ANÁLISIS DE GENOTIPIFICACIÓN POR SECUENCIA (GBS) EN UNA POBLACION DE *Solanum tuberosum* GRUPO PHUREJA (VARIEDAD CRIOLLA COLOMBIA).

DANIEL ALEJANDRO CRUZ CRUZ

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS
FACULTAD DE CIENCIAS Y EDUCACIÓN
PROYECTO CURRICULAR DE LICENCIATURA EN QUÍMICA
BOGOTÁ D.C.

ANÁLISIS DE GENOTIPIFICACIÓN POR SECUENCIA (GBS) EN UNA POBLACION DE *Solanum tuberosum* GRUPO PHUREJA (VARIEDAD CRIOLLA COLOMBIA).

DANIEL ALEJANDRO CRUZ CRUZ

Trabajo de grado para optar por el titulo de

LICENCIADO EN QUIMICA

DIRECTOR

LUIS FRANCISCO BECERRA GALINDO

Docente del proyecto curricular de licenciatura en química

Doctor en ciencias biológicas

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

FACULTAD DE CIENCIAS Y EDUCACIÓN

PROYECTO CURRICULAR DE LICENCIATURA EN QUÍMICA

BOGOTÁ D.C.

Nota de Aceptación

_____

Jurado (Nombre/Firma)

_____

Director (Nombre/Firma)

_____

Bogotá, Febrero 2024

AGRADECIMIENTOS

Primero que todo, quiero dar mis mas profundos agradecimientos al profesor Francisco Becerra por depositar en mi la confianza y concederme la oportunidad para realizar una investigación como esta, por confiar en mis capacidades, y sobre todo por hacer posible que el trabajo fuera finalizado y entregado en los tiempos establecidos.

A mi madre, por todo el apoyo y por estar presente día a día en la cotidianidad prestándome pequeñas ayudas en las actividades cotidianas que hicieron en todo este proceso una ayuda enorme para poder dedicar la mayoría de mi tiempo y energía en este proyecto. A mis hermanos quienes fueron parte fundamental de este proceso prestándome ayuda y algunas soluciones en momentos donde muchas veces la necesite; a mis tías y a mi familia en general por ser un apoyo incondicional tanto para mí como para mi madre en momentos complicados durante el proceso.

Un agradecimiento muy especial y muy importante para el licenciado Alexis Quintero López, por toda su ayuda y sus recomendaciones prestadas durante el proyecto, orientaciones, consejos y tiempo dedicado a esta investigación, gracias a las indicaciones otorgadas, a las tutorías a las que dedico mucho de su tiempo, a la resolución de dudas y a todo el acompañamiento que realizó, por todo esto fue posible para mi terminar el proyecto y aprender demasiado durante el proceso en un campo tan desconocido para mi como lo era la bioinformática y por hacer todo lo que era posible para terminar el proyecto, sin él nada de esto hubiera sido posible.

A mi mejor amiga Paola Zapata, con quien comparto este amor y esta vocación por la ciencia por prestarme su ayuda y sus contactos para solucionar los problemas presentados durante este proceso; a mis amigas Vanessa Rivera y Daniela Ramírez por ser las personas maravillosas con quienes compartí mi camino durante la carrera y mi paso por la Universidad, con quienes comparto vocación, y quienes fueron un apoyo incondicional y muy grande no solo durante el proceso de investigación sino durante todo la carrera cursada con sus palabras de aliento y de motivación, y demás compañeros y compañeras, personas con las que compartí una pequeña parte de mi vida y no solo estuvieron presentes en mi proceso de formación académica sino también fueron parte fundamental en mi desarrollo personal y humano.

Y finalmente uno de los agradecimientos mas importantes es para mí, por la constancia, la resiliencia, la determinación, los sacrificios, el sentido de responsabilidad, la disciplina y la

persistencia durante lo difícil que fue el proceso, este proyecto no solo me dio un gran aprendizaje académico y disciplinar sino también desarrollo y forjo mi desarrollo personal cualidades que no sabia que podía desarrollar como la confianza en mí mismo, la paciencia y la determinación, infinitas gracias.

CONTENIDO

LISTADO DE ILUSTRACIONES 8
LISTADO DE TABLAS 9
RESUMEN 10
PLANTEAMIENTO DEL PROBLEMA 12
HIPÓTESIS 14
OBJETIVOS 17
ESTADO DEL ARTE 18
ANTECEDENTES 18
MARCO TEÓRICO 22
PAPA CRIOLLA (*Solanum phureja*) 22
TAXONOMIA 23
CARACTERÍSTICAS GENÉTICAS 24
DORMANCIA 25
CARACTERÍSTICAS DE LA DORMANCIA. 26
MEJORAMIENTO GENETICO DE *Solanum phureja* EN COLOMBIA 28
GWAS (ESTUDIO DE ASOCIACION DEL GENOMA COMPLETO) 29
GENOTIPADO POR SECUENCIA (GBS) 30
POLIMORFISMO DE NUCLEOTIDO UNICO (SNP) 31
DESEQUILIBRIO DE LIGAMIENTO (LD) 32
METODOLOGÍA 34
BUSQUEDA DE DATOS GBS 34
CONTROL DE CALIDAD EN LAS SECUENCIAS 35
ALINEACION DE SECUENCIAS Y SELECCIÓN DEL GENOMA DE REFERENCIA 36
ORDENAMIENTO DE LAS ALINEACIONES DE CADA UNA DE LAS MUESTRAS 36
LLAMADA DE VARIANTES EN MUESTRAS INDIVIDUALES 37
FUSION DE MUESTRAS INDIVIDUALES Y GENOTIPADO POR MUESTRA 37
FUSION DE VARIANTES GENOTIPIFICADAS 38
ANOTACIÓN FUNCIONAL DE LAS VARIANTES 38
CONSTRUCCIÓN DE LA MATRIZ DE SNP 39
ANÁLISIS ESTADISTICO 39
RESULTADOS Y ANÁLISIS 41

DESCARGA DE DATOS GBS DE *Solanum phureja* 41
CONTROL DE CALIDAD EN LAS SECUENCIAS 42
ALINEACION DE LAS SECUENCIAS CON EL GENOMA DE REFERENCIA 46
LLAMADA DE VARIANTES, FUSION DE ARCHIVOS VCF Y GENOTIPADO 47
ANOTACION DE VARIANTES, MATRIZ SNP Y ANALISIS ESTADISTICO 53
CONCLUSIONES 56
RECOMENDACIONES 58
BIBLIOGRAFÍA 59
ANEXOS: 63
BITACORA DE COMANDOS: 63
1) Pretratamiento de datos: 63
2)Alineacion de secuencias 64
3)Organizacion de los alineamientos: 64
4)Conversion de formato SAM a formato BAM: 65
5)Lamada de varientes individual: 65
6) Indexacion con samtol de genoma de referencia: 65
8)Genotipado de variantes para cada una de las muestras: 66
9)Variantes fusionadas con informacion de genotipo: 67
10) Matriz con vcftools y plink 67
11)Analisis estadistico: 67
INFORMES DE CALIDAD: 69
2.1) INFORMES DE CALIDAD EN BRUTO 69
SRX8683637 69
SRX8683638 70
SRX8683639 71
SRX8683640 72
SRX8683641 73
SRX8683642. 74
2.2) INFORMES DE CALIDAD DESPUES DEL PROCESO DE TRIMMING 75
SRX8683637 75
SRX8683638 76
SRX8683639 77
SRX8683640 78

SRX8683641 79
SRX8683642 80

LISTADO DE ILUSTRACIONES

Ilustración 1 Resumen de estadísticas básicas y de la calidad por secuencia de la biblioteca SRX8683637.fastq sin proceso de trimming. 42
Ilustración 2 Resumen de estadísticas básicas y de la calidad por secuencia de la biblioteca SRX8683637_cleanpn.fastq recortada con nanofilt,Porechop y cutadapt. 43
Ilustración 3. Diferencia de las secuencias totales entre las muestras antes y después del proceso de trimming. 44
Ilustración 4 Grafica de contenido del adaptador de la biblioteca SRX8683637.fastq sin proceso de trimming. 45
Ilustración 5 Grafica de contenido del adaptador de la biblioteca SRX8683637_cleanpn.fastq después del proceso de trimming con cutadap 45
Ilustración 6 Visualización de alineamientos apilados y ordenados de las 6 bibliotecas analizadas, en un archivo BAM combinado con las 6 muestras. 46
Ilustración 7 Archivo.log con la información contenida durante el proceso de anotación funcional de variantes en archivos vcf fusionados 54

LISTADO DE TABLAS

Tabla 1 Taxonomía y clasificación de Solanum phureja. 24
Tabla 2 Informacion del conjunto de datos de la investigación Whole genome and transcriptome sequencing of S. tuberosum Group Phureja DM 1-3 516 R44. 41
Tabla 3 Estadísticas resumidas del proceso de genotipado en las 6 bibliotecas genómicas analizadas. 48
Tabla 4 Distribución de genotipos en cada una de las bibliotecas analizadas por categorías. 50

RESUMEN

La papa criolla *(Solanum phureja*) desempeña un papel vital en la alimentación y la economía de Colombia, siendo una de las variedades más emblemáticas y consumidas en el país. Su importancia radica en varios aspectos; es un elemento fundamental en la cocina colombiana, contribuyendo a la identidad cultural y gastronómica del país. Se utiliza en la preparación de platos tradicionales como el ajiaco y la bandeja paisa, entre otros, es rica en nutrientes esenciales como carbohidratos, vitaminas y minerales, su consumo aporta energía y elementos esenciales para una dieta equilibrada. Su cultivo diversifica las fuentes de alimentación y proporciona una fuente accesible de nutrientes, además, representa una parte importante de la diversidad agrícola colombiana y contribuye a la preservación de variedades y a la adaptación a diferentes condiciones climáticas y de suelo, dentro de sus principales características se encuentran la gran adaptabilidad a días cortos y brotación al momento de la cosecha, es decir ausencia del periodo de dormancia.

La ausencia de dormancia y el brote y crecimiento de los tubérculos en la papa pueden estar relacionados con la intervención de varios procesos bioquímicos que resultan ser nocivos para su calidad nutricional y procesamiento, razón por la cual esta especie es objeto de estudio y candidato a procesos de fitomejoramiento genético. El fitomejoramiento genético es todo un proceso complejo en el que implica la selección de características deseables y cruzamiento de plantas con características deseables para generar variedades mejoradas, entre otros; identificar rasgos específicos que se desean mejorar es crucial pues el uso de tecnologías genómicas y moleculares para identificar y seleccionar genes específicos asociados con rasgos deseables acelera el proceso de mejora y permite una mayor precisión en la selección.

Dentro de estas tecnologías genómicas se encuentra la genotipificación por secuencia (GBS), un método de secuenciación genómica a gran escala y bajo costo, es por esto que en la presente investigación se realizó un análisis de genotipificación por secuencia de 6 accesiones de *Solanum tuberosum gurpo phureja* con el fin de identificar polimorfismos de nucleótido único (SNP) de nucleótido cerca de regiones codificantes y cerca a regiones de interés, para lo cual se realizó un protocolo para un estudio GWAS usando datos GBS, que consistió en un

alineamiento de las secuencias con el genoma de referencia actualizado DM_1-3_516_R44_potato_genome_assembly.v6.1, llamado de variantes, genotipado y anotación funcional de variantes con NGSEPcore, para generar la matriz se planeó utilizar vcftools y PLINK para finalmente realizar el análisis de asociación genética con la herramienta de Rstudio GAPIT usando el modelo lineal mixto para evitar falsos positivos. Los resultados muestran una limitación en el proceso de anotación funcional de variantes, limitación atribuida principalmente al archivo de anotación con los modelos genéticos para *Solanum tuberosum grupo phureja* y a la naturaleza de los datos y su calidad previa al momento de la secuenciación.

PLANTEAMIENTO DEL PROBLEMA

Solanum tuberosum grupo Phureja es una variedad de papa conocida comúnmente como papa criolla colombiana; la importancia a nivel comercial es bastante grande tanto a nivel nacional como a nivel internacional, pues se le atribuyen propiedades organolépticas particulares y distintivos de la cocina colombiana tradicional, además de contar con un valor nutricional superior a la papa tradicional o papa de año. (Ñústez y Rodríguez, 2021)

El departamento de Cundinamarca es el que presenta la mejor tecnología en cuanto a técnicas e instrumentos para la producción de la papa criolla, por lo que la mayor producción de este cultivo se encuentra en esta zona y sus alrededores como en Boyacá; las características principales de este tubérculo son la resistencia a las plagas y la ausencia en el periodo de dormancia. Esta última es la razón por la cual esto representa uno de los mayores problemas y limitantes para ser comercializada, generando hasta un 40% en perdidas postcosecha (Parra Galindo, 2019).

La dormancia en los tubérculos es definida como el periodo de tiempo que existe entre la cosecha y la aparición de brotes. Durante este periodo los tubérculos de la papa se encuentran inactivos y de forma latente, incluso, si las condiciones externas como lo son la temperatura, la humedad, la luz, son favorables para dicho crecimiento y brote. La ausencia de dormancia y el brote y crecimiento de los tubérculos en la papa está relacionado con la intervención de varios procesos bioquímicos que resultan ser nocivos para su calidad nutricional y procesamiento (Toro Pantoja, 2015).

La salida del periodo de dormancia de la papa está intrínsecamente relacionada con procesos bioquímicos y metabólicos como se había mencionado anteriormente. A pesar de que en este periodo el metabolismo es bajo, la actividad hormonal en el tubérculo tiene un efecto antagónico, evitando así la división celular y aumentando la respiración; sin embargo, a pesar de la estrecha relación que existe entre la disminución del ácido abscísico (ABA) y el aumento de producción de giberelinas (hormonas vegetales), sus funciones específicas aún son tema de estudio (Gong et al., 2021). Al igual que como ocurre con el ácido abscísico y con las giberelinas, los tubérculos de papa producen una cantidad determinada de etileno, dicha cantidad aumenta después de la cosecha y posteriormente disminuye nuevamente. La función fisiológica entre el etileno y el periodo de dormancia no es clara aun, pero bajo ciertas

condiciones y tratamientos continuos con etileno han logrado inhibir el periodo de brote en esta especie de tubérculo (Rodríguez y Moreno, 2010).

Caracterizar la diversidad génica del tubérculo es de gran importancia para poder analizar la compleja relación que existe entre la dormancia y los factores que pueden afectarla, tales como: rutas metabólicas (que implican la actividad enzimática y contenido hormonal) y la expresión de genes. Gracias a las diferentes técnicas de secuenciación del genoma completo es posible ensamblar genomas por separado para identificar variedades puntuales y específicas como lo son los SNP; estos ensamblajes pueden proporcionar información valiosa acerca de los genes candidatos para el estudio de la regulación y aumento en el periodo de dormancia como para los genes relacionados con la producción de la papa.

HIPÓTESIS

H_0= Hay presencia de polimorfismos de un solo nucleótido (SNPs) cerca a regiones genómicas de interés relacionadas con genes que modulan el periodo de dormancia.

H_1= No hay presencia de polimorfismos de un solo nucleótido (SNPs) cerca a regiones genómicas de interés relacionadas con genes que modulan el periodo de dormancia.

PREGUNTA DE INVESTIGACIÓN

¿Existen cambios polimórficos cerca de regiones genómicas relacionadas con la actividad de fitohormonas que influyan en el periodo de dormancia y en genes candidatos de estudio en la población de *Solanum tuberosum* grupo Phureja analizada?

OBJETIVOS

1. Objetivo general:

- Evaluar la diversidad genética en una población de *Solanum tuberosum* grupo Phureja mediante el estudio de asociación del genoma completo (GWAS) utilizando datos GBS para identificar variantes de polimorfismos a través del genoma.

2. Objetivos específicos:

- Caracterizar la diversa población de genes presentes en *Solanum tuberosum* grupo Phureja.
- Identificar variantes polimórficas cerca a regiones génicas relacionadas con la dormancia.

ESTADO DEL ARTE

ANTECEDENTES

La presente investigación toma como antecedentes estudios, artículos e investigaciones relacionadas con el estudio del genoma y genotipificación de la especie *Solanum tuberosum* centrándose en el grupo Phureja de esta especie. Dentro de estos se encuentran:

El consorcio de secuenciación del genoma de la papa (PGSC, por sus siglas en inglés The Potato Genome Sequencing Consortium) fue el primero en proporcionar el genoma de referencia en el año 2011. Utilizando el método de secuenciación de escopeta logran organizar con precisión alineando los segmentos de ADN secuenciados; se ensambló el 86% del genoma, identificando 39.031 genes que codifican para diferentes proteínas dentro del tubérculo (Xu et al., 2011). Los grupos investigadores que lideran y han obtenido avances potencialmente importantes en esta área provienen de Holanda y de países bajos; institutos como Wageningen University & Research (WUR) y el Instituto Holandés de Investigación Agrícola (Wageningen Plant Research) son reconocidos a nivel mundial por su destacada investigación en el cultivo de la papa (*Solanum tuberosum*) y el uso de tecnologías avanzadas en genética y genómica. En Colombia hay algunos estudios relacionados con la mejora genética de la papa con diferentes técnicas, incluyendo el GBS, usado principalmente en la búsqueda de genes que presenten resistencia a enfermedades, a la búsqueda de marcadores moleculares de interés específico y en el mejoramiento genético en la especie *Solanum tuberosum* L, pero hay muy pocos con relación a la especie *Solanum tuberosum* grupo Phureja.

En mayo del 2019, en la facultad de ciencias agrarias de la Universidad Nacional de Colombia, se presentó un estudio de asociación amplia del genoma, basándose en la identificación de polimorfismos de un solo nucleótido (SNP por sus siglas en inglés); donde se utilizó la técnica de genotipificación por secuencia (GBS) en la especie *Solanum tuberosum* grupo Phureja para identificar regiones genómicas asociadas con la resistencia al Tizón tardío o gota en papa como comúnmente se conoce, causada por el patógeno *Phytophthora infestans;* en la investigación se asoció en rasgo de la resistencia con los polimorfismos de un solo nucleótido distribuidos por todo el genoma; además, se identificaron 16 locus de rasgo cuantitativo (QTL por sus siglas en inglés Quantitative Trait Locus) que presentan una relación directa con la resistencia a la enfermedad de gota explicando del 13,7% al 50,9% la varianza fenotípica presentada en la

especie, lo que permite ver la complejidad genómica de la especie *S. phureja* (Juyo Rojas et al., 2019).

En diciembre del mismo año (2019) se realizó una segunda investigación en la misma facultad de la Universidad Nacional de Colombia. Se publicó un estudio de asociación de genoma completo (GWAS) en la especie *Solanum tuberosum* grupo Phureja con el fin de identificar el conjunto de genes que participan tanto de forma directa como de forma indirecta en la ruta biosintética de antioxidantes presentes en la papa, tal como lo son los polifenoles y compuestos antocianicos debido a su importancia nutricional como una de las fuentes principales de antioxidantes implementadas en la dieta humana (Galindo, 2019). Para determinar el contenido de compuestos antocianicos se utilizó cromatografía liquida de alta precisión (HPLC) y para el genotipado se utilizó la técnica de genotipificación por secuencia (GBS), donde se utilizaron distintos marcadores polimórficos y modelos de rutas biosintéticas con las regiones genómicas promotoras de cada gen relacionado con esta ruta. Los resultados obtenidos permitieron identificar una región de interés en el cromosoma 10, un locus de 4 Mb con genes asociados a la ruta biosintética de antioxidantes en la papa (Galindo, 2019). Esta investigación permite identificar nuevas regiones genómicas de interés para el estudio de las rutas biosintéticas y permitió dar un nuevo enfoque en investigaciones que se puedan realizar más adelante en programas de mejoramiento genético de la especie *S. phureja*.

El departamento de biología vegetal de la Universidad Estatal de Michigan en septiembre de 2020 publicó un artículo titulado “Construction of a chromosome-scale long-read reference genome assembly for potato”, el cual presenta una actualización del ensamblaje del genoma de referencia (DM1-3 516 R44 (v6.1) de la especie *S. phureja*, realizando un ensamblaje a nivel cromosómico. Este nuevo ensamblaje secuenció y ensambló 741,5 Mb, lo que representa el 87,8 % del genoma que se estimaba ensamblar, de los cuales 731,2 Mb son secuencias ancladas a los 12 cromosomas del clon monoploide secuenciado. Se encontró 32.917 genes que codifican para diferentes proteínas de alta confianza; estos genes compartían homología de secuencia con 44.851 modelos de genes que presentaban una importante y significativa mejora en comparación con los anteriores genes registrados en el primer genoma de referencia publicado. Además de esto el nuevo ensamblaje aumentó los contings para alargar el tamaño del andamiaje del genoma e identificar la ubicación de los centrómeros en los cromosomas de esta especie, lo que le proporciona la mejora de la calidad de este en comparación con el primer genoma de referencia; la nueva construcción y actualización del genoma de referencia de la

especie *Solanum tuberosum* grupo Phureja en investigaciones como estas resulta de gran importancia para estudios a futuro ya que brindan información más detallada de la evolución y los cambios genéticos en la especie. Esta información también puede ser utilizada en futuros estudios de genotipado y en la mejora de rasgos agronómicos para incrementar su productividad (G. M. Pham et al., 2020).

