconductor. doi: 10.18129/B9.BIOC.DESEQ2

National Human Genome Research Institute. (2019, 03). *Transcriptoma.* Descargado de `https://www.genome.gov/es/about-genomics/fact-sheets/Transcriptoma`

National Human Genome Research Institute. (2022). *Polimorfismo de nucleótido único (SNP).* Descargado de `https://www.genome.gov/es/genetics-glossary/Polimorfismos-de-nucleotido-%C3%BAnico`

Ñústez-López, C. (2011). *Variedades colombianas de papa.* Bogotá: Universidad Nacional de Colombia, Facultad de Agronoma, Red Latinpapa.

Ñústez-López, C., y Rodríguez-Molano, L. (2020). *Papa criolla (Solanum tuberosum Grupo Phureja): Manual de recomendaciones técnicas para su cultivo en el departamento de Cundinamarca.* Corredor Tecnológico Agroindustrial CTA-2. Descargado de `http://investigacion.bogota.unal.edu.co/visibilidad/publicaciones/manuales-dervivado-2/papa-criolla-solanum-tuberosum-grupo-phureja-manual-de-recomendaciones-tecnicas-para-su-cultivo-en-el-departamento`

-de-cundinamarca/

Parra Pérez, E. M., y Ortíz Arévalo, I. L. (2017, Sep). *Identificación de CYP707A1 en la Síntesis de ABA involucrada en la respuesta por estrés hídrico en un cultivo de papa criolla (Solanum tuberosum grupo Phureja) irradiada con cobalto 60 ubicado en el municipio El Rosal, Finca El Pino Km 16 vía Subachoque, Cundinamarca* (Inf. Téc.). Descargado de http://hdl.handle.net/11349/6583

Petek, M., Zagorščak, M., Ramšak, Ž., Sanders, S., Tomaž, Š., Tseng, E., ... Gruden, K. (2020, jul). Cultivar-specific transcriptome and pan-transcriptome reconstruction of tetraploid potato. *Scientific Data*, *7*(1). doi: 10.1038/s41597-020-00581-4

Pillai, R. (2005, dec). MicroRNA function: Multiple mechanisms for a tiny RNA? *RNA*, *11*(12), 1753–1761. doi: 10.1261/rna.2248605

Pinzón, A., Barreto, E., Bernal, A., Achenie, L., González Barrios, A. F., Isea, R., y Restrepo, S. (2009). Computational models in plant-pathogen interactions: the case of Phytophthora infestans. *Theoretical Biology and Medical Modelling*, *6*(1). doi: 10.1186/1742-4682-6-24

Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E., y Zanini, F. (2022, mar). Analysing high-throughput sequencing data in python with HTSeq 2.0. *Bioinformatics*, *38*(10), 2943–2945. doi: 10.1093/bioinformatics/btac166

Ramamoorthy, C. V., y Li, H. F. (1977, mar). Pipeline architecture. *ACM Computing Surveys*, *9*(1), 61–102. doi: 10.1145/356683.356687

Riascos Chica, M., Gutiérrez Sánchez, P. A., y Marín Montoya, M. A. (2018). Identificación molecular de Potyvirus infectando cultivos de papa en el oriente de Antioquia (Colombia). *Acta Biológica Colombiana*, *23*(1), 39–50. doi: 10.15446/abc.v23n1.65683

Rodriguez, L. E., y Moreno, P. (2010). Agronomía Colombiana. *Factores y mecanismos relacionados con la dormancia en tubérculos de papa. Una revisión*, *28*(2), 189–197. Descargado de https://revistas.unal.edu.co/index.php/agrocol/article/view/18022

Rodríguez, L. E., Ñustez, C. E., y Estrada, N. (2009). Criolla Latina, Criolla Paisa y Criolla Colombia, nuevos cultivares de papa criolla para el departamento de An-

tioquia (Colombia). *Agronomía Colombiana*, *27*(3), 289–303. Descargado de https://revistas.unal.edu.co/index.php/agrocol/article/view/13204

Roy, S., Coldren, C., Karunamurthy, A., Kip, N. S., Klee, E. W., Lincoln, S. E., ... Carter, A. B. (2018, jan). Standards and guidelines for validating next-generation sequencing bioinformatics pipelines. *The Journal of Molecular Diagnostics*, *20*(1), 4–27. doi: 10.1016/j.jmoldx.2017.11.003

Shumway, M., Cochrane, G., y Sugawara, H. (2009, dec). Archiving next generation sequencing data. *Nucleic Acids Research*, *38*(suppl_1), D870–D871. doi: 10.1093/nar/gkp1078

Smith, B. (2016, 03). *Reproducible bioinformatics pipelines using Make.* Descargado de http://byronjsmith.com/make-bml/

The state of food security and nutrition in the world 2020. (2020). FAO, IFAD, UNICEF, WFP and WHO. doi: 10.4060/ca9692en

Sánchez Álvarez, E. L., y Baracaldo Pinto, D. M. (2019, Oct). *Análisis de la expresión del Gen CDPK7 y evaluación del Gen EIN2 en papa criolla Solanum Tuberosum Vf. Phureja, (Variedad Criolla Colombiana) Irradiada con Cobalto 60* (Inf. Téc.). Descargado de http://hdl.handle.net/11349/23235

Taiz, L., y Zeiger, E. (2010). *Plant physiology 5th edition sinauer associates.* Inc. Publishunder land Massac husetts.