Una pequeña revisión publicada en la revista cientifica Plant Biotechnology Reports en el año 2021 presenta a los genes StZEP y StNCED, encargados de la producción de algunas fitohormonas como ABA y StCYP y StSUT4, como genes candidatos responsables de la salida de la dormancia del tubérculo. De igual manera, los genes ARF1, STIAA y STGA20x son candidatos causantes de los brotes del tubérculo de la especie *S. phureja*; mediante la información obtenida en el sumario se concluyó que el proceso de salida de la latencia y el brote del tubérculo está regulado por múltiples genes que actúan de manera organizada para dar paso a este proceso (Gong et al., 2021). La información del sumario suministrada es de gran utilidad para un estudio en el que se requiera identificar los diferentes genes candidatos para mejoras genéticas.

Finalmente, en agosto de 2022, el Departamento de Ciencias de las Plantas, Universidad McGill, Sainte-Anne-de-Bellevue, QC, Canadá en colaboración con el Centro de Investigación y Desarrollo de Fredericton, Agricultura y Agroalimentación de Canadá publican un estudio titulado “Genome sequencing of adapted diploid potato clones”, en el que se realizó el ensamblaje de la secuencia del genoma de la papa diploide de nueve clones, donde los tamaños del genoma oscilaron entre 712 Mb y 948 Mb. Se realizó una comparación con el genoma de referencia seleccionado (DM1-3 516 R44 (v6.1), genoma de referencia actualizado (mencionado anteriormente) presentando una heterocigosidad fluctuante entre 1% y 9, 04%. Esto se realizó con el fin de encontrar alelos perniciosos, sin embargo, en el estudio se detectó una variación del gen StCDF1, asociado a la prematura tuberización, no obstante, estas variantes le confirieron al tubérculo días más largos (Achakkagari et al., 2022). En la literatura se registra que la sobreexpresión de las variantes del gen se ha asociado con una mayor duración de la dormancia en *Solanum tuberosum phureja*.

MARCO TEÓRICO

PAPA CRIOLLA (*Solanum phureja*)

La papa criolla es uno de los alimentos con mayor importancia a nivel mundial; es el cuarto producto agrícola con mayor producción del mundo, después del maíz, el arroz y el trigo. De igual manera, ocupa el primer lugar como producto agrícola no cereal en el mundo. En Colombia es considera uno de los alimentos fundamentales y básicos en los platos típicos de las diferentes regiones del país debido a sus fortalezas como lo son sus propiedades organolépticas, su alto valor nutricional, su tiempo de cosecha corto (a diferencia de *Solanum tuberosum* L., papa de año), su inferior costo de producción, y su alto estatus como producto de exportación exótico, razón por la cual este cultivo ocupa 170.000 hectáreas del terreno colombiano para su siembra y su posterior cosecha (Ligarreto & Suárez, 2003).

Colombia es uno de los principales países de Latinoamérica en producir papa criolla. El clima frio de sus diferentes zonas como lo son Antioquia, Cundinamarca y Nariño es el adecuado para el desarrollo del tubérculo, teniendo una adaptabilidad de un rango de altura entre los 2.600-3.000 msnm. Además, el país también es el principal exportador y abastecedor de papa criolla a nivel mundial, pues sus principales exportaciones son realizadas a países europeos, Estados Unidos y Japón, puesto que la industria internacional se ha expandido y globalizado ampliamente para el procesamiento y distribución principalmente de snacks y productos fritos a base de papa en las presentaciones de papa precocida congelada y en casquitos, encurtido, pues se caracteriza por presentar agradable sabor y textura de fácil preparación y en un segundo plano de conservas y enlatados del tubérculo (Socorro et al., 2022).

Sus características principales aparentemente visibles y su morfología se describen como tubérculos redondos con carne y piel de color amarillo de fenotipo comúnmente conocido en la comunidad científica como “yema de huevo”. A diferencia de la especie originaria predominante *Solanum tuberosum* L., especie tetraploide, *Solanum phureja* es una especie nativa diploide (*2*n = 2X = 24 cromosomas) que tiene gran adaptabilidad a días cortos y brotación al momento de la cosecha. De la misma forma, el cultivo de papa criolla también presenta un alto grado de mezcla de variedades nativas ya que el cultivo crece extensamente en los andes, desde el occidente de Venezuela, hasta el centro de Bolivia, por lo que, en términos de genética, el cultivo presenta una alta heterocigocidad, además de ser propenso a

tener una alta perecibilidad, que lo hace foco y objetivo de programas de mejoramiento genético de rasgos agronómicos (Ñústez López & Rodríguez Molano, 2020).

Su composición nutricional es prominente y destaca principalmente por esta razón, pues presenta un elevado índice de antioxidantes y un elevado índice de fibra, almidón y proteína en comparación con la especie de origen, además de contener cantidades más altas de fibra cruda, sodio y potasio, tal como lo describe Penagos y cito textualmente: "El almidón, el principal representante de los carbohidratos digeribles, está presente en un rango ligeramente mayor en los tubérculos de papa del grupo Phureja (8,7 - 22,7 g/100 g BH) que en el Grupo Tuberosum (9,1 - 19,5 g/100 g BH), a pesar de que el contenido de proteína puede ser menor (grupo Phureja (1,7 - 2,8 g/100 g BH) - Grupo Tuberosum (1,7 - 4,2 g/100 g BH)), clones que fueron sometidos a mejoramiento genético de *Solanum phureja* presentan índices de proteína mucho más altos en comparación con cultivares comerciales cuando son sometidos a ebullición; el grupo Phureja puede proporcionar un alto contenido de potasio [563 mg/100 g BH (Ritter et al., 2008)] y sodio (18 mg/100 g BH) en comparación con el Grupo Tuberosum (5,3 - 651,9 mg/100 g BH y 0,1 - 6,7 mg/100 g BH respectivamente). de igual manera otros macronutrientes como Hierro, fosforo y calcio no varias significativamente entre especies" (Penagos Beltrán, 2020).

TAXONOMIA

La papa diploide es la especie de papa silvestre que se ha cultivado a lo largo de los andes, y es una especie que pertenece al grupo de la familia de las Solanaceae, y se distribuyen de la siguiente forma; tal como indica Bonierbale: La papa cultivada de la zona andina pertenece a un conjunto de especies diploides (*S. ajanhuiri, S. phureja, S. stenotomun*), triploides (*S. chaucha, S. juzepzuckii*), tetraploides (*S. tuberosum L subsp andigena Hawkes*) y pentaploides (*S. curtilobum*), mientras que la papa cultivada en la región litoral del sur de Chile pertenece a la especie tetraploide *S. tuberosum* L subsp. *tuberosum* (Bonierbale et al., 2004)

La taxonomía de las diferentes especies del tubérculo sigue siendo foco de debate, sin embargo, de los 8 grupos propuestos y mencionados anteriormente por sus características morfológicas, actualmente se toman en cuenta 4 grupos dentro de los cuales se encuentra el grupo Phureja en la cual Spooner y su equipo en 2007 eliminaron las anteriores clasificaciones (Barrera Alvarado & Avellaneda Ramírez, 2016).

GÉNERO	Solanum Linneus
SUBGÉNERO	Potatoe (G. Don) Arcy
SECCIÓN	PetotaDumortier
SUBSECCIÓN	Potatoe G. Don
SUPERSERIE	RotataHawkes (corola rotácea)
SERIE	XVI Tuberosa (Rybd.) Hawkes

Tabla 1 Taxonomía y clasificación de Solanum phureja.

Fuente: Huamán y Spooner (2002)

CARACTERÍSTICAS GENÉTICAS

La domesticación de la papa es el resultado de un cruce entre varias y diferentes especies de papa silvestre, tal y como lo explica Sponner en 2001:

Las primeras papas domesticadas pertenecieron a la especie diploide *Solanum stenotomum*, la cual derivó de *Solanum leptophyes.* A su vez, *S. stenotomum* probablemente se cruzó con *Solanum sparsipilum*, una especie silvestre, diploide seguida de una duplicación cromosómica lo cual daría origen a la especie cultivada tetraploide *Solanum tuberosum* subsp. *Andigena* (Spooner & Hijmans, 2001)**.**

Las diferentes especies de papa surgen a partir de la poliploidización sexual que sufrió la especie del grupo *Stenotomum*, a partir del cruce genético con otras especies y diferentes factores ambientales y epigenéticos generando una adaptación a los diferentes entornos donde se cultivaba, claramente la especie *S. phureja*, surge a raíz de estos cambios y por algunas mutaciones, en la distribución a lo largo de la zona Andina, desde el norte de Bolivia hasta el suroeste venezolano; la gran diversidad que existe en este cultivo es lo que dificulta el proceso de mejoramiento genético del tubérculo, dado que al haber sido una especie que presento demasiados cruces, presentar una alta heterocigosidad y gran poliploidía se dificulta aplicar técnicas de mejoramiento genético convencionales, pues esto complica el proceso para obtener conocimiento completo de la base genética de muchos de los genes que interfieren en procesos metabólicos, hormonales y de desarrollo del tubérculo como la salida de la dormancia y tuberización precoz que sufre la papa criolla. Sin embargo, se han desarrollado diferentes técnicas de mapeo genético que permiten el estudio de especies con alto grado de complejidad

utilizando marcadores moleculares que permiten identificar regiones genéticas de interés y algunos QTL para el mejoramiento genético de especies agronómicas con rasgos genéticos y poliploidía compleja (Martinez Corcino, 2012).

DORMANCIA

La dormancia o también conocida como latencia es uno de los fenómenos más estudiados en el ámbito del mejoramiento genético de la agronomía, y podría definirse como la fase o el periodo de tiempo en que una semilla detiene su desarrollo y crecimiento, es decir, se ve pausado o estancado, aun estando en un ambiente favorable para su crecimiento; este mecanismo se ha desarrollado debido a la adaptabilidad que ha tenido la semilla con el entorno, y que, no solo debe estar relacionado con la ausencia de germinación o de brotes visibles, sino también debe ser tomada en cuenta como una cualidad propia de la semilla en la que biológica y ambientalmente determina las condiciones requeridas para germinar y establecer una nueva generación de plantas de su especie (Finch-Savage & Leubner-Metzger, 2006).

La latencia también es una respuesta fisiológica ante el peligro que puede suponer la especie en el entorno, reconocido también como mecanismo de protección, evita la germinación temprana para asegurar la supervivencia de las plantas en diferentes condiciones de estrés, ya sea biótico o abiótico, no hay mucha información sobre como ocurre este proceso, sin embargo, a nivel molecular la semilla no se transporta a ningún nutriente almacenado y no hay proceso de división celular; por otro lado, los genes encargados de la germinación y el crecimiento no se expresan de manera activa durante este periodo, no obstante, la estructura de la cromatina es la encargada de regular expresión de los genes, razón por la cual en los últimos años se han estudiado los genes encargados de la modificación de la estructura de la cromatina, esto y otros factores como algunas fitohormonas en la semilla (ABA, ácido abscísico), también están involucrados en la transformación de la cromatina, de manera que se puede decir que lo genes relacionados con la regulación epigenética de las semillas son indispensables en el análisis de la germinación en esta misma. (Shu et al., 2016)

La dormancia en la especie *Solanum tuberosum* L. no está específicamente definida, sin embargo, es importante resaltar que son demasiados procesos dentro de los cuales se encuentra la actividad y expresión de los genes involucrados en el proceso de la salida del periodo de dormancia y la concentración de fitohormonas, no obstante, el estudio genético de la especie

Solanum tuberosum L. no se ha profundizado micho debido a la poliploidía que esta presenta (especie tetraploide, sin embargo, su relación con la salida de la latencia y la dormancia se ha estudiado con la producción de la fitohormonas asociadas a la salida de la latencia, las cuales son ácido abscísico (ABA), zeatina (ZT) y ácido giberélico (GA_3), que se producen a través de la vía del ácido mevalónico, concentraciones altas tanto de ABA como de GA_3 se reportan con bastante incremento justo después de la cosecha, luego las concentraciones bajan durante periodos de almacenamiento del tubérculo y nuevamente aumentan durante el crecimiento y desarrollo de los tubérculos (Haider et al., 2022), sin embargo, esto no sucede con el grupo Phureja de la especie, si bien es una especie que presenta también poliploidía (*Solanum phureja* es diploide*)*, no presenta una ploidía tan compleja, lo que permite analizar a la especie genéticamente en cuestión, diferenciando los genes responsables de las producción de las fitohormonas anteriormente mencionadas (ácido abscísico (ABA), zeatina (ZT) y ácido giberélico (GA_3)), y realizar una modificación que permita tener una latencia más prolongada sin recurrir a una latencia inducida.

CARACTERÍSTICAS DE LA DORMANCIA.

Durante el periodo de dormancia la semilla experimenta 2 fases que podría clasificarse como: latencia primaria y latencia secundaria; la latencia primaria es el estado postcosecha, donde la semilla permanece en un estado inactivo, producido por la acción y participación de ABA, una vez perdida la latencia primaria ya sea por condiciones ambientales o factores diferentes externos diferentes, la latencia secundaria puede ser inducida por condiciones ambientales no favorables para el desarrollo de la semilla, es decir ausencia de nutrientes y alteración del suelo, condiciones de luz y nitratos, cambios ambientales y estacionales hacen que la latencia se pueda perder y de igual manera inducir nuevamente, este proceso fisiológico desarrollado por la planta es un mecanismo de supervivencia que permite la germinación en condiciones óptimas y cuando estas entes presentes sin embargo, este proceso en la especie *Solanum phureja* es mucho más complejo de lo que parece. (Finch-Savage & Leubner-Metzger, 2006)

Según Lang en 1987, durante el desarrollo de los tubérculos presentan 3 etapas en el periodo de dormancia: paradormancia, endodormancia y ecodormancia. La paradormancia es el periodo en el cual se inhibe el crecimiento de las yemas de los tubérculos, está relacionado de forma más directa con la dormancia apical, la endodormancia es inducida debido al aumento de temperatura durante el verano, está relacionada con la acumulación de frio, por lo que

durante este periodo los meristemos están inactivos y los tubérculos no brotan, permanecen dentro de las yemas; durante esta etapa la semilla o tubérculo detiene su crecimiento incluso cuando las condiciones resulten favorables para su desarrollo; finalmente la ecodormancia está regulada por los factores ambientales relacionados con el calor acumulado necesario para que el tubérculo pueda generar brotes (Lang et al., 2022). Por lo mencionado anteriormente, la floración depende principalmente de los requerimientos de frio y calor necesario para la semilla, cuando se alcanzan los requerimientos ocurre el brote y el desarrollo de los tubérculos. (Prudencio et al., 2021)

Solanum phureja al igual que el resto de los tubérculos y papas, no sufren de deshidratación como acurren en el caso de las semillas de los cereales, por lo que necesitan pasar por un periodo de dormancia (así sea breve, como en el caso de esta especie) para poder iniciar el brote del tubérculo como tal, diferentes técnicas como control de latencia con inhibidores químicos, y control de latencia con disminución de la temperatura en frio han sido los más usados momentáneamente, sin embargo presentan problemas como el consumo de papa sin residuos químicos y el endulzamiento de la papa con la terapia en frio hacen de estas técnicas un proceso problemático y una solución no muy viable. (Naz et al., 2018)

La característica principal observada en los diferentes estudios realizados sobre la duración del periodo de latencia en los tubérculos es la actividad hormonal que este presenta, durante la latencia la actividad metabólica es baja, existe una disminución en la concentración de ABA y un aumento en las giberelinas, además de un aumento en la tasa de respiración.(Toro Pantoja, 2015),no obstante, Suttle afirma que pese a que la actividad hormonal en el tubérculo está estrechamente vinculada con la salida del estado de latencia del tubérculo, el proceso es más complejo de lo que se piensa y no es la única razón por la cual los tubérculos salen de este estado, por ejemplo pese a que la concentración de ABA disminuye progresivamente durante su almacenamiento, no existe una concentración preestablecida que prediga la salida del estado latente del tubérculo, por lo que esto indica la presencia de otros reguladores en el control e inactividad (Hartmann et al., 2011), como lo son factores genéticos.

La latencia de los tubérculos es genéticamente dependiente, Bisognin 2018, también plantea que las hormonas juegan un papel fundamental en este proceso, sin embargo, la relación que se presenta es entre algunos QTL presentes en los cromosomas de la papa, donde platea lo siguiente: El cromosoma 2 está relacionado con genes inducidos por ABA, mientras que en el

cromosoma 3 se correlaciona con genes de respuesta a ABA y genes de señalización, el cromosoma 7 se asocia con genes de señalización IAA, ABA y giberelinas (GA_3). Diferentes tipos de genes están conectados y relacionados con la biosíntesis de Etileno, ABA y giberelinas tales como StZEP, StNCED, StCYP, StSUT4, ARF1, STIAA y STGA20x por lo que la tasa de brotación está estrechamente relacionada con la herencia genética del tubérculo. (Gong et al., 2021)

MEJORAMIENTO GENETICO DE *Solanum phureja* EN COLOMBIA

La papa criolla representa el representa el 10% de la producción total de papa, y debió a las principales diferencias con la papa blanca que presenta (tiempo de cosecha de 120 días y día corto) y sus propiedades organolépticas su producción va creciendo poco a poco gracias a las condiciones climáticas que ofrece el terreno colombiano para su siembra (JUYÓ ROJAS, 2012). En los programas de mejoramiento genético agrícola el genotipo de *Solanum phureja* ha sido muy utilizada para la búsqueda e introgresión de genes importantes en la agronomía de la especie *Solanum tuberosum* L (Straadt & Rasmussen, 2003).

En 1934 se reporta el inicio de la colecta para conformar la Colección Central Colombiana de papa y en 1945 se registra el primer clon de papa de genotipo *Solanum phureja*, posteriormente en 1996 réplicas del genotipo de *Solanum phureja* fueron entregadas a la Universidad Nacional de Colombia como un recurso genético para hacer parte del programa de mejoramiento genético de la universidad; el mejoramiento genético ha sido fundamental y un punto clave para el desarrollo y comercialización de la papa criolla, pues ha desarrollado 5 cultivares comerciales dentro de los que se encuentra el grupo Phureja, el propósito del mejoramiento esta es aumentar el periodo de dormancia, generar un alto potencial de rendimiento y resistencia a algunas enfermedades como lo es el virus de la vena amarilla. (Rodríguez et al., 2009)

A partir del primer mapeo genético por asociación en papas tetrapolares realizado por Gebhardt en 2004, se ha ido implementando paulatinamente en los programas de mejoramiento genético debido a que permite el desarrollo de marcadores moleculares explorando la diversidad alélica a lo largo del genoma; los estudios de mapeo por asociación resultan ser de gran ayuda ya que las asociaciones entre marcadores de ADN y los rasgos agronómicos estudiados permiten validar genes que pueden ser posibles candidatos y responsables de los caracteres agronómicos de interés, facilitar el mapeo para reconocer QTL de alta resolución, permite realizar una

evolución genotípica previa a la evaluación fenotípica e identificar los alelos que presentan rasgos predominantes en el estudio de la población en cuestión. (Gebhardt et al., 2004)

GWAS (ESTUDIO DE ASOCIACION DEL GENOMA COMPLETO)

Los estudios de asociación del genoma completo o GWAS (por sus siglas en ingles Genome-wide association study) es un enfoque de investigación que consiste en el análisis de la variabilidad genética con el fin de identificar polimorfismos de nucleótido único SNP (single nucleotide Polymorphism) relacionados estadísticamente a un rasgo determinado, haciendo uso del desequilibro de ligamiento (LD) y de marcadores molecular (o también conocidos como QTL, loci de rasgos cuantitativos) para asociar algunos segmentos en cromosomas (genotipo) con la expresión de dicho rasgo determinado (fenotipo) (Hayes, 2013), es decir, se realiza una asociación de los alelos en diferentes segmentos de cromosoma o loci.

El enfoque de GWAS ha tenido un éxito en lo que respecta al mapeo de genes y de variantes genéticas que siguen la genética mendeliana, y en la identificación de mutaciones causadas por los polimorfismos de nucleótido único (SNPs), aun se discute su imparcialidad y eficacia, pues, aunque muestra buenos resultados, los GWAS se basan en LD entre SNP genotipados y variantes causales no genotipadas, es decir si no han sido genotipadas anteriormente es puede ser probable que no se detecten ciertos loci que pueden resultar importantes en estos estudios, por lo que las variables que se pueden detectar por este método son relativamente comunes en la población estudiada (Visscher et al., 2012) . Los estudios GWAS se acercan mucho más a una comprensión de rasgos complejos, en papa ya se ha realizado una caracterización donde se han reportado un aproximado de 47 QTL relacionados a la resistencia con el estrés por sequía (López Contreras, 2020). El genotipado para el GWAS se puede realizar por diferentes técnicas como lo son PCR, sondas ASO e hibridación en Micromatrices de ADN o esferas, secuenciación de ADN por diferentes métodos como lo es el método Sanger, la secuenciación paralela masiva o también se denomina secuenciación de próxima generación (NGS) y el GBS (genotipificación basado secuencia), siendo este último uno de los más utilizados en este enfoque investigativo. (Visscher et al., 2012)

GENOTIPADO POR SECUENCIA (GBS)

El genotipado por secuencia o genotipado basado en secuencia (por sus siglas en ingles genotyping by sequencing) es una técnica que busca la caracterización de variaciones genéticas presentadas en la especie estudiada en cuestión, fundamentado en una cantidad considerable de datos o cientos de marcadores moleculares utilizando la secuenciación masiva de ADN; la versatilidad y el bajo costo lo hacen una excelente técnica para el avance, descubrimiento e identificación de miles de polimorfismo de nucleótido único relacionados con rasgos de interés presentes en todo el genoma (Kumar et al., 2012). Las pipeline para el análisis avanzado de datos son un herramienta necesaria para limpiar lecturas, separar, analizar, clasificar y alinear los datos obtenidos a partir de la secuenciación masiva, y no es la excepción una pipeline para GBS, ya que la función de esta va encaminada a descartar lecturas de baja calidad separar las lecturas por individuos o grupos , alinear lecturas con el genoma de referencia del organismo estudiado para el descubrimiento de SNPs y la identificación de loci y alelos que pueden ser nuevos, es decir, el análisis de datos GBS se puede aplicar mediante dos enfoques: de novo-basado y basado en referencias. (Torkamaneh et al., 2016)

Al compara la técnica de GBS con otros métodos como con RAD -seq, GBS presenta una ventaja grande y una reducción en su complejidad al generar datos de SNPs de manera rápida, también presentan una metodología relativamente simple, pues el enfoque utiliza enzimas de restricción reducel falsos positivos en el momento de la secuenciación en regiones repetitivas, es decir reduce la complejidad del genoma facilitando el posterior análisis bioinformático además de reducir la complejidad en genomas muy grandes y extensos(Wong et al., 2015).El mayor desafío que presenta la técnica es la limitación de lectura producía por la baja profundidad en las lecturas en especies que presentan un elevado grado de heterocigocidad, debido a que no se detectan SNPs con precisión en el mapeo genético de especies con decendencia silvestre por la complejidad en la extensión del genoma y la diversidad génica que presenta, sin embargo hay reporte de estudios en la literatura a cerca del uso de GBS en cultivos domesticados con ancestros silvestres para construir la relación filogenética y la diversidad genética (Hamwieh et al., 2005)

Por lo mencionado anteriormente el método GBS resulta tener un gran potencial aplicable en estudios de mapeo genético y particularmente en la caracterización genómica en especies

altamente heterocigóticas con genomas extensos y complejos como lo es la especie *Solanum phureja*.

POLIMORFISMO DE NUCLEOTIDO UNICO (SNP)

La variación genómica presentada en una base en la cadena de ADN es lo que comúnmente se conoce como polimorfismo de nucleótido único o SNP (por sus siglas en inglés (Single Nucleotide Polymorphism), estas pequeñas diferencias es lo que hace que se presente gran diversidad fenotípica en diferentes especies, generalmente no tiene consecuencias biológicas, sin embargo usarlos como marcadores genéticos estos nos pueden proporcionar información sobre los patrones de herencia de algunos loci trasmitidos y conservados de generación en generación, además de servir como herramientas en el estudio de factores genéticos asociados a enfermedades, problemas metabólicos y hormonales; gracias al gran interés por el estudio de polimorfismos de nucleótido únicos se ha desarrollo áreas de detección muy amplias como la exploración de secuencias de ADN en busca de polimorfismos previamente desconocidos y la detección (genotipado) de individuos en busca de polimorfismos conocido. (Gao et al., 2005)

Los avances en las técnicas de detección de SNPs han permitido realizar posteriormente estudios de asociación del genoma completo (GWAS) a gran escala, este último enfoque ha hecho posible asociar los SNPs con diversas enfermedades y rasgos complejos expresados en diferentes especies (Zhu & Zhou, 2020). En la fase de descubrimiento después del enfoque GWAS, el análisis de SNPs se realiza de forma individual, lo que implica la expresión del fenotipo en cada SNP y la generación de un valor p como parámetro de referencia, de manera que los SNPs identificados se clasifican en función de este valor p y se establece un límite, donde valores inferiores a ese límite serán posteriormente validados y verificados, sin embargo, el análisis individual de SNPs puede resultar limitado por loque para evitar falsos positivos hay que ser riguroso en su estudio. (Wu et al., 2010).

En *Solanum phureja* hay reporte de diferentes estudios de detección de SNPs e identificación de genes relacionados con la resistencia el tizón tardío, en el que gracias al mapeo (GWAS) de asociaciones en germoplasma de *Solanum phureja*, que identificó entre 29 candidatos dos genes asociados con resistencia cuantitativa al tizón tardío.

DESEQUILIBRIO DE LIGAMIENTO (LD)

En los diferentes estudios de genética realizados, el LD o desequilibrio de ligamiento (por sus siglas en ingles Linkage Disequilibrium) se pude definir como la asociación no aleatoria de los alelos en diferentes loci establecido por diferentes factores tales como lo son la evolución y las mutaciones usado ampliamente por genetistas como un herramienta que observaron y que podían explotar una a su favor para el mapeo y asociación de genes, tal como Bodmer en 1986 explico que mediante el mapeo de asociación de poblaciones se podría llegar más fácil a entre la causa de una mutación y el marcador vinculado a esta región genómica. (Visscher et al., 2012)

La terminología resulta ser un poco problemática, pues LD no expresa la vinculación ni la falta de equilibro, simplemente el termino surgió en 1960 y fue usado por la mayoría de genetistas mientras su definición matemática fuera concreta y clara, de allí en más el termino perduro y simplemente referencia a loci estrechamente vinculados, sin embargo, en la década de 1980 su importancia comenzó a tomar relevancia como se había mencionado anteriormente, ya que proporciona información acerca de eventos pasados y restringe la respuesta a la selección tanto natural como artificial, reflejando toda la historia de la población a lo largo del genoma, de la misma forma, refleja la selección natural, las mutaciones, la conversión de genes y cada factor evolutivo en la frecuencia de los genes en cada región genómica analizada, tal que, los factores que afecten el LD de un loci en particular, depende únicamente de la tasa de recombinaciones locales, lo que la hace una teoría ampliamente desarrollada para proveer información acerca de la historia evolutiva de la especie estudiada en cuestión, y como herramienta fundamental en estudios GWAS y en el mapeo de genes tanto nuevos como ya regiones genómicas conocidas. (Slatkin, 2008)

El LD utilizado como herramienta para el descubrimiento de nuevos QTL’s es bastante popular dentro de los estudios de mapeo y asociación genética, consiste en determinar la disminución de la distancia de los SNP’s, es decir, la decadencia del LD con respecto a la distancia genética determinan tanto la resolución del mapeo genético realizado, como la cantidad de SNP utilizados en la matriz a emplear; datos reales registrados en estudios sobre mapeo genético en *Solanum tuberosum* L. demuestran como estimar e interpretar la descomposición del LD donde se pueda observar la cantidad de haplotipos en la población, es decir la variabilidad genética y el porcentaje de SNP’s específicos de cada haplotipo, aplicando este método, Vos en 2017

utilizo esta metodología para un mapeo de asociación en papa tetraploide demostrando que varios de los factores tomados en cuenta para el decaimiento del LD dependían directamente de la cantidad de haplotipos presentes, sin embargo, el valor de decaimiento de LD no tuvo una influencia significativa por el número de haplotipos. (Vos et al., 2017)

En los programas de mejoramiento genético agronómico el LD cumple un papel indispensable y de suma importancia, pues brinda información detallada sobre la diversidad y estructura génica de la especie, tal como lo es la colección de mejoramiento genético en papa de la universidad de Nariño en Colombia compuesta por especies tanto diploides como tetraploides. Uno de sus estudios más recientes presenta un análisis de 144 de estos genotipos con una matriz de 8303 SNP's que permitieron diferencias entre 2 y 3 subpoblaciones que diferencias los genotipos diploides de los tetraploides; el estudio revelo altos niveles de diversidad y menor LD en especies tetraploides, mientras que en especies diploides se observó un bajo nivel de diversidad y un mayor LD con respecto a las especies tetraploides; mediante el análisis genético realizado se pudo identificar, reajustar y seleccionar errores de clasificación en la especie relacionados con la poliploidía y genotipos con un gran potencial uso para la hibridación génica por las distancia génica que presenta su LD en los haplotipos.(Berdugo-Cely et al., 2021)

Los estudios mencionados anteriormente presentan algunos factores por los que se puede ver afectado el aumento o la decadencia del LD, razón por la cual no es posible proponer un método estandarizado universal para determinar este, por lo que cada estimación debe realizarse de manera rigurosa y teniendo en cuenta las características genéticas de la especie muestreada y analizada en cuestión.

METODOLOGÍA

BUSQUEDA DE DATOS GBS

Los datos GBS escogidos para el análisis fueron obtenidos de la investigación Whole genome and transcriptome sequencing of *S. tuberosum* Group Phureja DM 1-3 516 R44 descargados directamente de la base de datos y el servidor NCBI (National Center for Biotechnology Information); donde los archivos fueron convertidos a formato .sra a formato fastq en el proceso de descarga mediante el uso de la la herramineta SRA toolkit, implementa en un scrip que automatizo el proceso mediante la siguiente linea de comandos:

mkdir WGSdata_directorio

cd WGSdata_directorio

nano experiment.txt

- SRX8683642
- SRX8683641
- SRX8683640
- SRX8683639
- SRX8683638
- SRX8683637

Nano download_data.sh

```
#!/bin/bash

# Verificar si se proporcionó un argumento para la ruta del archivo
if [ "$#" -ne 1 ]; then
    echo "Uso: $0 <ruta_al_archivo_txt>"
    exit 1
fi

# Ruta al archivo de texto
archivo_txt=$1

# Verificar si el archivo existe
if [ ! -f "$archivo_txt" ]; then
```

```
    echo "Error: El archivo '$archivo_txt' no existe."
    exit 1
fi

# Iterar sobre cada línea del archivo
while IFS= read -r linea; do
    # Verificar si la línea está vacía
    if [ -z "$linea" ]; then
        echo "Advertencia: Línea vacía encontrada en el archivo."
        continue
    fi

    echo "Descargando ... $linea"
# Agregar control de errores para el comando fasterq-dump
    if ! fasterq-dump "$linea"; then
        echo "Error al descargar $linea"
        # Puedes personalizar cómo manejar el error aquí
    fi
done < "$archivo_txt"
chmod +x download_data.sh
./download_data.sh
```

CONTROL DE CALIDAD EN LAS SECUENCIAS

Para visualizar la calidad de las secuencias en cada una de las muestras en un informe detallado se utilizó la herramienta fastqc. Se produjo un informe de calidad inicial con cada una de las muestras mediante el siguiente comando:

```
fastqc -t 6 SRX8683637.fastq -o informes
```

Visualización de comandos completa para cada muestra en Anexos: Bitácora de comandos

Con el timming buscamos principalmente mejorar la calidad por secuencia de bases, eliminar secuencias sobrerrepresentadas y eliminar adaptadores para lo cual utilizamos 3 programas para cada uno de estos módulos: Nanofilt, Porechop y Cutadapt, respectivamente, los parámetros establecidos para cada una de las secuencias fueron los siguientes:

1.1) cutadapt

cutadapt -a AAAAAAAAAA -o SRX8683637_clean.fastq SRX8683637.fastq

1.2)Porechop

~/Documents/Porechop/porechop-runner.py -i SRX8683637_clean.fastq -o SRX8683637_cleanp.fastq --check_reads 10000

1.3)Nanofilt

pipx run nanofilt --headcrop 50 --tailcrop 20 SRX8683637_cleanp.fastq > SRX8683637_cleanpn.fastq

Visualización de comandos completa para cada muestra en Anexos: Bitácora de comandos

ALINEACION DE SECUENCIAS Y SELECCIÓN DEL GENOMA DE REFERENCIA

Para realizar el proceso de alineación de las secuencias con el genoma de referencia se seleccionó el genoma DM 1-3 516 R44. (v6.1); genoma ensamblado y mejorado a escala cromosómica de lectura larga para la papa diploide *Solanum tuberosum phureja*, porporcionado por el laboratorio de Buell de la Universidad de Georgia (G. M. Pham et al., 2020). Para el alineamiento de las secuencias con el genoma se utilizó el software bwa mem, se indexo el genoma de referencia y posteriormente se realizaron las alineaciones de cada una de las muestras con archivo de salida en formato SAM, se realizó mediante los siguientes comandos respectivamente:

bwa index -a bwtsw 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa'

bwa mem -t 4 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' SRX8683637_cleanpn.fastq > alineacion_37.sam

Visualización de comandos completa para cada muestra en Anexos: Bitácora de comandos

ORDENAMIENTO DE LAS ALINEACIONES DE CADA UNA DE LAS MUESTRAS

Para realizar el ordenamiento y de cada una de las muestras se utilizó la herramienta samtools diseñada especificamente para almacenar alineaciones de lectura frente a secuencias de referencia, y admite lecturas cortas y largas (hasta 128 Mbp) producidas por diferentes plataformas de secuenciación (Li et al., 2009), se utilizó la herramienta el comando siguiente comando para ordenar cada una de las muestras:

samtools sort -o sorted_alineacion_37.sam alineacion_37.sam
Visualización de comandos completa para cada muestra en Anexos: Bitácora de comandos

Posteriormente, después de ordenar cada una de las secuencias se realizó la conversión de formato SAM a formato BAM para el consecutivo análisis, para esto se ejecutó el siguiente comando:

samtools view -bS sorted_alineacion_37.sam > sorted_alineacion_37.bam
Visualización de comandos completa para cada muestra en Anexos: Bitácora de comandos

LLAMADA DE VARIANTES EN MUESTRAS INDIVIDUALES

Para la llamada de variantes por muestras individuales se utilizó se utilizó el software NGSEPcore, software encargado del proceso bioinformático para el análisis de datos de secuenciación de próxima generación que realiza un análisis preciso eficiente y fácil de usar para el análisis de datos genómicos (Perea et al., 2016). Para realizar el llamado de variantes en muestras individuales se utilizó el módulo SingleSampleVariantsDetector encargado de realizar llamadas a SNP, pequeños indeles y estructurales, variantes a partir de datos de secuenciación de individuos individuales, según la documentación encontrada en el sitio web github, para cada una de las muestras se utilizó el siguiente comando:

NGSEPcore SingleSampleVariantsDetector -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_37.bam -o variantes_37
Visualización de comandos completa para cada muestra en Anexos: Bitácora de comandos

FUSION DE MUESTRAS INDIVIDUALES Y GENOTIPADO POR MUESTRA

Para realizar la fusión y el genotipado por muestra se utilizó el softwar NGSEPcore, el proceso de divide en 2 partes: fusión de archivos VCF con las variantes identificadas por cada uno de los archivos sin información de genotipo y genotipar cada una de las muestras.
Para la fusión de los archivos VCF generados anteriormente en la llamada de variantes se utilizó el módulo MergeVariants mediante el siguiente comando:

NGSEPcore MergeVariants -s 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa.fai' -o variantes_merged.vcf variantes_37.vcf variantes_38.vcf variantes_39.vcf variantes_40.vcf variantes_41.vcf variantes_42.vcf

Y para el genotipado de cada una de las muestras se utilizó el módulo SingleSampleVariantsDetector ejecutando el siguiente comando:

nohup java -jar /apps/NGSEPcore/NGSEPcore.jar SingleSampleVariantsDetector -maxAlnsPerStartPos 2 -minQuality 40 -knownVariants variantes_merged.vcf -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_37.bam -o genotipos_37.vcf > genotipo_37.log 2>&1 &
Visualización de comandos completa para cada muestra en Anexos: Bitácora de comandos

FUSION DE VARIANTES GENOTIPIFICADAS

Para realizar la fusión cada uno de los archivos vcf con la información de genotipo por muestra se utilizó el software NGSEPcore con el módulo VCFMerge ejecutando el siguiente comando:

NGSEPcore VCFMerge -s 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa.fai' -o variantes_genotipadas genotipos_37.vcf.vcf genotipos_38.vcf.vcf genotipos_39.vcf.vcf genotipos_40.vcf.vcf genotipos_41.vcf.vcf genotipos_42.vcf.vcf

ANOTACIÓN FUNCIONAL DE LAS VARIANTES

La anotación funcional de las variantes se realizó con el software NGSEPcore utilizando el módulo de VCFAnnotate, modulo que proporciona la información funcional relacionada con cada variante, se realizó uso de un archivo en formato GFF3 con información de anotación de modelos genéticos de alta confianza (G. M. Pham et al., 2020), se ejecutó el siguiente comando:

NGSEPcore VCFAnnotate -i variantes_genotipadas.vcf -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -t DM_1-3_516_R44_potato.v6.1.hc_gene_models.gff3 -o anotaciones_variantes_genotipadas.vcf

CONSTRUCCIÓN DE LA MATRIZ DE SNP

Para la construcción de la matriz se utilizó el software vcftools, dicho paquete se encarga de implementar varias utilidades para procesar archivos VCF, incluida la validación, fusión, comparación y también proporciona una API Perl general (Danecek et al., 2011), se realizó un filtrado de variantes, elimina variantes que no son SNP y crea un nuevo archivo VCF recorde.vcf; el comando ejecutado fue el siguiente:

vcftools --vcf variantes_genotipadas.vcf --remove-indels --min-alleles 2 --max-alleles 2 --recode --out variantes_genotipadas_R

Luego se ejecutó un segundo comando utilizando el software PLINK, software encargado análisis de asociación de genoma completo y de enlace basado en poblaciones, para transformar de formato VCF a formato PLINK comando para crear archivos de formato VCF a formato PLINK (un formato común para datos genéticos) y crea dos archivos: salida.map y salida.ped (Purcell et al., 2007); el comando ejecutado fue el siguiente:

vcftools --vcf variantes_genotipadas_R.recode.vcf --plink --out variantes_genotipadas_RMX

Finalmente, después de la conversión de los archivos a los diferentes formatos, para terminar el proceso de crear la matriz SNP se ejecutó el siguiente comando que permite obtener la matriz en formato BED, un formato que permite el análisis estadístico de manera más sencilla, el comando ejecutado fue el siguiente:

plink --file variantes_genotipadas_RMX --make-bed --out matriz_de_genotipos

ANÁLISIS ESTADISTICO

Para realizar el análisis estadístico se usó la herramienta Genome Association and Prediction Integrated Tool (GAPIT) del software Rstudio. Para formato de datos de genotipo GAPIT usa la matriz generada anteriormente en formato numérico; los homocigotos se indican con "0" y "2" y los heterocigotos se indican con "1" en el archivo "GD". Cualquier valor numérico entre "0" y "2" puede representar genotipos de SNP imputados. El archivo "GM" contiene el nombre y la ubicación de cada SNP. La primera columna es la identificación del SNP, la segunda columna es el cromosoma y la tercera columna es la posición del par de bases (Parra Galindo,

2019) . El comando utilizado después de definir las variantes reconocidas por GAPIT es el siguiente:

```
myGD=read.table("matriz_de_genotipos",head=T)
myGM=read.table("anotación_variantes_genotipadas",head=T)

 myGAPIT_MLM <- GAPIT(
 GD=myGD,
 GM=myGM,
 model="MLM",
 PCA.total=5,
 file.output=T)
```

Cabe mencionar que el archivo suministrado remplazado en la variante myGM para realizar el análisis estadístico debe encontrarse en formato de texto pano para que GAPIT pueda indetificarlo, leerlo y posteriormente realizar el analisis estadistico, generar gráficos manhattan, hapmaps y gráficos de dispersión que complementan el análisis asociando las variaciones con las regiones de interés. Durante el análisis de asociación, GAPIT puede utilizar la información de LD entre variantes genéticas para tener en cuenta la correlación entre marcadores cercanos en el genoma. Esto permite identificar asociaciones entre variantes.