Tange, O. (2011, Feb). GNU parallel - the command-line power tool. *The USENIX Magazine*, *36*(1), 42-47. doi: 10.5281/ZENODO.16303

The Potato Genome Sequencing Consortium. (2011, jul). Genome sequence and analysis of the tuber crop potato. *Nature*, *475*(7355), 189–195. doi: 10.1038/nature10158

Tinjacá, S., y Rodríguez, L. (2015). *Catálogo de papas nativas de nariño - colombia.* Bogotá: Mc Gill.

Tiwari, J. K., Buckseth, T., Singh, R. K., Zinta, R., Saraswati, A., Kumar, M., y Chakrabarti, S. K. (2021, jan). Methylome and transcriptome analysis reveals candidate genes for tuber shape variation in tissue culture-derived potato. *Plant Growth Regulation*, *93*(3), 319–332. doi: 10.1007/s10725-020-00690-5

Tiwari, J. K., Buckseth, T., Zinta, R., Saraswati, A., Singh, R. K., Rawat, S., ... Chakra-

barti, S. K. (2020, jan). Transcriptome analysis of potato shoots, roots and stolons under nitrogen stress. *Scientific Reports*, *10*(1). doi: 10.1038/s41598-020-58167-4

Universidad Nacional de Colombia. (2014, 09). *Dormancia.* Descargado de https://web.archive.org/web/20140910200052/http://www.virtual.unal.edu.co/cursos/ciencias/2000024/lecciones/cap02/02_04_15.htm

Valencia, R. A., Lobo Arias, M., y Ligarreto, G. A. (2010, jun.). Estado del arte de los recursos genéticos vegetales en Colombia: Sistema de Bancos de Germoplasma. *Ciencia & Tecnología Agropecuaria*, *11*(1), 85–94. Descargado de http://revistacta.agrosavia.co/index.php/revista/article/view/198 doi: 10.21930/rcta.vol11_num1_art:198

Veen, A. H. (1986, dec). Dataflow machine architecture. *ACM Computing Surveys*, *18*(4), 365–396. doi: 10.1145/27633.28055

Veramendi, S., y Gabriel, J. (2015). *Selección asistida por marcadores moleculares para resistencia a enfermedades en un programa práctico de mejoramiento genético de papa* (Inf. Téc.). Descargado de https://www.fontagro.org/wp-content/uploads/PROINPA_compendio_2011-2014_ProyectoPAG20.pdf

Virus del amarillamiento en papa: Una amenaza que se puede controlar. (2013). Descargado de https://www.ica.gov.co/periodico-virtual/prensa/2013-(2)/virus-del-amarillamiento-de-las-venas-de-la-papa,#:%7E:text=El%20amarillamiento%20de%20las%20venas,de%20vigor%20de%20la%20planta.

Visscher, P. M. (2008, may). Sizing up human height variation. *Nature Genetics*, *40*(5), 489–490. doi: 10.1038/ng0508-489

Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., ... Samani, N. J. (2008, apr). Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics*, *40*(5), 575–583. doi: 10.1038/ng.121

Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S., y Burbano, H. A. (2018). nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics*, *19*(1). doi: 10.1186/s12859-018-2128-z

Wratten, L., Wilm, A., y Göke, J. (2021, sep). Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods*, *18*(10),

1161–1168. doi: 10.1038/s41592-021-01254-9

Yalamanchili, H. K., Wan, Y.-W., y Liu, Z. (2017, sep). Data analysis pipeline for RNA-seq experiments: From differential expression to cryptic splicing. *Current Protocols in Bioinformatics*, *59*(1). doi: 10.1002/cpbi.33

Yizhen, L., Lin, L., y Jun, W. (2011, aug). The application of pipeline technology: An overview. En *2011 6th international conference on computer science & education (ICCSE)*. IEEE. doi: 10.1109/iccse.2011.6028582

Zabala Pardo, D. M. (2015, Mar). *Efecto de diferentes dosis de radiación gamma (?) sobre la expresión fenotípica en papa criolla (solanum tuberosum grupo phureja, variedad criolla Colombia)* (Inf. Téc.). Descargado de `http://hdl.handle.net/11349/863`

Zhang, R., Marshall, D., Bryan, G. J., y Hornyik, C. (2013, feb). Identification and characterization of miRNA transcriptome in potato by high-throughput sequencing. *PLoS ONE*, *8*(2), e57233. doi: 10.1371/journal.pone.0057233