RESULTADOS Y ANÁLISIS

DESCARGA DE DATOS GBS DE *Solanum phureja*

Para obtener los datos GBS de *Solanum tuberosum* grupo Phureja se descargaron 6 archivos SRA de la plataforma de datos NCBI (National Center for Biotechnology Information) con el número de adhesión de SRX8683637 a SRX8683642; la investigacion de donde provienen los datos se conoce como Whole genome and transcriptome sequencing of *S. tuberosum* Group Phureja DM 1-3 516 R44 realizada por Gina M Pham, John P Hamilton, Joshua C Wood, Joseph T Burke, Hainan Zhao, et al; en los laboratorios de la Universidad Estatal de Michigan en Estados Unidos, el 23 de septiembre de 2020, identificado mediante doi: 10.1093/gigascience/giaa100.

Los datos descargados están disponibles en la colección de datos biologicos de la plataforma NCBI; filtro Bioproject donde están almacenadas las siguientes lecturas:

Run	**# of Spots**	**# of Bases**	**Size**	**Published**
SRX8683642	536,768	9.4G	8Gb	2020-07-13
SRX8683641	388,107	6,9G	5,9 GB	2020-07-13
SRX8683640	447,909	7,7G	6,5 GB	2020-07-13
SRX8683639	292,014	5,4G	4,6 GB	2020-07-13
SRX8683638	464.750	8,2G	7GB	2020-07-13
SRX8683637	235.873	4,6G	3,9 GB	2020-07-13

Tabla 2 Informacion del conjunto de datos de la investigación Whole genome and transcriptome sequencing of S. tuberosum Group Phureja DM 1-3 516 R44.

Fuente: Autor.

Cabe mencionar que el Bioproyect cuenta con 18 experimentos reportados en la plataforma NCBI, de los cuales 4 son datos de carácter transcriptómica y 14 de carácter genómico, de los 14 experimentos 8 no fue posibles descargarlos ya que son experimentos vacíos, es decir, están reportados, pero no se encuentran publicados en la base de datos de Bioproyect de NCBI, fueron secuenciados con el Kit de secuenciación de ligadura de nanoporos Oxford, SQK-LSK109. (G. M. Pham et al., 2020)

CONTROL DE CALIDAD EN LAS SECUENCIAS

La construcción de las bibliotecas se realizó utilizando el enfoque shotgun del genoma completo en la plataforma Illumina TruSeq DNA Nano y para correcion de errores en el porceso de secuenciacion se utilizó el sistema Illumina HiSeq 2500 (G. M. Pham et al., 2020). Durenate el proceso de conversión de SRA a formato Fastq se generó únicamente un solo archivo correspondiente a la muestra descargada del Bioproyect del NCBI, biblioteca que es de diseño único o single.

El informe generado por Fastqc evalúa varios parámetros como: calidad de secuencia por base, puntuación de calidad por secuencia, porcentaje de guanina y citocina (%GC), secuencias sobrerrepresentadas y contenido de adaptadores, además de dar información adicional como la cantidad de secuencias, la longitud por medio, entre otras. En las siguientes figuras se observa la información descrita anteriormente:

Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683637.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	235873
Bases totales	4,5 GBP
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-305105
%GC	36

Calidad por secuencia de bases

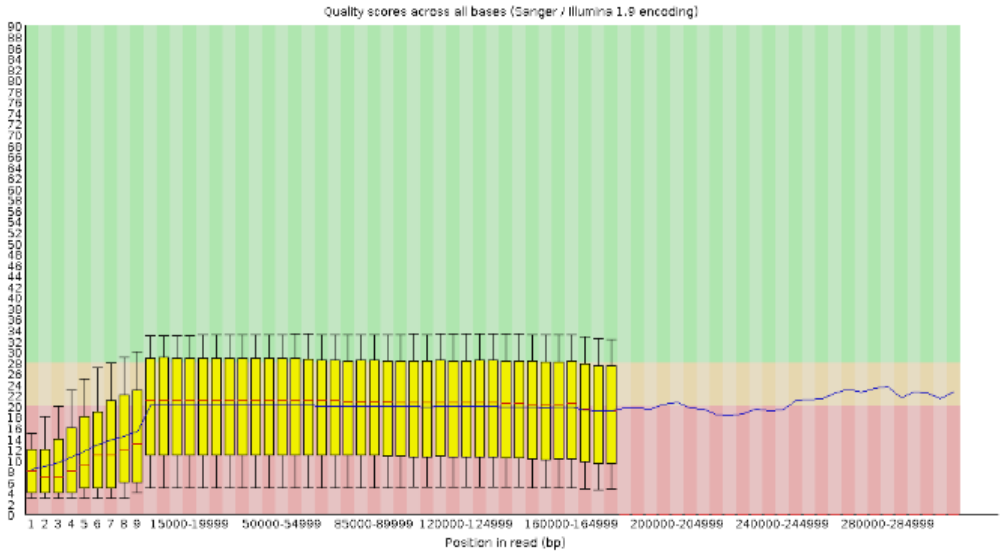


Ilustración 1 Resumen de estadísticas básicas y de la calidad por secuencia de la biblioteca SRX8683637.fastq sin proceso de trimming.

Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683637_cleanpn.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	226454
Bases totales	636,7 Mbps
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-148068
%GC	38

Calidad por secuencia de bases

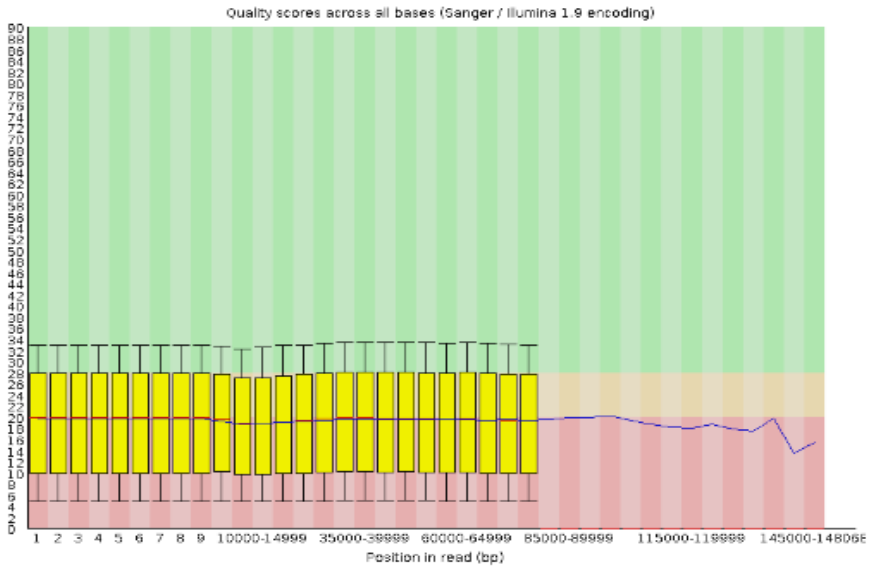


Ilustración 2 Resumen de estadísticas básicas y de la calidad por secuencia de la biblioteca SRX8683637_cleanpn.fastq recortada con nanofilt,Porechop y cutadapt.

El informe de calidad de cada uno de los módulos presentados por Fastqc se puede observar en los *Anexos 2: informes de calidad por muestra*. Al realizar la comparación entre el grafico de la Figura 1 y la figura dos observamos que el módulo de calidad por secuencia de basse sigue presentando una falla; según el instructivo de calidad de secuencia por mosaico, esto se genera si el cuartil inferior de cualquier base es menor que 5 o si la mediana de cualquier base es menor que 20; la línea azul representa la calidad media, como se observa en la figura 1 la calidad media de las lecturas incamente se encuentra limitando con el cuartil inferior, es decir, muestra una tendencia entre 20 y menor a 20, sin embargo, se recortaron secuencias de mucho más baja calidad y dejaron las que cumplían con el criterio de secuencias que mostraran una tendencia de calidad de secuencia limitando con 20; este proceso de filtrado se realizó con el software nanofilt, herramienta encargada para filtrar y recortar datos de secuenciación de nanoporos Oxford, recortar más secuencias para eliminar la playa que presenta el módulo implicaría

perder más del 50% de las secuencias totales lo que significa una pérdida significativa y de información valiosa para el proceso de genotipado. (McClinton et al., 2023)

En la comparación de los informes realizados por Fastqc para el módulo de estadísticas básicas se observa que hubo una disminución del 4% de secuencias con respecto a las secuencias totales antes del proceso de trimming con los 3 software mencionados anteriormente (cutadapt,Porechop y nanofilt); el recorte estas secuencias principalmente se debe al uso de las herramientas Porechop y Nanofilt para eliminar secuencias sobrerrepresentadas, el recorte de posibles adaptadores presentes en medio de las secuencias y realizar un filtrado en la calidad de las secuencias de cada muestra respectivamente (McClinton et al., 2023). La información después del proceso de trimming sobre el módulo de secuencias sobrerrepresentadas para cada una de las bibliotecas analizadas se visualiza en Anexos*: Informes de calidad por muestra.*

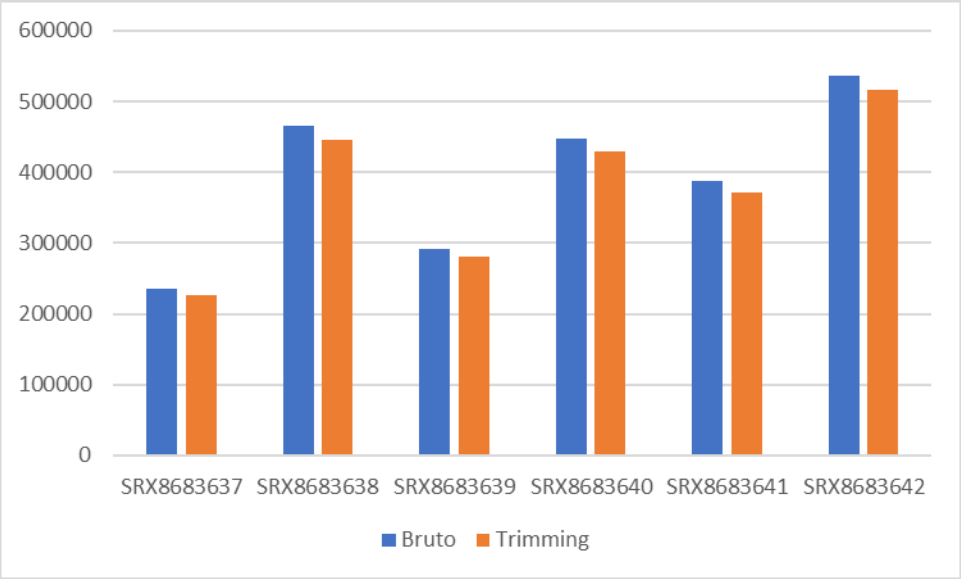


Ilustración 3. Diferencia de las secuencias totales entre las muestras antes y después del proceso de trimming.

Como se observa en la figura 3, el grafico expresa el cambio expresado posteriormente al proceso de trimming y recorte en las bibliotecas analizadas, para ninguna de las 6 bibliotecas analizadas el porcentaje de secuencias recortadas fue mayor al 5%, los recortes que se realizaron fueron mínimos dentro de los que se encuentran un conjunto de secuencias de baja calidad, adaptadores presentes en a los extremos de las secuencias y en medio de estas. Esto se realizó con el fin de no realizar un filtrado agresivo y perder demasiada información que puede contribuir a la investigación.

El trimming realizado para el módulo de contenido del adaptador se trabajó con cutadapt, todas las bibliotecas analizadas presentaban presencia de Poly (A) tal como se observa en las siguientes figuras:

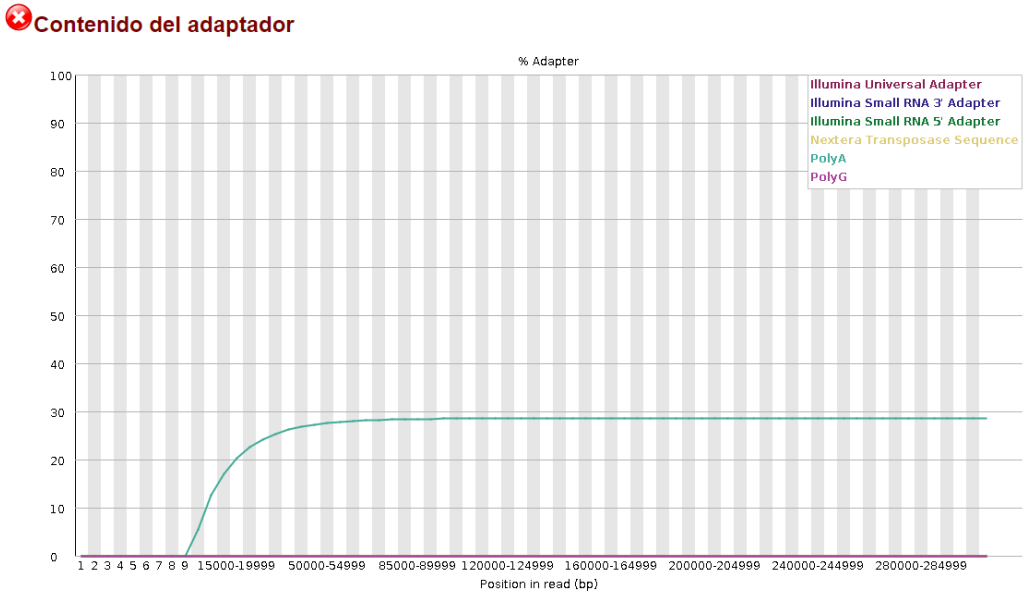


Ilustración 4 Grafica de contenido del adaptador de la biblioteca SRX8683637.fastq sin proceso de trimming.

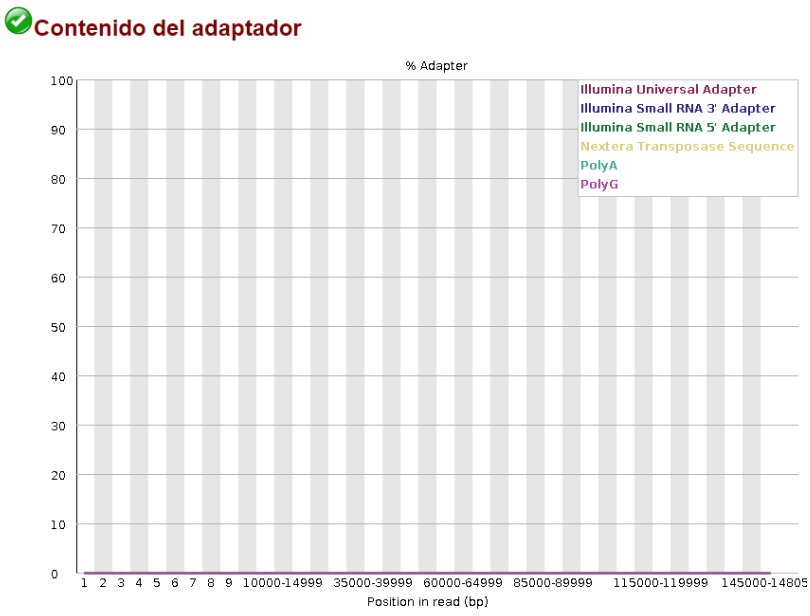


Ilustración 5 Grafica de contenido del adaptador de la biblioteca SRX8683637_cleanpn.fastq después del proceso de trimming con cutadap

Al compara la figura 4 con la figura 5 se observa la disminución y el aplanamiento que tuvo la curva de la gráfica contenido del adaptador entre los datos brutos, de la biblioteca sin tratar,

y despues del proceso de trimming para la biblioteca SRX8683637; para eliminar la inconsistencia que presentaba el módulo contenido del adaptador se utilizó cutadapt para eliminar las colas de Poly (A) presentes en cada una de las bibliotecas, sobre la gráfica del módulo contenido del adaptador para cada una de las bibliotecas analizadas se visualiza en *Anexos: Informes de calidad por muestra*. Cutadapt fue el encargado de eliminar la poliadenilacion presente en los extremos de cada una de las muestras, principalmente son los extremos quienes se ven afectados por este fenómeno debido a la baja eficacia de la polimerasa tanto al inicio como al final de la secuenciación. (He et al., 2020)

ALINEACION DE LAS SECUENCIAS CON EL GENOMA DE REFERENCIA

Posterior al proceso de trimming, las bibliotecas filtradas se alinearon con el genoma de referencia a escala de cromosomas de lectura larga actualizado (v6.1) para la papa monoploide doble *S. tuberosum* Group Phureja DM 1-3 516 R44, descargo de repositorio http://spuddb.uga.edu/dm_v6_1_download.shtml. Para realiza el alineamiento de las 6 bibliotecas analizadas se utilizó el software de mapeo BWA-MEM, herramienta ideal para optimizar CPU y tiempos de proceso con análisis de datos extensos como lo son datos de origen genómico; la salida del archivo se obtiene en formato SAM, el formato SAM contiene la información de las lecturas ya alineadas, es un archivo de texto plano, es importante realizar el ordenamiento por cromosoma y por coordenada cromosómica previo a la conversión a formato BAM, este formato equivalente con una ligera diferencia, la información contenida y almacenada en lenguaje binario, esto automatiza procesos y mejora el rendimiento y la eficiencia. (M. Pham et al., 2023)



Ilustración 6 Visualización de alineamientos apilados y ordenados de las 6 bibliotecas analizadas, en un archivo BAM combinado con las 6 muestras.

El software samtools permite la visualización del alineamiento de cada una de las bibliotecas con el genoma de referencia, tal como se observa en la figura 6, de igual manera, el comando ejecutado permite ver las variaciones genéticas que se presentan en las muestras en una región

o coordenada especifica de un cromosoma en particular, resulta ser una herramienta útil para verificar si paso previo al genotipado se encuentran variaciones cerca a los genes de interés o a los genes a tipificar. En la primera línea de la figura 6 se observan los dígitos de las coordenadas cromosómicas, la segunda linea nos presenta la secuencio o genoma de referencia contra el que se realizó el alineamiento, la tercera línea es generada por defecto, representa la codificación de símbolos para identificar si hay indels o si se las secuencias de las lecturas presentan coincidencia con las secuencias del genoma de referencia, cada punto o coma que se observa significa una coincidencia en la secuencia de la biblioteca alineada con el genoma de referencia, la tercera línea es el genoma de referencia en este lenguaje de puntos; de allí en para debajo de la línea continua cada fila representa una biblioteca alineada con el genoma, los asteriscos representan deleciones en el genoma, y para las inserciones y variantes una letra indica un nucleótido alternativo a la referencia para ese biblioteca en concreto. (Li et al., 2009)

LLAMADA DE VARIANTES, FUSION DE ARCHIVOS VCF Y GENOTIPADO

Para la llamada de variantes se utilizó el software NGSEPcore con el módulo SingleSampleVariantsDetector para cada una de las bibliotecas analizadas, modulo encargado de realizar llamas a SNP, pequeños indels, y variantes estructurales en individuos de manera aislada y por separado. El archivo que se genera contiene las llamadas de base para cada muestra individual, es decir 6 muestras generadas. Para el proceso de fusión de variantes y proceso de genotipado se realiza en dos pasos importantes, después de generar dicho proceso de manera independiente para cada muestra con el módulo anteriormente mencionado es generar el archivo que incluye todas las muestras y todo el conjunto de variantes llamadas de manera individual, generándose un archivo VCF con las llamadas de variantes de las 6 bibliotecas analizadas pero sin información de genotipo en este, para después realizar el genotipado correspondiente de cada muestra de manera independiente y una vez se generó el segundo conjunto de archivos VCF viene la unión de los archivos VCF pero con información del genotipo. Es importante realizar esto ya que para muestras multiplexadas el módulo de NGSEPcore MultisampleVariantsDetector es ideal para realizar análisis de genotipo por secuenciación ya que este módulo genera un archivo VCF de salida con variantes descubiertas y llamadas de genotipo, simplificando el proceso para el genotipado.

En la tabla 3 se observan las estadísticas generales resumidas del proceso de genotipado para las 6 bibliotecas analizadas en el cual podemos observar principalmente los valores de las Variantes identificadas, las llamadas de genotipo y el porcentaje de datos perdidos.

Count	Biallelic SNVs	Biallelic Indels	Biallelic STRs	Other biallelic	Multi allelic SNVs	Multi allelic Indels	Multi allelic STRs	Other Multi allelic
Variants	181070	113362	31931	0	8795	15237	21405	0
Genotype calls	760385	204906	78955	0	43310	36778	63251	0
Percentage of missing data	30.01%	69.87%	58.79%	0.0%	17.93%	59.77%	50.75%	0.0%
Coding variants	0	0	0	0	0	0	0	0
Variants with at least 20 samples genotyped	0	0	0	0	0	0	0	0

Tabla 3 Estadísticas resumidas del proceso de genotipado en las 6 bibliotecas genómicas analizadas.

La categoría muestra una cantidad significativa de Biallelic SNVs muestra una cantidad significativa de (181,070), lo que indica una diversidad sustancial en las variantes de un solo nucleótido. La cantidad de llamadas de genotipos (760,385) sugiere que estas variantes han sido bien caracterizadas en la población analizada. Con 113,362 Biallelic Indels, se observa una presencia significativa de inserciones y deleciones de un solo nucleótido. Las llamadas de genotipos (204,906) indican que estas variantes también han sido evaluadas y caracterizadas, aunque la cantidad de llamadas puede sugerir cierta complejidad. La presencia de 31,931 Biallelic STRs destaca la variabilidad en las repeticiones cortas. Las llamadas de genotipos (78,955) indican una caracterización adecuada, pero la relación entre variantes y llamadas sugiere cierta complejidad en la interpretación de estas regiones. La ausencia de variantes y llamadas de genotipos en la categoría de "Other Biallelic" sugiere que esta categoría está vacía o no se ha definido.

La presencia de 8,795 Multiallelic SNVs sugiere la existencia de variantes de un solo nucleótido con más de dos alelos. Las llamadas de genotipos (43,310) indican que estas

variantes han sido caracterizadas, aunque la cantidad de alelos podría añadir complejidad. Con 15,237 Multiallelic Indels, se identifican inserciones y deleciones con más de dos alelos. Las llamadas de genotipos (36,778) indican un nivel de caracterización, aunque la variabilidad en los alelos podría complicar la interpretación. La presencia de 21,405 Multiallelic STRs indica variabilidad en las repeticiones cortas con múltiples alelos. Las llamadas de genotipos (63,251) sugieren una caracterización, pero la cantidad de alelos podría añadir complejidad a la interpretación. Similar a la categoría de "Other Biallelic", la ausencia de variantes y llamadas de genotipos sugiere que esta categoría no está definida o está vacía.

La diferencia de datos entre las variantes y las mamadas de genotipo se debe a varios factores, dentro de ellos variantes múltiples en una posición, es decir puede haber múltiples variantes coexistentes, esto significa que una posición específica puede tener más de una variante presente en la población; variantes con múltiples alelos, donde una posición del genoma puede tener más de dos alelos, cada alelo adicional se considera una variante única; complejidad en repeticiones cortas (STRs), pues la variabilidad genética puede ser alta, con diferentes longitudes de repeticiones y, por lo tanto, varios alelos y diferencias en cobertura genómica, ya que algunas regiones del genoma pueden tener una cobertura más baja, lo que puede resultar en una subrepresentación de variantes; todo esto puede contribuir a un mayor número de llamadas de genotipos en comparación con el número de variantes identificadas. (McClinton et al., 2023)

El porcentaje de datos perdidos representa el valor porcentual de datos que falta en cada categoría. Los valores altos presentes se pueden deben principalmente a la falta de llamadas de genotipo para algunas bibliotecas analizadas, de igual manera el módulo de variantes codificantes expresa que NGSEPcore no encontró variantes que se encuentren cerca de las regiones codificantes de los genes que modulan el metabolismo en *Solanum tuberosum phureja*. El módulo Variantes con al menos 20 muestras genotipadas también mostro un valor de 0 en todas las categorías ya que NGSEPcore no encontró variantes genotipadas en 20 muestras ya que para la investigación presente se trabajó con menos de 20 datos, únicamente con 6 bibliotecas genómicas. (Perea et al., 2016)

El 30.01% de datos faltantes en Biallelic SNVs indica que alrededor de un tercio de las posiciones de nucleótidos evaluadas para variaciones de un solo nucleótido presentan falta de datos. Esta pérdida puede afectar la precisión en la identificación de variantes genéticas y la

comprensión de la variabilidad en estas posiciones. Con un elevado 69.87% de datos faltantes en Biallelic Indels, la capacidad para identificar inserciones y deleciones de un solo nucleótido se ve significativamente comprometida. Esto podría limitar la comprensión de eventos de inserción y deleción en el genoma. El 58.79% de datos faltantes en Biallelic STRs indica que más de la mitad de las posiciones de Short Tandem Repeats evaluadas carecen de información. Esto puede afectar la caracterización de repeticiones cortas y su papel en la variabilidad genética. La ausencia de datos faltantes (0%) en la categoría de "Other Biallelic" sugiere que esta categoría está bien caracterizada y completamente informada, lo cual es positivo para la interpretación de datos

Con un 17.93% de datos faltantes en Multiallelic SNVs, la capacidad para identificar variantes de un solo nucleótido con más de dos alelos se ve moderadamente afectada. Esto puede influir en la precisión en la identificación de polimorfismos de nucleótido único con múltiples variantes. Similar a los Biallelic Indels, un 59.77% de datos faltantes en Multiallelic Indels sugiere que la capacidad para caracterizar inserciones y deleciones con más de dos alelos está significativamente comprometida. Con un 50.75% de datos faltantes en Multiallelic STRs, más de la mitad de las posiciones de repeticiones cortas con múltiples alelos carecen de información. Esto impacta la capacidad para entender la variabilidad en estas regiones. Al igual que en la categoría de Biallelic, la ausencia de datos faltantes (0%) en la categoría de "Other Multiallelic" sugiere que esta categoría está bien caracterizada y completamente informada.

La tabla 4 presenta la distribución de genotipos posterior a la llamada de variantes y de qué manera se distribuyen las variantes encontradas por categorías:

Samples genotyped	**Biallelic SNVs**	**Biallelic Indels**	**Biallelic STRs**	**Other biallelic**	**Multi allelic SNVs**	**Multi allelic Indels**	**Multi allelic STRs**	**Other Multi allelic**
SRX8683637	32345	67014	12395	0	419	5397	5320	0
SRX8683638	15308	21378	6576	0	561	3816	4611	0
SRX8683639	14812	11911	4949	0	646	2626	3912	0
SRX8683640	17176	7412	3529	0	915	1724	3038	0
SRX8683641	24290	4127	2447	0	1353	1069	2323	0
SRX8683642	77139	1520	2035	0	4901	605	2201	0

Tabla 4 Distribución de genotipos en cada una de las bibliotecas analizadas por categorías.

La tabla presenta los diferentes marcadores moleculares encontrados en las bibliotecas analizadas, un elevado número de polimorfismos encontrados, como los que se observan en las bibliotecas, implican una nueva manera de identificarlos tal y como se observa en el informe proporcionado por NGSEPcore, polimorfismos de un solo nucleótido (SNPs o SNVs son marcadores de polimorfismos de conformación de cadena única), inserciones/deleciones (indels) de longitudes variables (marcadores de regiones amplificadas caracterizadas por secuencia) y recuentos de repeticiones en tándem (número variable de marcadores de repeticiones en tándem); particularmente para *Solanum tuberosum* resulta ser una especie polimórfica de 3 vías, es decir que en su proceso de genotipado se encuentran los 3 tipos de marcadores mencionados anteriormente, en especies que presenten una alta heterocigosidad es común encontrar en el genotipado 2 o más alelos presentes en la población. (Toal et al., 2016)

La categoría que presento un número mayor con respecto a las demás categorías después del proceso de genotipado fue Biallelic SNVs; lo que representa un factor positivo dentro de a investigación debido a que los SNVs (Single Nucleotide Variants) bialélicos, que representan variaciones en un solo nucleótido con dos alelos alternativos, factores importantes dentro de la genética y la conservación de la diversidad genética en poblaciones de *Solanum tuberosum Grupo Phureja*. Estos SNVs son esenciales para prevenir la depresión endogámica, un fenómeno negativo asociado con la pérdida de variabilidad genética en especies con poblaciones que para su cultivo se propaga clonalmente; los SNVs bialélicos en Solanum tuberosum Grupo Phureja representan diferencias genéticas en la población, permitiendo la adaptación a cambios ambientales, resistencia a enfermedades. (Madroñero Cristina et al., 2013)

La naturaleza bialélica de las STR facilita la identificación y diferenciación de individuos dentro de la población. Cada individuo puede tener combinaciones únicas de alelos en las regiones de STR, lo que facilita la construcción de perfiles genéticos distintivos. Las STR bialélicas proporcionan una ventana única hacia la variabilidad genética en poblaciones de Solanum tuberosum Grupo Phureja. La presencia de dos alelos o más (Biallelic STR & Multiaellic STR) en estas regiones refleja la diversidad genética dentro de la población, permitiendo la identificación de individuos únicos y la caracterización de su perfil genético. Las secuencias de repetición tándem (SRTs), justamente su alta variabilidad se debe a las altas tasas de mutación que las regiones STRs pueden presentar y la variabilidad en la longitud del fragmento en los STRs puede hacer que la interpretación de los resultados sea más compleja

en comparación con marcadores con variaciones más simples, como los SNPs. (Toal et al., 2016)

Los Indels (Inserciones y Deleciones) bialélicos y multialélicos en el genoma pueden tener impactos significativos en la interpretación genética y en la aplicación de datos genómicos, pueden llegar a causar alteraciones sustanciales en la estructura genómica, interrumpiendo la secuencia normal de genes y elementos reguladores. Esto podría afectar la función génica y la expresión de características importantes del individuo en cuestión; así mismo, como la anotación de Indels múltiples presenta desafíos adicionales en la interpretación genómica. La identificación precisa de las regiones afectadas y la asignación de funciones biológicas se vuelven más complejas, dificultando la interpretación de los resultados que pueden traducirse en cambios en el marco de lectura, produciendo variantes proteicas no funcionales o con funciones alteradas. Esto podría tener consecuencias negativas en la viabilidad y la adaptabilidad de la población, además de presentar inconvenientes entre la distinción entre variantes causales y aquellas con efectos neutros se vuelve desafiante, lo que puede afectar la identificación precisa de loci asociados a características de interés. (Wang et al., 2022)

Por todo lo mencionado anteriormente la accesión que presenta mejor rendimiento en cuanto a los parámetros mencionados anteriormente es la **SRX8683642,** como se observa en la tabla 4. Presenta un valor de 77139 para variantes bialelicas y 4901 para variantes multialelicas, importantes para la conservación de a diversidad genética en la especie, de igual manera, el bajo número de indels bialelicos y multialelicos con un numero de 1520 y 605 respectivamente representan una pérdida de información baja con respecto a las otras accesiones de *solanum tuberosum grupo phureja* analizadas, lo que proporciona al genotipado una veracidad mas fuerte al momento de identificar loci asociados a zonas de interés. Por otro lado el número de esta accesión también presenta niveles bajos de STR bialelicas y multialeicas, con un valor para cada una de 2035 y de 2201 respectivamente, a pesar de presentar los valores más bajos en estas categorías con respecto a las demás accesiones, esto no representa un inconveniente al presentar una simplicidad en anotación e interpretación de los datos, menos alelos en estas regiones facilitan la identificación y la asignación de funciones biológicas y, a su vez, puede facilitar la asociación de variantes con características fenotípicas en estudios de asociación genética. La reducción de la complejidad puede mejorar la precisión en la identificación de loci asociados.

La accesión con los valores y el rendimiento más bajo según lo observado en la tabla 4. Los presenta la **SRX8683637** con un número muy elevado de indels bialelicos y multialelicos, es la accesión que mayor cantidad de indels presenta entre las 6 analizadas, esto se traduce en una pérdida de información puede representar un sesgo muy grande al momento de realizar el análisis de asociación genómica, principalmente al momento de realizar la alineación genómica para el genotipado pues la falta de información puede generar una interrupción en la estructuración genómica que se está realizando. A pesar de ser la segunda accesión con mayor cantidad de SNV biaelilicos y multialelicos, el elevado número de indels presentado aumenta la dificultad para el análisis y la precisión al momento de identificar regiones si estas fueron afectadas; de igual manera la veracidad se ve compromotida por la gran posibilidad de los cambios que se pueden generar en el cambio de lectura. Las STR bialelicas y multialelicas si bien presentan los valores más elevados en esta accesión, para realizar le genotipado estas variaciones presentan una dificultad grande por la cantidad de mutaciones que pueden presentar, la identificación y asignación de funciones biológicas a múltiples alelos en estas regiones se vuelve desafiante, dificultando la comprensión precisa de la variabilidad genéticas, por lo que a pesar de ser importantes y suministrar información adicional no pueden tomarse únicamente en cuenta para el estudio de asociación.

Dentro de las accesiones mejor calificadas con una distribución mucho más pareja con relación todas las categorías son **SRX8683641** y **SRX8683640**, presentando valores bajo de indels bialelicos y multialelicos, a pesar de no tener valores tan elevados en las demás categorías, la información se ha conservado y es lo que permite realizar el análisis de una manera mas eficiente. Para las accesiones **SRX8683639** y **SRX8683638** los valores también son bajos y el número elevado en las categorías de indels bialelicos y multialelicos representan un desafío para el análisis de asociación.

ANOTACION DE VARIANTES, MATRIZ SNP Y ANALISIS ESTADISTICO

El software NGSEPcore no realizo de manera correcta la anotación funcional de las variantes y presentaba un mensaje de advertencia, es el que se muestra en la siguiente imagen:

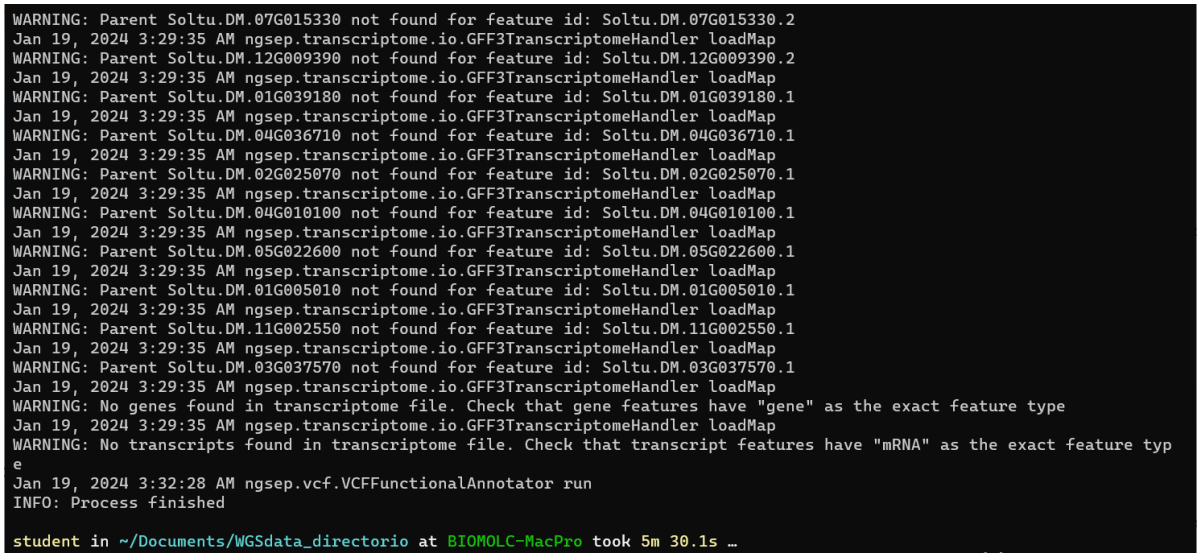


Ilustración 7 Archivo.log con la información contenida durante el proceso de anotación funcional de variantes en archivos vcf fusionados

Como se observa en la imagenLa advertencia "WARNING: Parent Soltu.DM.05G014690 not found for feature id: Soltu.DM.05G014690.1" que menciona sugiere que hay una discrepancia o falta de información en la anotación genómica proporcionada a NGSEPcore. Esta advertencia indica que NGSEPcore no puede encontrar la información del "padre" asociada a la característica genómica con el identificador en el archivo GFF3. Unas posibles razones por las que esto pudo haber ocurrido es la falta de anotación del gen "Padre", es decir, la advertencia puede deberse a que la anotación genómica proporcionada no incluye información sobre el gen "padre" de la característica específica para *Solanum tuberosum* DM.05G014690.1.

El hecho de que las accesiones analizadas para el Genotipado por secuencia (GBS) sean las mismas con las que se ensambló el genoma de referencia no debería, en principio, causar problemas con respecto a la identificación de los "padres" de las características genómicas durante la anotación con NGSEPcore. La anotación funcional con NGSEPcore se basa en archivos GFF3 que contienen información sobre las características genómicas, como genes y sus transcriptos asociados. Como se observó en la tabla 3, no se encontraron variantes codificantes, es posible que la falta de variantes codificantes en el proceso de genotipado pueda influir en la anotación funcional, pero no debería ser la única razón para que NGSEPcore no pueda realizar la anotación de manera correcta. La anotación funcional generalmente se realiza sobre todas las variantes detectadas, ya sean codificantes o no codificantes, para proporcionar información detallada sobre su impacto biológico. (Díaz et al., 2021)

La razón principal puede deberse a errores de formato, sintaxis o información faltante en el archivo GFF3; el mensaje de advertencia " No se encontraron transcripciones en el archivo transcriptoma. Verifique que las características de la transcripción tengan 'mRNA' como el tipo de característica exacto" indica que el software NGSEPcore no ha encontrado transcritos en el archivo de transcriptoma y sugiere que los transcritos deben tener el tipo de característica "mRNA" en el archivo GFF3, por lo que el error al momento de realizar las anotaciones puede atribuirse a un error en lectura del contenido en el archivo de anotación de modelos genéticos de alta confianza en formato GFF3, tal como lo solicita la documentación de NGSEPcore.

La información de anotación funcional se utiliza para preparar la matriz de SNP. Se incluyen características adicionales, como la ubicación de las variantes en relación con los genes y las regiones genómicas importantes. Al comprender la función de las variantes, se reduce la posibilidad de identificar falsos positivos o negativos en el análisis de asociación genómica y puede mejorar la potencia estadística del análisis de asociación genómica al considerar no solo la presencia o ausencia de variantes, sino también su impacto biológico. De igual manera, para realizar el análisis de asociación genómica la anotación funcional de variantes es crucial, ya que permite la priorización de variantes para el GWAS. La anotación funcional ayuda a centrarse en variantes que podrían tener un impacto significativo en la función génica o en la regulación biológica, ayuda a comprender cómo las variantes podrían afectar a los genes, permite eliminar variantes que no son relevantes para el estudio, facilita la inferencia de los efectos genéticos de las variantes y permite la reproducibilidad de los resultados del GWAS al proporcionar un contexto biológico detallado de las variantes estudiadas. (Díaz et al., 2021)

CONCLUSIONES

El análisis de genotipificación por secuencia es un método bastante utilizado para una gran cantidad de muestras de ADN de forma simultánea generando miles de marcadores SNP que son posteriormente analizados mediante un estudio de asociación genética en el se permita explorar y descubrir nuevas regiones genómicas de interés particular, resultando un método de secuenciación a gran escala y a bajo costo. El protocolo y flujo de trabajo descrito en la metodología permite observar que se abarco el proyecto desde la obtención de datos hasta el análisis estadístico y el estudio de asociación del genoma, el ultimo mencionado no se llevo a cabo debido a las implicaciones que se tuvo durante el proceso de anotación funcional de las variantes, paso decisivo y crucial al momento de realizar el estudio de asociación genómica para garantizar el éxito y la interpretación adecuada de un GWAS. Sin esta información, el análisis de asociación genómica podría pasar por altas variantes importantes o generar asociaciones espurias.

El mapeo y estudios de asociación genética presenta un panorama muy amplio del genoma completo, es una técnica que puede ser aplicada a múltiples especies, algunas limitaciones que presenta la técnica es que debido a la cantidad de información que se procesa durante el estudio dificulta que sea aplicable a especies con una ploidía muy elevada, principalmente se realiza en especies diploides, tal como lo es *Solanum tuberosum grupo phureja*. Sin embargo, también se puede realizar en especies poliploides resultado un poco más complejo su análisis. Para el análisis bioinformático el software NGSEPcore resulto ser un software dinámico y relativamente sencillo al momento de ejecutarse, su instructivo es bastante claro permite automatizar datos de manera rápida y eficiente generando archivos de salida en distintos formatos para que sea compatible con otras plataformas; de igual manera Samtools y BWA MEM resultaron ser herramientas importantes en el proceso de alineamiento, indexación y ordenamiento de la secuencias permitiendo ser herramientas versátiles y compatibles con el uso de otros softwares.

El proceso de genotipado permitió identificar por categorías las variantes encontradas en cada categoría, en la población analizada no se encontraron variantes en regiones codificantes por lo que se intuye de las variantes encontradas se localizan cerca de exones o en regiones no codificantes, es importante considerar la naturaleza específica de las variantes detectadas, como la presencia de variantes múltiples y alelos adicionales, al interpretar estos resultados. Además,

la calidad de los datos, los criterios de filtrado y la cobertura genómica son aspectos claves que pueden influir en estas discrepancias. Este análisis detallado proporciona una visión completa de la diversidad genética en cada categoría de variantes, destacando áreas de complejidad que pueden requerir atención adicional en futuros análisis y estudios genéticos.

RECOMENDACIONES

Para próximos trabajos de investigación:

- Se recomienda trabajar con la población mutante irradiada con Co-60 para identificar posibles polimorfismos causados por la radiación.

- Buscar un software alternativo a NGSEPcore para realizar el proceso de genotipado y anotación de funcional de variantes para poder establecer una comparación entre varias herramientas de análisis que permitan el desarrollo de mejores compilaciones de estas para disminuir errores de tipo metodológico.

- Enforcarse en la creación de un archivo con anotaciones funcionales y modelos genéticos de interés encaminado en los genes analizados por el grupo de investigación Biomol, en genes que regulan la producción de fitohormonas y genes candidatos implicados en dormancia.

BIBLIOGRAFÍA

Achakkagari, S. R., Kyriakidou, M., Gardner, K. M., De Koeyer, D., De Jong, H., Strömvik, M. V., & Tai, H. H. (2022). Genome sequencing of adapted diploid potato clones. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.954933

Barrera Alvarado, J. A., & Avellaneda Ramírez, M. A. (2016). *RESPUESTA DE LA PAPA CRIOLLA (SOLANUM PHUREJA) A DIFERENTES APLICACIONES DE FERTILIZACIÓN ORGÁNICO MINERAL EN BOGOTA, CUNDINAMARCA*.

Berdugo-Cely, J. A., Martínez-Moncayo, C., & Lagos-Burbano, T. C. (2021). Genetic analysis of a potato (Solanum tuberosum L.) breeding collection for southern Colombia using Single Nucleotide Polymorphism (SNP) markers. *PLoS ONE*, *16*(3 March). https://doi.org/10.1371/journal.pone.0248787

Bonierbale, M., Amoros, W., Espinoza, J., Mihovilovich, E., Roca, W., & Gómez, R. (2004). Recursos genéticos de la papa: Don del pasado, legado para el futuro. *Revista Latinoamericana de La Papa*.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15). https://doi.org/10.1093/bioinformatics/btr330

Díaz, P., Sarmiento, F., Mathew, B., Ballvora, A., & Vásquez, T. M. (2021). Genomic regions associated with physiological, biochemical and yield-related responses under water deficit in diploid potato at the tuber initiation stage revealed by GWAS. *PLoS ONE*, *16*(11 November). https://doi.org/10.1371/journal.pone.0259690

Finch-Savage, W. E., & Leubner-Metzger, G. (2006). Seed dormancy and the control of germination. In *New Phytologist* (Vol. 171, Issue 3). https://doi.org/10.1111/j.1469-8137.2006.01787.x

Gao, X. L., Jing, F. X., Yang, J. B., & Zhao, J. L. (2005). Detection for single nucleotide polymorphisms. In *Yi chuan = Hereditas / Zhongguo yi chuan xue hui bian ji* (Vol. 27, Issue 1). https://doi.org/10.21775/cimb.005.043

Gebhardt, C., Ballvora, A., Walkemeier, B., Oberhagemann, P., & Schüler, K. (2004). Assessing genetic potential in germplasm collections of crop plants by marker-trait association: A case study for potatoes with quantitative variation of resistance to late blight and maturity type. *Molecular Breeding*, *13*(1). https://doi.org/10.1023/B:MOLB.0000012878.89855.df

Gong, H. L., Dusengemungu, L., Igiraneza, C., & Rukundo, P. (2021). Molecular regulation of potato tuber dormancy and sprouting: a mini-review. In *Plant Biotechnology Reports* (Vol. 15, Issue 4). https://doi.org/10.1007/s11816-021-00689-y

Haider, M. W., Nafees, M., Ahmad, I., Ali, B., Maryam, B., Iqbal, R., Vodnar, D. C., Marc, R. A., Kamran, M., Saleem, M. H., Al-Ghamdi, A. A., Al-Hemaid, F. M., & Elshikh, M. S. (2022). Postharvest dormancy-related changes of endogenous hormones in relation to different dormancy-breaking methods of potato (Solanum tuberosum L.) tubers. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.945256

Hamwieh, A., Udupa, S. M., Choumane, W., Sarker, A., Dreyer, F., Jung, C., & Baum, M. (2005). A genetic linkage map of Lens sp. based on microsatellite and AFLP markers and the localization of fusarium vascular wilt resistance. *Theoretical and Applied Genetics*, *110*(4). https://doi.org/10.1007/s00122-004-1892-5

Hartmann, A., Senning, M., Hedden, P., Sonnewald, U., & Sonnewald, S. (2011). Reactivation of meristem activity and sprout growth in potato tubers require both cytokinin and gibberellin. *Plant Physiology*, *155*(2). https://doi.org/10.1104/pp.110.168252

Hayes, B. (2013). Overview of statistical methods for genome-wide association studies (GWAS). *Methods in Molecular Biology*, *1019*. https://doi.org/10.1007/978-1-62703-447-0_6

He, B., Zhu, R., Yang, H., Lu, Q., Wang, W., Song, L., Sun, X., Zhang, G., Li, S., Yang, J., Tian, G., Bing, P., & Lang, J. (2020). Assessing the Impact of Data Preprocessing on Analyzing Next Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology*, *8*. https://doi.org/10.3389/fbioe.2020.00817

JUYÓ ROJAS, D. K. (2012). *DIVERSIDAD GENÉTICA Y ESTRUCTURA POBLACIONAL EN GENOTIPOS DIPLOIDES DE PAPA*.

Juyo Rojas, D. K., Soto Sedano, J. C., Ballvora, A., Léon, J., & Mosquera Vásquez, T. (2019). Novel organ-specific genetic factors for quantitative resistance to late blight in potato. *PLOS ONE*, *14*(7), e0213818. https://doi.org/10.1371/journal.pone.0213818

Kumar, S., Banks, T. W., & Cloutier, S. (2012). SNP discovery through next-generation sequencing and its applications. In *International Journal of Plant Genomics* (Vol. 2012). https://doi.org/10.1155/2012/831460

Lang, G. A., Early, J. D., Martin, G. C., & Darnell, R. L. (2022). "Endo-, Para, and Ecodormancy: Physiological Terminology and Classification for Dormancy Research." *HortScience*, *22*(5). https://doi.org/10.21273/hortsci.22.5.701b

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16). https://doi.org/10.1093/bioinformatics/btp352

Ligarreto, G. A., & Suárez, M. N. (2003). *Evaluation of the potential of genetics resources of Creole potato (Solanum phureja) for industrial quality* (Vol. 21, Issue 2).

López Contreras, L. M. (2020). *Arquitectura genética de la tolerancia al estrés por déficit hídrico en papa diploide Genetic architecture of tolerance to water deficit stress in diploid potato*.

Madroñero Cristina, I., Rodríguez Ernesto, L. M., Fernando Navia, J. E., & Andrés Benavides, C. (2013). *CARACTERIZACIÓN MORFOAGRONOMICA DE GENOTIPOS PROMISORIOS DE PAPA CRIOLLA (Solanum tuberosum L.Grupo Andigenum) EN NARIÑO. MORPHO-AGRONOMIC CHARACTERIZATION OF PROMISING NATIVE CREOLE POTATO GENOTYPES (Solanum tuberosum L. Andigenum group) IN NARIÑO*.

Martinez Corcino, D. S. (2012). *Ubicación de secuencias genómicas de Solanum phureja en un mapa genético utilizando marcadores microsatélites.*

McClinton, B., Crinnion, L. A., McKibbin, M., Mukherjee, R., Poulter, J. A., Smith, C. E. L., Ali, M., Watson, C. M., Inglehearn, C. F., & Toomes, C. (2023). Targeted nanopore sequencing enables complete characterisation of structural deletions initially identified using exon-based short-read sequencing strategies. *Molecular Genetics and Genomic Medicine*, *11*(6). https://doi.org/10.1002/mgg3.2164

Naz, R. M. M., Li, M., Ramzan, S., Li, G., Liu, J., Cai, X., & Xie, C. (2018). QTL mapping for microtuber dormancy and GA3 content in a diploid potato population. *Biology Open*, *7*(1). https://doi.org/10.1242/bio.027375

Ñústez López, & Rodríguez Molano. (2020). *Papa criolla (Solanum tuberosum Grupo Phureja)*.

Parra Galindo, M. A. (2019). *Análisis de rutas biológicas y estudio de asociación amplia del genoma identificaron factores genéticos del cromosoma 10 asociados con la variación de antocianinas en papa Biological Pathway Analysis and Genome Wide Association Study identified genetic factors on chromosome 10 associated with anthocyanin variation in potato*.

Penagos Beltrán, A. M. (2020). *COMPOSICIÓN EN NUTRIENTES Y BIOACTIVOS Y PROPUESTA DE BIORREFINERÍA DE PAPA Solanum tuberosum Grupo Phureja*.

Perea, C., De La Hoz, J. F., Cruz, D. F., Lobaton, J. D., Izquierdo, P., Quintero, J. C., Raatz, B., & Duitama, J. (2016). Bioinformatic analysis of genotype by sequencing (GBS) data with NGSEP. *BMC Genomics*, *17*. https://doi.org/10.1186/s12864-016-2827-7

Pham, G. M., Hamilton, J. P., Wood, J. C., Burke, J. T., Zhao, H., Vaillancourt, B., Ou, S., Jiang, J., & Robin Buell, C. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. *GigaScience*, *9*(9). https://doi.org/10.1093/gigascience/giaa100

Pham, M., Tu, Y., & Lv, X. (2023). Accelerating BWA-MEM Read Mapping on GPUs. *Proceedings of the International Conference on Supercomputing*. https://doi.org/10.1145/3577193.3593703

Prudencio, Á. S., Hoeberichts, F. A., Dicenta, F., Martínez-Gómez, P., & Sánchez-Pérez, R. (2021). Identification of early and late flowering time candidate genes in endodormant and ecodormant almond flower buds. *Tree Physiology*, *41*(4). https://doi.org/10.1093/treephys/tpaa151

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3). https://doi.org/10.1086/519795

Rodríguez, L., Ñustez, C., & Estrada, N. (2009). Criolla Latina, Criolla Paisa y Criolla Colombia, nuevos cultivares de papa criolla para el departamento de Antioquia (Colombia). *Agronomía Colombiana*, *27*(3).

Shu, K., Liu, X. D., Xie, Q., & He, Z. H. (2016). Two Faces of One Seed: Hormonal Regulation of Dormancy and Germination. In *Molecular Plant* (Vol. 9, Issue 1, pp. 34–45). Cell Press. https://doi.org/10.1016/j.molp.2015.08.010

Slatkin, M. (2008). Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. In *Nature Reviews Genetics* (Vol. 9, Issue 6). https://doi.org/10.1038/nrg2361

Socorro, D., Lasso, C., Contreras, L. P., Magdalena, A., Holguín, G., & Ortega, J. G. (2022). *PAPA NATIVA DIPLOIDE*.

Spooner, D. M., & Hijmans, R. J. (2001). Potato systematics and germplasm collecting, 1989-2000. *American Journal of Potato Research*, *78*(4). https://doi.org/10.1007/BF02875691

Straadt, I. K., & Rasmussen, O. S. (2003). AFLP analysis of Solanum phureja DNA introgressed into potato dihaploids. *Plant Breeding*, *122*(4). https://doi.org/10.1046/j.1439-0523.2003.00878.x

Toal, T. W., Burkart-Waco, D., Howell, T., Ron, M., Kuppu, S., Britt, A., Chetelat, R., & Brady, S. M. (2016). Indel group in genomes (IGG) molecular genetic markers. *Plant Physiology*, *172*(1). https://doi.org/10.1104/pp.16.00354

Torkamaneh, D., Laroche, J., & Belzile, F. (2016). Genome-wide SNP calling from genotyping by sequencing (GBS) data: A comparison of seven pipelines and two

sequencing technologies. *PLoS ONE*, *11*(8). https://doi.org/10.1371/journal.pone.0161333

Toro Pantoja, Mi. D. C. (2015). Evaluación del efecto de 3 inhibidores de brotación en papa criolla (Solanum phureja) variedad criolla colombia aplicados en el proceso de poscosecha. *Universidad Nacional*.

v28n2a08. (n.d.).

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. In *American Journal of Human Genetics* (Vol. 90, Issue 1). https://doi.org/10.1016/j.ajhg.2011.11.029

Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G. F., van Eck, H. J., & van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theoretical and Applied Genetics*, *130*(1). https://doi.org/10.1007/s00122-016-2798-8

Wang, N., Lysenkov, V., Orte, K., Kairisto, V., Aakko, J., Khan, S., & Elo, L. L. (2022). Tool evaluation for the detection of variably sized indels from next generation whole genome and targeted sequencing data. *PLoS Computational Biology*, *18*(2). https://doi.org/10.1371/journal.pcbi.1009269

Wong, M. M. L., Gujaria-Verma, N., Ramsay, L., Yuan, H. Y., Caron, C., Diapari, M., Vandenberg, A., & Bett, K. E. (2015). Classification and characterization of species within the genus lens using genotyping-by-sequencing (GBS). *PLoS ONE*, *10*(3). https://doi.org/10.1371/journal.pone.0122025

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., & Lin, X. (2010). Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *American Journal of Human Genetics*, *86*(6). https://doi.org/10.1016/j.ajhg.2010.05.002

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J., Orjeda, G., Guzman, F., Torres, M., Lozano, R., Ponce, O., Martinez, D., De La Cruz, G., Chakrabarti, S. K., Patil, V. U., … Visser, R. G. F. (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, *475*(7355). https://doi.org/10.1038/nature10158

Zhu, H., & Zhou, X. (2020). Statistical methods for SNP heritability estimation and partition: A review. In *Computational and Structural Biotechnology Journal* (Vol. 18). https://doi.org/10.1016/j.csbj.2020.06.011

ANEXOS:

BITACORA DE COMANDOS:

1) Pretratamiento de datos:

1.1) cutadapt

cutadapt -a AAAAAAAAAA -o SRX8683637_clean.fastq SRX8683637.fastq
cutadapt -a AAAAAAAAAA -o SRX8683638_clean.fastq SRX8683638.fastq
cutadapt -a AAAAAAAAAA -o SRX8683639_clean.fastq SRX8683639.fastq
cutadapt -a AAAAAAAAAA -o SRX8683640_clean.fastq SRX8683640.fastq
cutadapt -a AAAAAAAAAA -o SRX8683641_clean.fastq SRX8683641.fastq
cutadapt -a AAAAAAAAAA -o SRX8683642_clean.fastq SRX8683642.fastq

1.2)Porechop

~/Documents/Porechop/porechop-runner.py -i SRX8683637_clean.fastq -o SRX8683637_cleanp.fastq --check_reads 10000
~/Documents/Porechop/porechop-runner.py -i SRX8683638_clean.fastq -o SRX8683638_cleanp.fastq --check_reads 10000
~/Documents/Porechop/porechop-runner.py -i SRX8683639_clean.fastq -o SRX8683639_cleanp.fastq --check_reads 10000
~/Documents/Porechop/porechop-runner.py -i SRX8683640_clean.fastq -o SRX8683640_cleanp.fastq --check_reads 10000
~/Documents/Porechop/porechop-runner.py -i SRX8683641_clean.fastq -o SRX8683641_cleanp.fastq --check_reads 10000
~/Documents/Porechop/porechop-runner.py -i SRX8683642_clean.fastq -o SRX8683642_cleanp.fastq --check_reads 10000

1.3)Nanofilt

pipx run nanofilt --headcrop 50 --tailcrop 20 SRX8683637_cleanp.fastq > SRX8683637_cleanpn.fastq
pipx run nanofilt --headcrop 50 --tailcrop 20 SRX8683638_cleanp.fastq > SRX8683638_cleanpn.fastq

pipx run nanofilt --headcrop 50 --tailcrop 20 SRX8683639_cleanp.fastq > SRX8683639_cleanpn.fastq
pipx run nanofilt --headcrop 50 --tailcrop 20 SRX8683640_cleanp.fastq > SRX8683640_cleanpn.fastq
pipx run nanofilt --headcrop 50 --tailcrop 20 SRX8683641_cleanp.fastq > SRX8683641_cleanpn.fastq
pipx run nanofilt --headcrop 50 --tailcrop 20 SRX8683642_cleanp.fastq > SRX8683642_cleanpn.fastq

2)Alineacion de secuencias

bwa mem -t 4 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' SRX8683637_cleanpn.fastq > alineacion_37.sam
bwa mem -t 6 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' SRX8683638_cleanpn.fastq > alineacion_38.sam
bwa mem -t 6 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' SRX8683639_cleanpn.fastq > alineacion_39.sam
bwa mem -t 6 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' SRX8683640_cleanpn.fastq > alineacion_40.sam
bwa mem -t 6 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' SRX8683641_cleanpn.fastq > alineacion_41.sam
bwa mem -t 6 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' SRX8683642_cleanpn.fastq > alineacion_42.sam

3)Organizacion de los alineamientos:

samtools sort -o sorted_alineacion_37.sam alineacion_37.sam
samtools sort -o sorted_alineacion_38.sam alineacion_38.sam
samtools sort -o sorted_alineacion_39.sam alineacion_39.sam
samtools sort -o sorted_alineacion_40.sam alineacion_40.sam
samtools sort -o sorted_alineacion_41.sam alineacion_41.sam
samtools sort -o sorted_alineacion_42.sam alineacion_42.sam

4)Conversion de formato SAM a formato BAM:

samtools view -bS sorted_alineacion_37.sam > sorted_alineacion_37.bam

samtools view -bS sorted_alineacion_38.sam > sorted_alineacion_38.bam

samtools view -bS sorted_alineacion_39.sam > sorted_alineacion_39.bam

samtools view -bS sorted_alineacion_40.sam > sorted_alineacion_40.bam

samtools view -bS sorted_alineacion_41.sam > sorted_alineacion_41.bam

samtools view -bS sorted_alineacion_42.sam > sorted_alineacion_42.bam

5)Lamada de varientes individual:

NGSEPcore SingleSampleVariantsDetector -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_37.bam -o variantes_37

NGSEPcore SingleSampleVariantsDetector -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_38.bam -o variantes_38

NGSEPcore SingleSampleVariantsDetector -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_39.bam -o variantes_39

NGSEPcore SingleSampleVariantsDetector -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_40.bam -o variantes_40

NGSEPcore SingleSampleVariantsDetector -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_41.bam -o variantes_41

NGSEPcore SingleSampleVariantsDetector -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_42.bam -o variantes_42

6) Indexacion con samtol de genoma de referencia:

samtools faidx 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa

7)Variantes fusionadas:

NGSEPcore MergeVariants -s 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa.fai' -o variantes_merged.vcf variantes_37.vcf variantes_38.vcf variantes_39.vcf variantes_40.vcf variantes_41.vcf variantes_42.vcf

8)Genotipado de variantes para cada una de las muestras:

nohup java -jar /apps/NGSEPcore/NGSEPcore.jar SingleSampleVariantsDetector -maxAlnsPerStartPos 2 -minQuality 40 -knownVariants variantes_merged.vcf -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_37.bam -o genotipos_37.vcf > genotipo_37.log 2>&1 &

nohup java -jar /apps/NGSEPcore/NGSEPcore.jar SingleSampleVariantsDetector -maxAlnsPerStartPos 2 -minQuality 40 -knownVariants variantes_merged.vcf -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_38.bam -o genotipos_38.vcf > genotipo_38.log 2>&1 &

nohup java -jar /apps/NGSEPcore/NGSEPcore.jar SingleSampleVariantsDetector -maxAlnsPerStartPos 2 -minQuality 40 -knownVariants variantes_merged.vcf -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_39.bam -o genotipos_39.vcf > genotipo_39.log 2>&1 &

nohup java -jar /apps/NGSEPcore/NGSEPcore.jar SingleSampleVariantsDetector -maxAlnsPerStartPos 2 -minQuality 40 -knownVariants variantes_merged.vcf -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_40.bam -o genotipos_40.vcf > genotipo_40.log 2>&1 &

nohup java -jar /apps/NGSEPcore/NGSEPcore.jar SingleSampleVariantsDetector -maxAlnsPerStartPos 2 -minQuality 40 -knownVariants variantes_merged.vcf -r 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_41.bam -o genotipos_41.vcf > genotipo_41.log 2>&1 &

nohup java -jar /apps/NGSEPcore/NGSEPcore.jar SingleSampleVariantsDetector -maxAlnsPerStartPos 2 -minQuality 40 -knownVariants variantes_merged.vcf -r 'DM_1-

3_516_R44_potato_genome_assembly.v6.1 (1).fa' -i sorted_alineacion_42.bam -o genotipos_42.vcf > genotipo_42.log 2>&1 &

9)Variantes fusionadas con informacion de genotipo:

NGSEPcore VCFMerge -s 'DM_1-3_516_R44_potato_genome_assembly.v6.1 (1).fa.fai' -o variantes_genotipadas genotipos_37.vcf.vcf genotipos_38.vcf.vcf genotipos_39.vcf.vcf genotipos_40.vcf.vcf genotipos_41.vcf.vcf genotipos_42.vcf.vcf

10) Matriz con vcftools y plink

comando para el filtrado de variantes, elimina variantes que no son SNP y crea un nuevo archivo VCF recorde.vcf:

vcftools --vcf variantes_genotipadas.vcf --remove-indels --min-alleles 2 --max-alleles 2 --recode --out variantes_genotipadas_R

comando para crear archivos de formato VCF a formato PLINK (un formato común para datos genéticos) y crea dos archivos: salida.map y salida.ped:

vcftools --vcf variantes_genotipadas_R.recode.vcf --plink --out variantes_genotipadas_RMX

comando para crear la matriz:

plink --file variantes_genotipadas_RMX --make-bed --out matriz_de_genotipos

11)Analisis estadistico:

```
myGD=read.table("matriz_de_genotipos",head=T)
myGM=read.table("anotación_variantes_genotipadas",head=T)

 myGAPIT_MLM <- GAPIT(
 GD=myGD,
 GM=myGM,
```

```
model="MLM",
PCA.total=5,
file.output=T)
```

INFORMES DE CALIDAD:

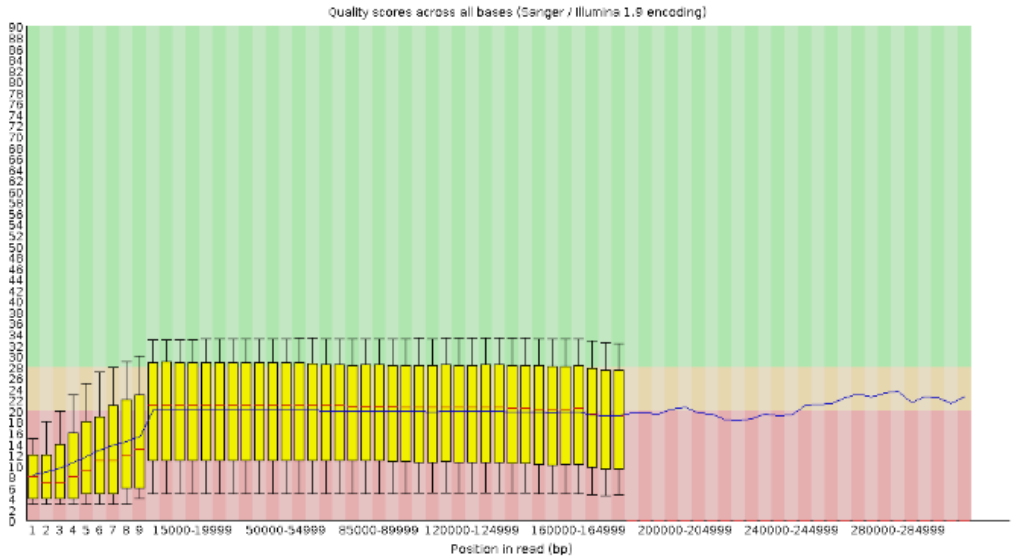
2.1) INFORMES DE CALIDAD EN BRUTO

SRX8683637

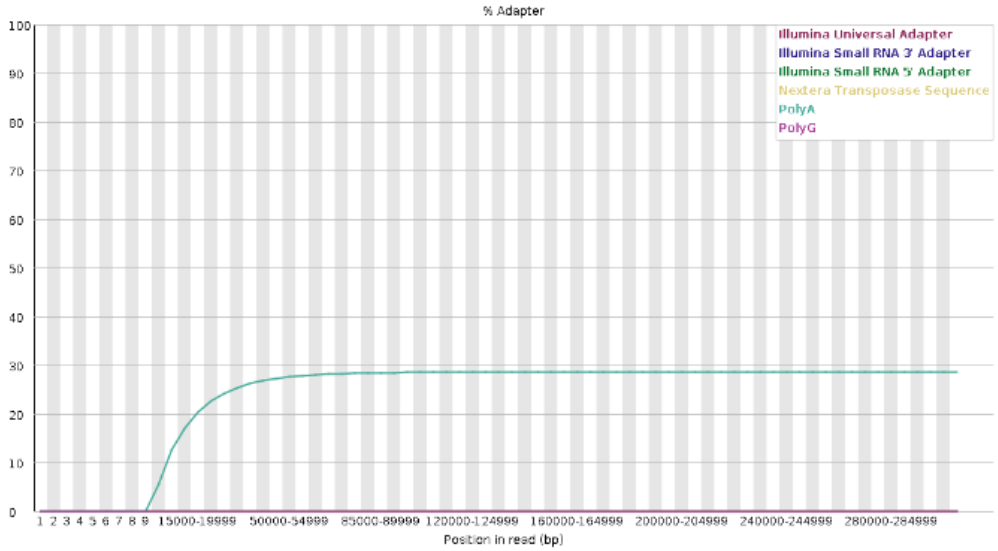
Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683637.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	235873
Bases Totales	4,5 GBP
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-305105
%GC	36

Calidad por secuencia de bases



Contenido del adaptador

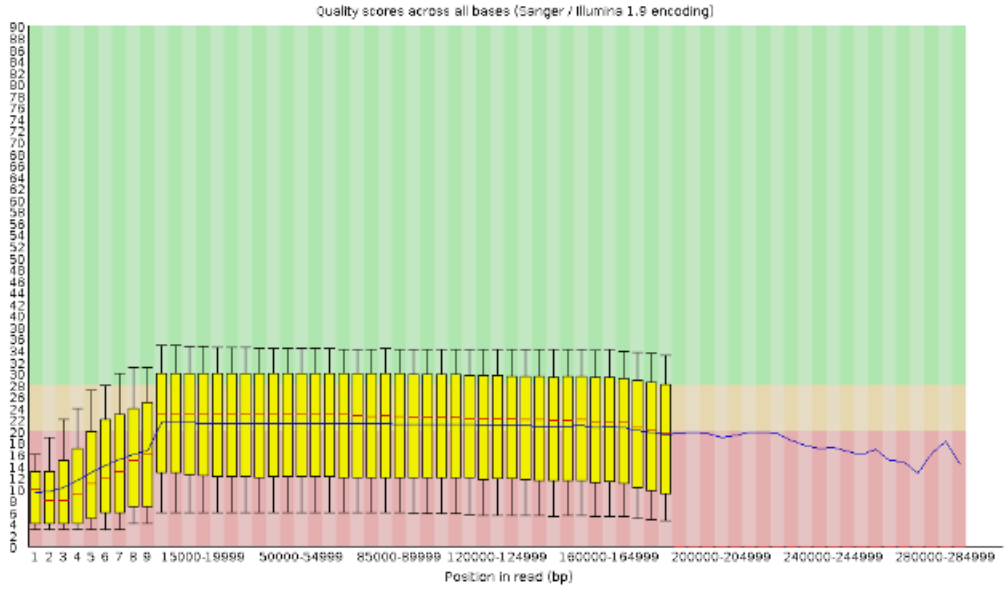


SRX8683638

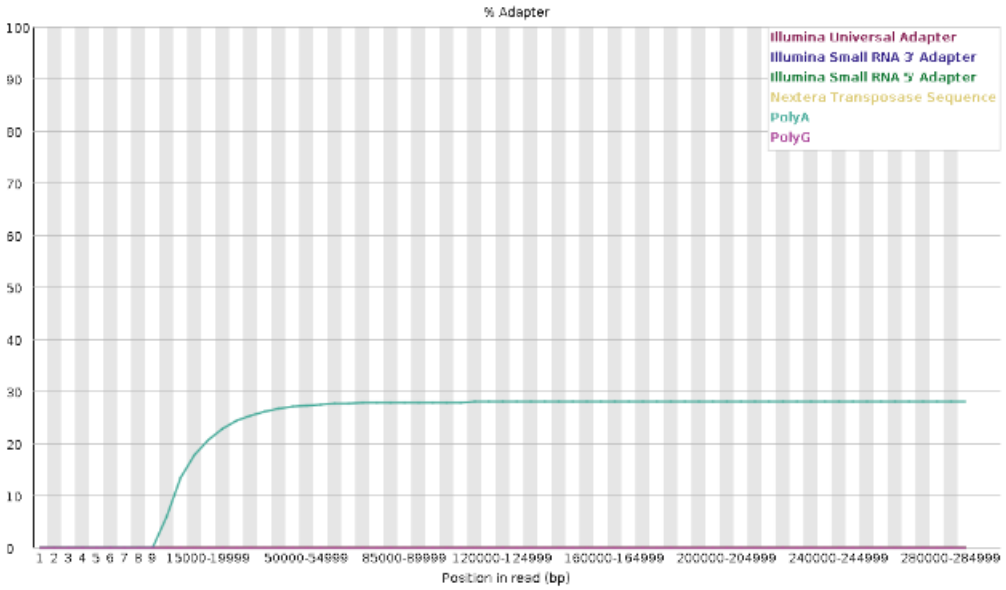
Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683638.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	464750
Bases Totales	8,2 GBP
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-289040
%GC	36

Calidad por secuencia de bases



Contenido del adaptador

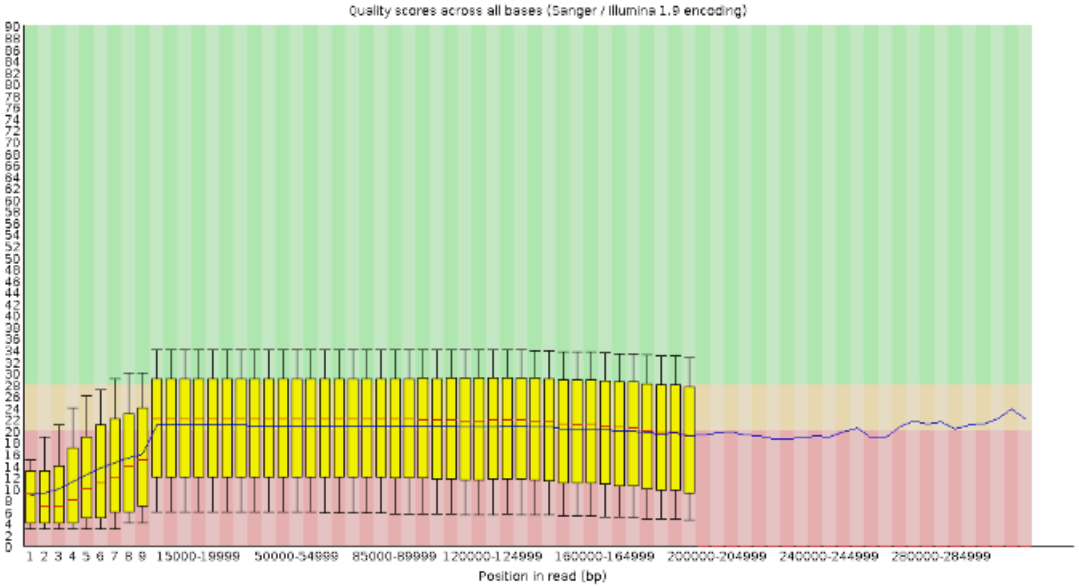


SRX8683639

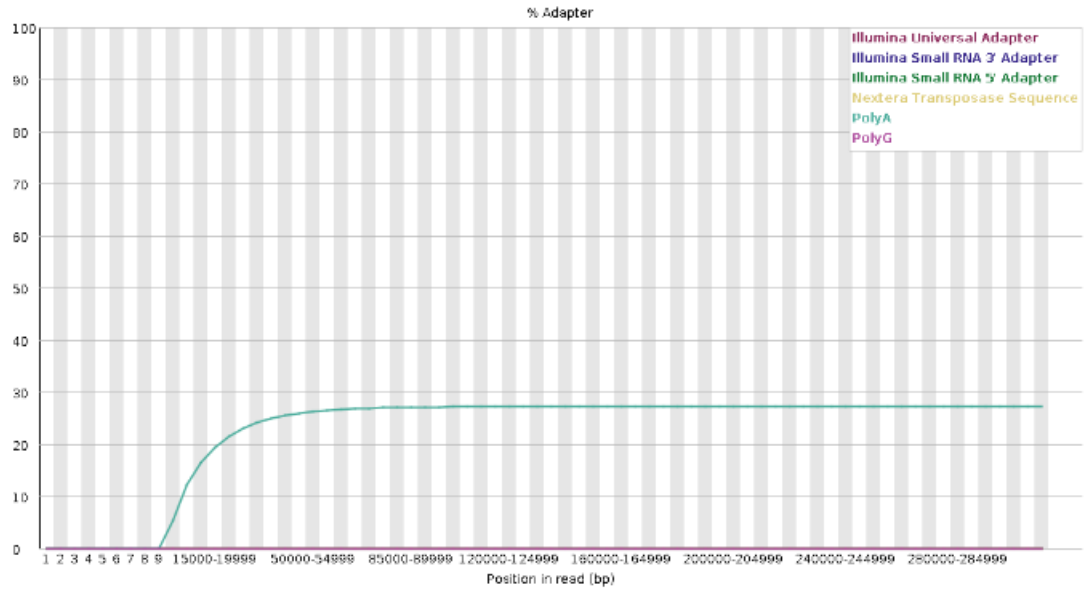
Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683639.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	292014
Bases Totales	5,4 GBP
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-312344
%GC	36

Calidad por secuencia de bases



Contenido del adaptador

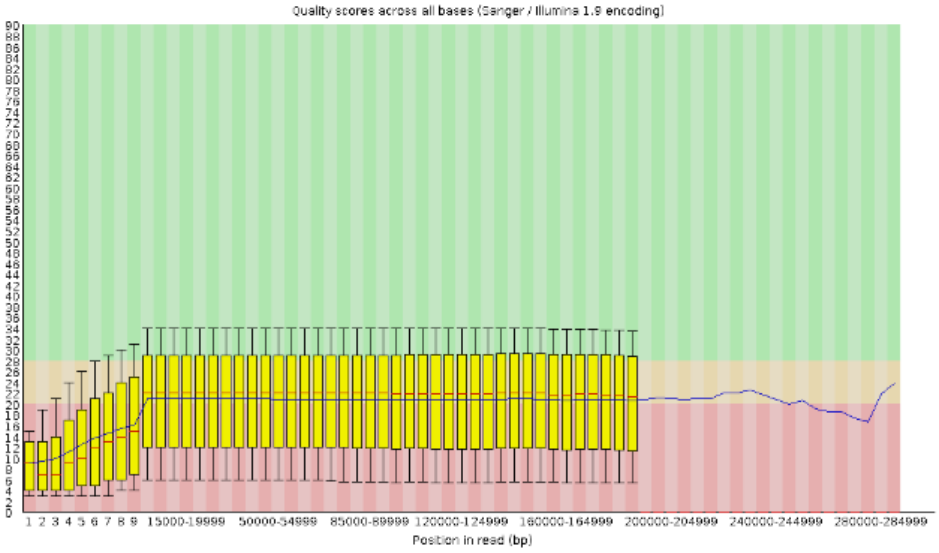


SRX8683640

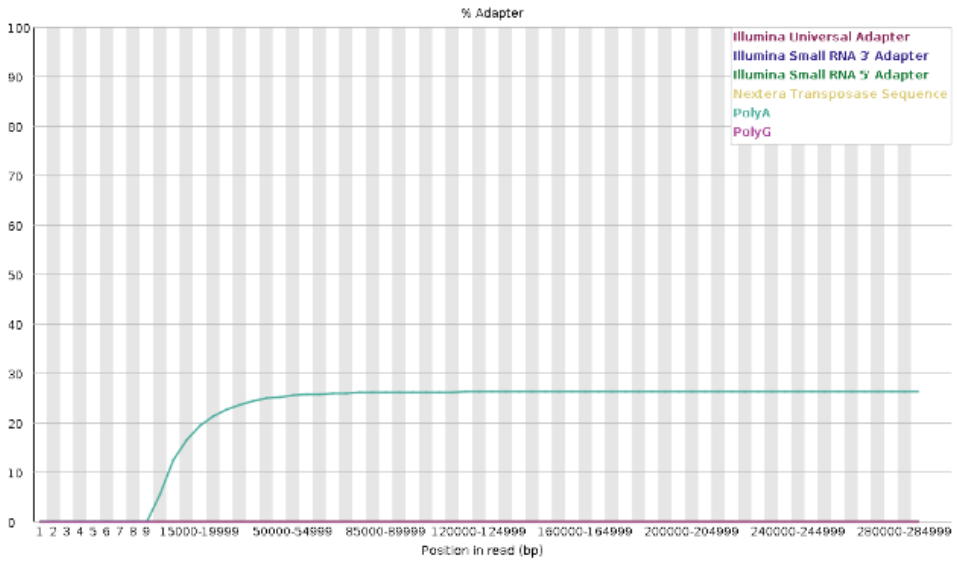
Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683640.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	447989
Bases Totales	7,6 GBP
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-286923
%GC	36

Calidad por secuencia de bases



Contenido del adaptador

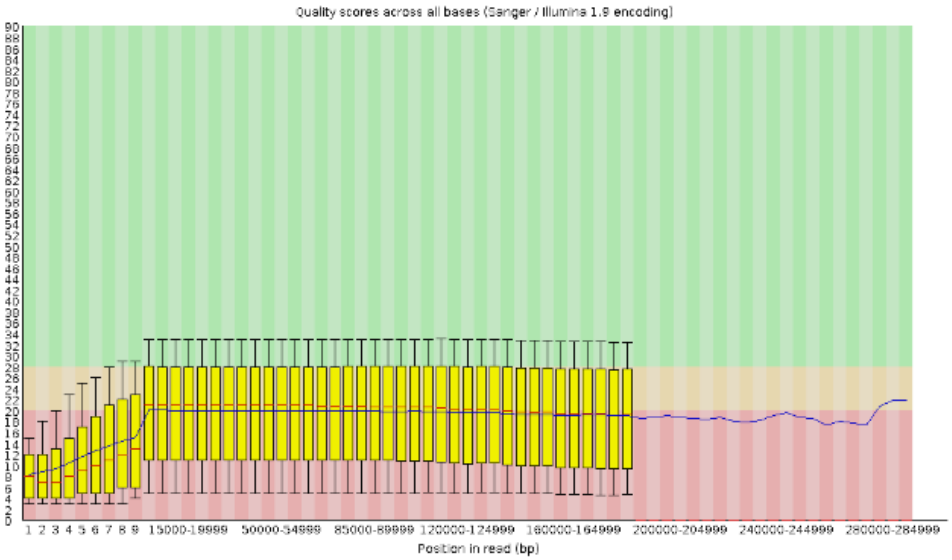


SRX8683641

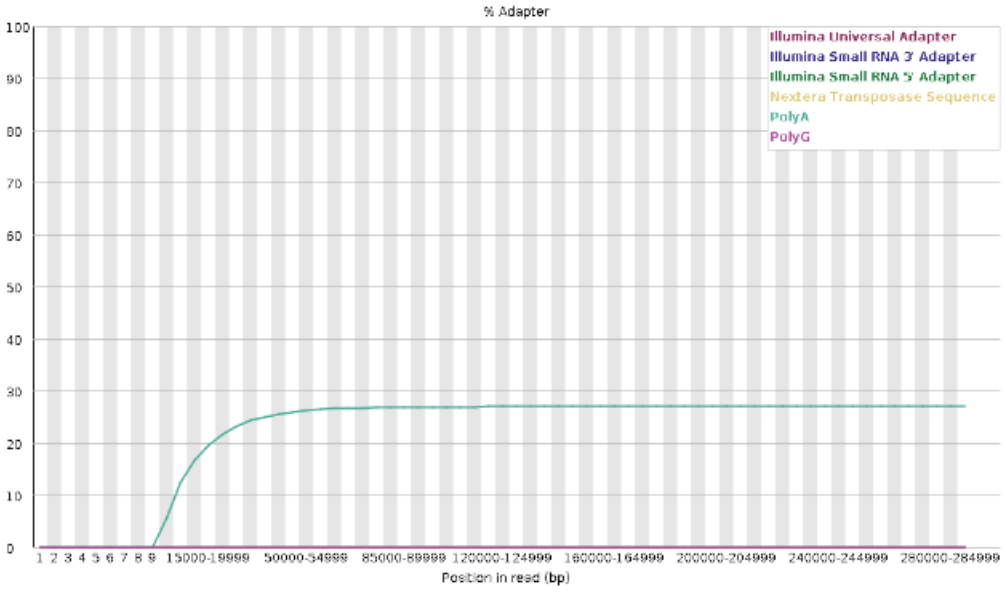
Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683641.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	388107
Bases Totales	6,9 GBP
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-289169
%GC	35

Calidad por secuencia de bases



Contenido del adaptador

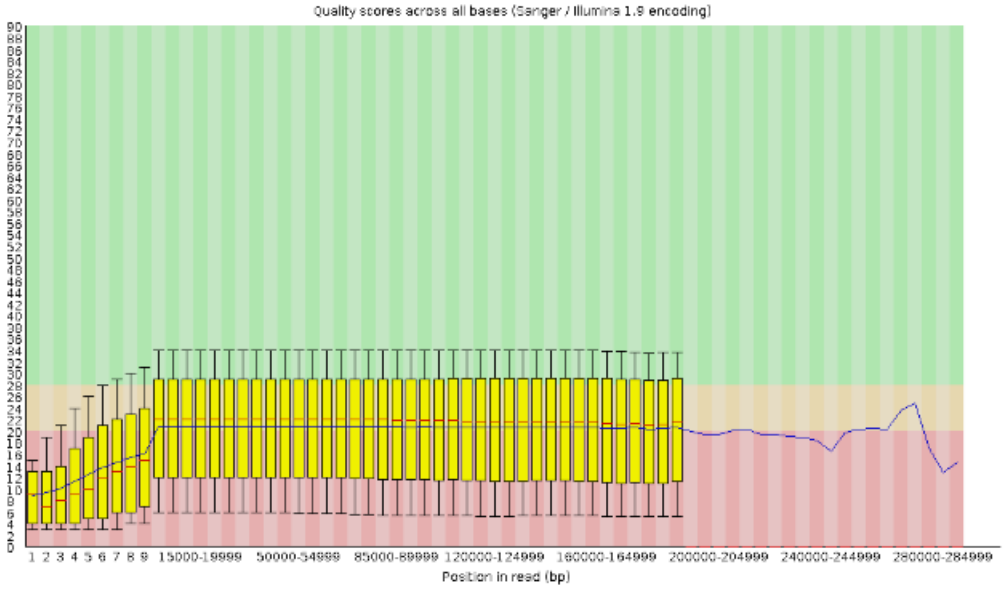


SRX8683642.

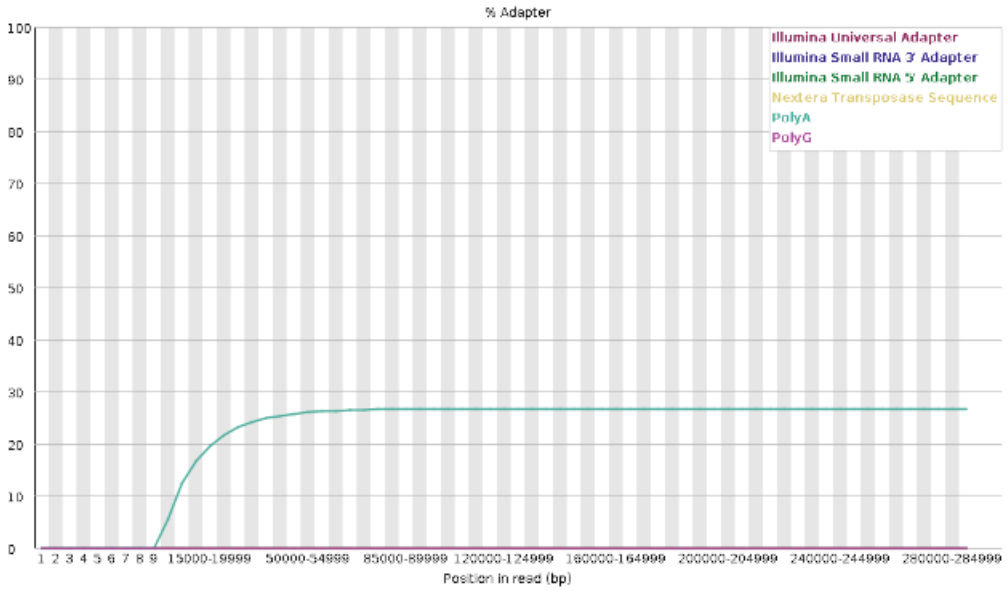
Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683642.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	536768
Bases Totales	9,4 GBP
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-285536
%GC	36

Calidad por secuencia de bases



Contenido del adaptador



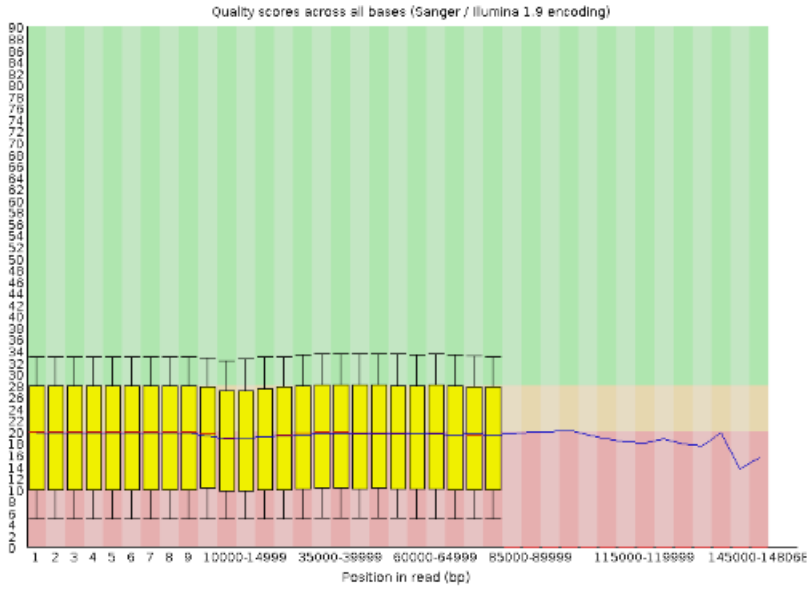
2.2) INFORMES DE CALIDAD DESPUES DEL PROCESO DE TRIMMING

SRX8683637

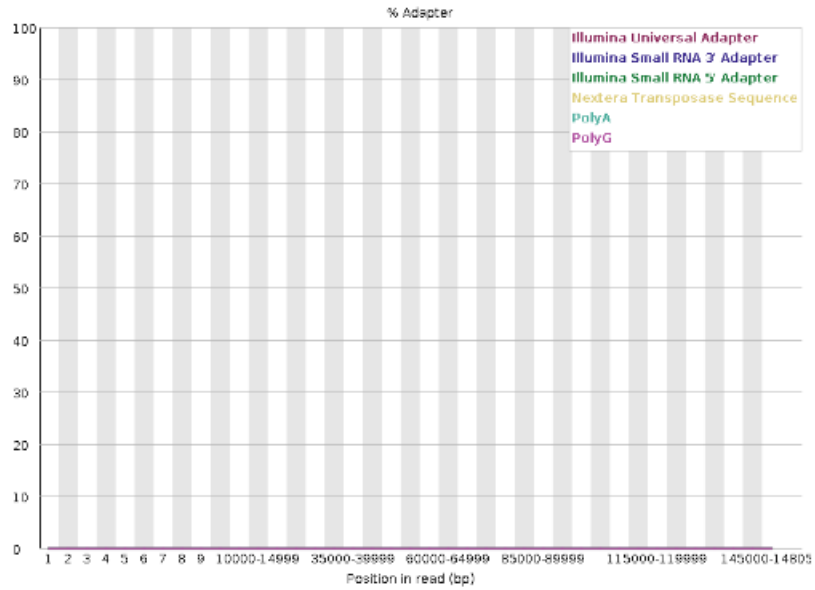
Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683637_cleanpn.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	226454
Bases Totales	636,7 Mbps
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-148068
%GC	38

Calidad por secuencia de bases



Contenido del adaptador

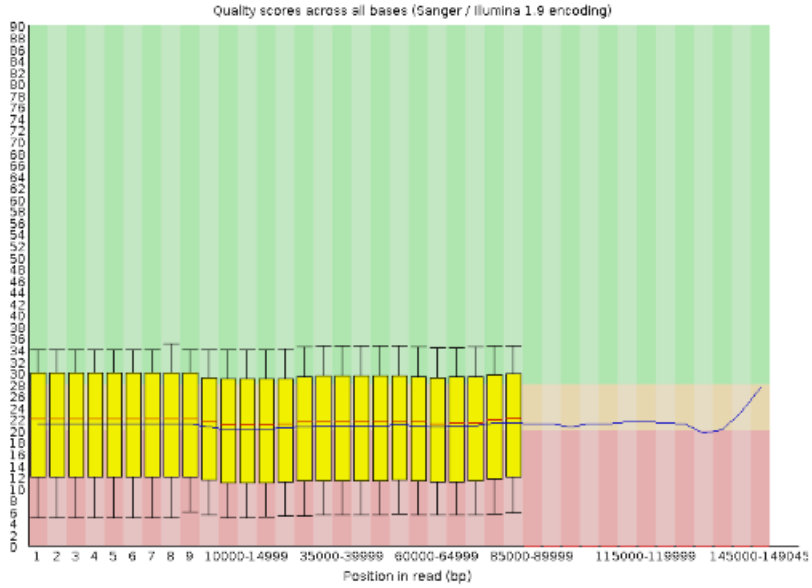


SRX8683638

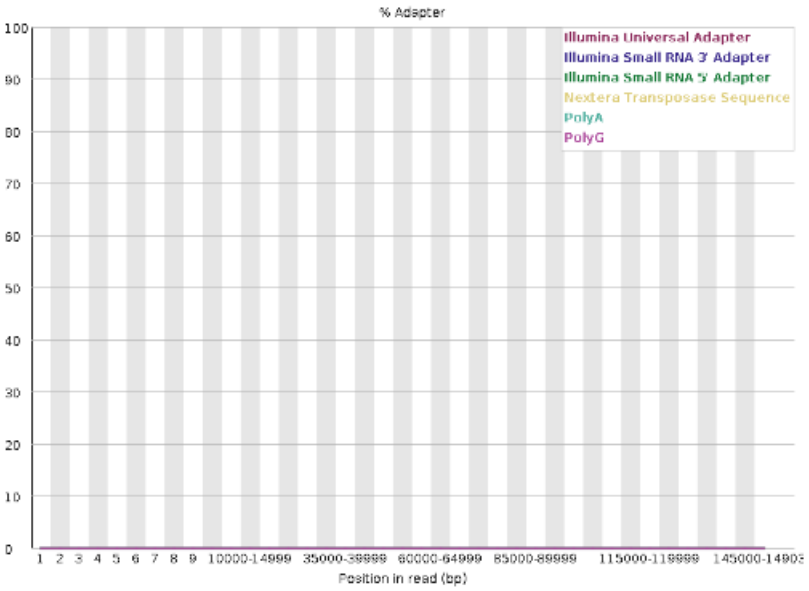
Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683638_cleanpn.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	445694
Bases Totales	1,2 GBP
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-149045
%GC	38

Calidad por secuencia de bases



Contenido del adaptador

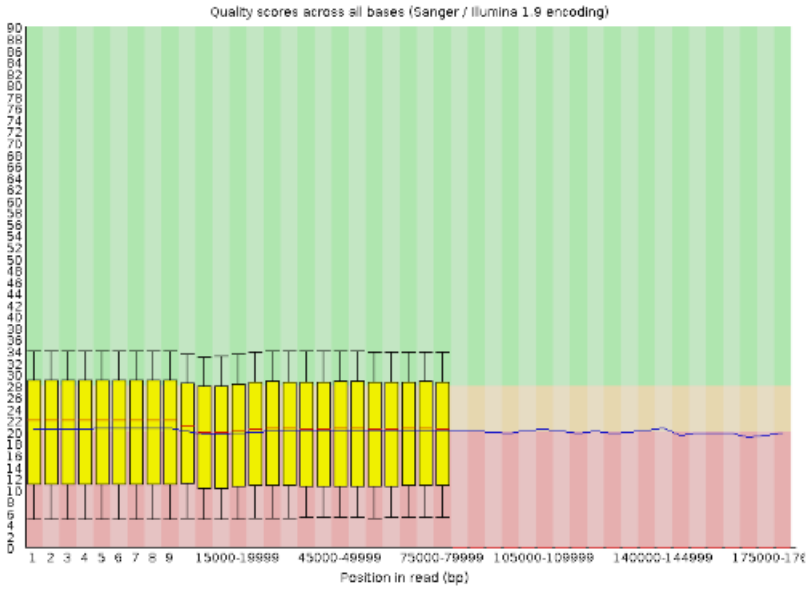


SRX8683639

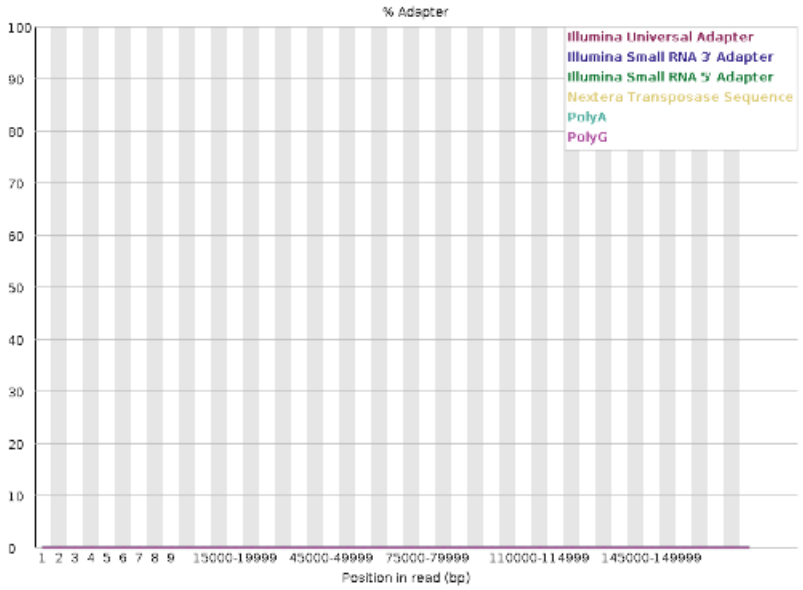
Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683639_cleanpn.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	288141
Bases Totales	772 Mbps
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-176530
%GC	38

Calidad por secuencia de bases



Contenido del adaptador

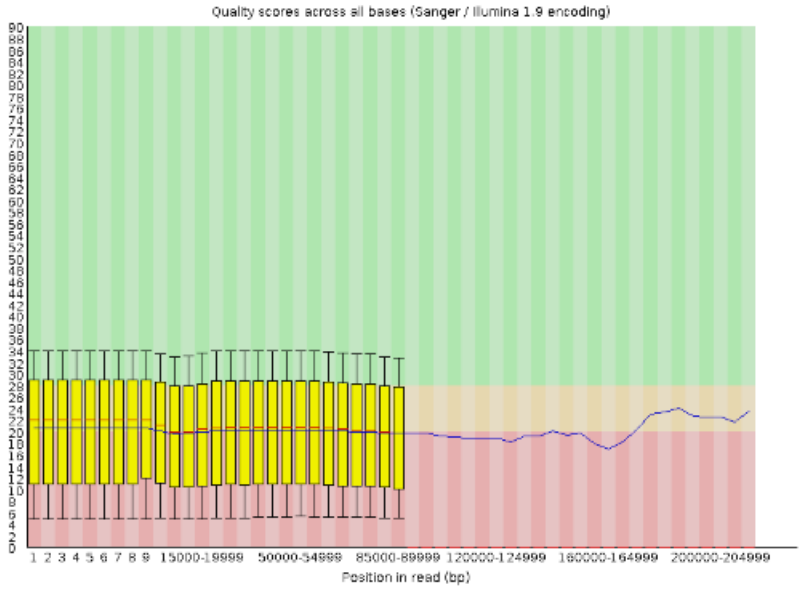


SRX8683640

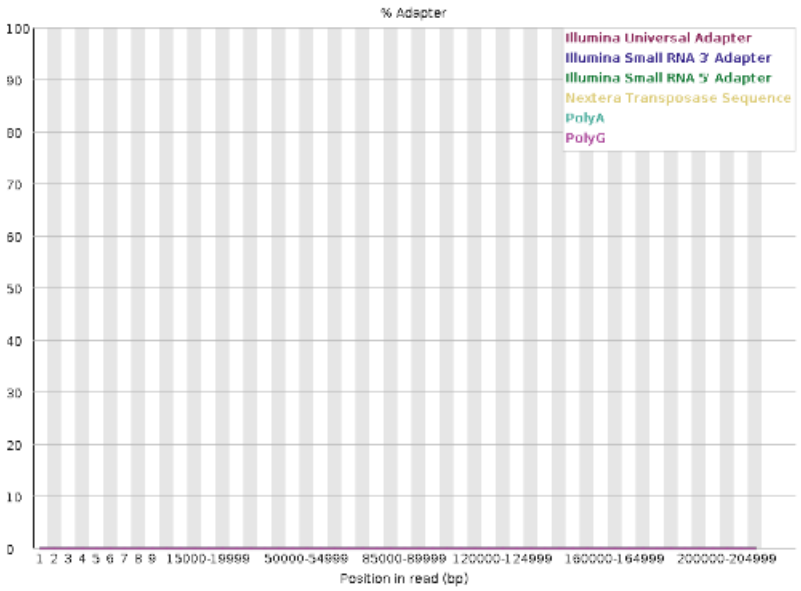
Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683640_cleanpn.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	429586
Bases Totales	1,1 GBP
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-211190
%GC	38

Calidad por secuencia de bases



Contenido del adaptador

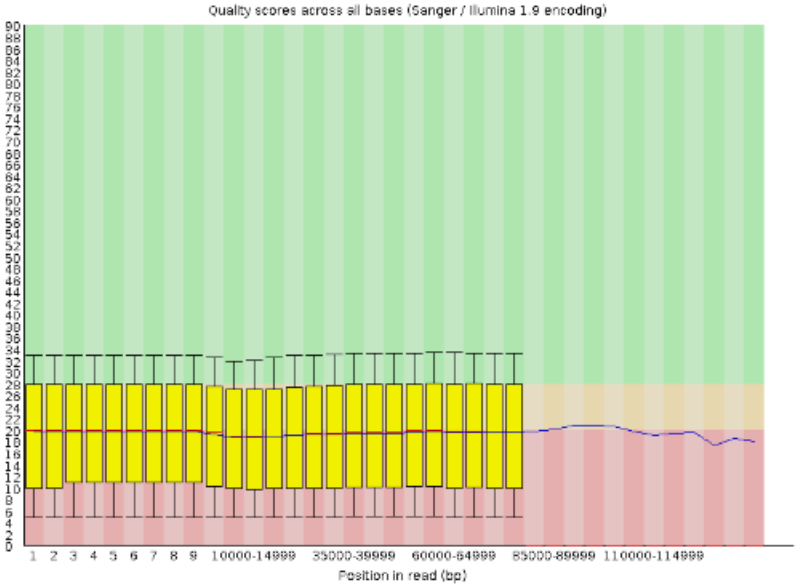


SRX8683641

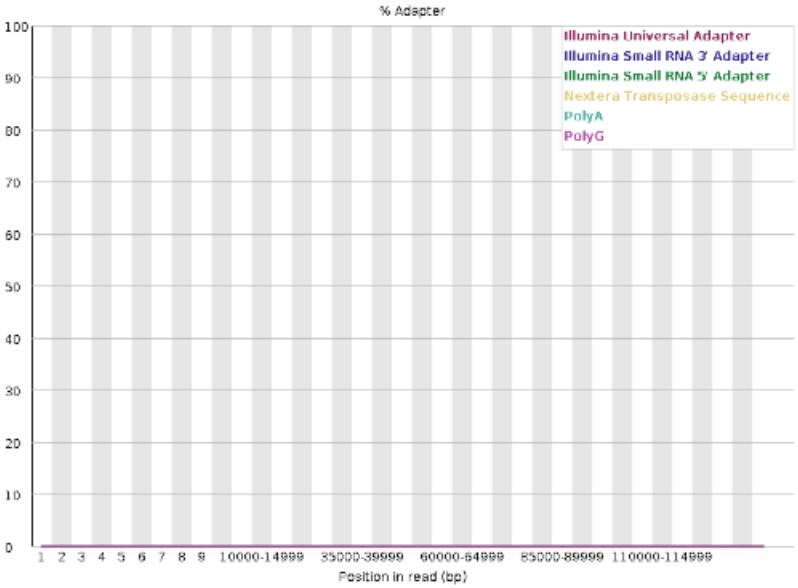
Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683641_cleanpn.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	372073
Bases Totales	1 GBP
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-139866
%GC	38

Calidad por secuencia de bases



Contenido del adaptador

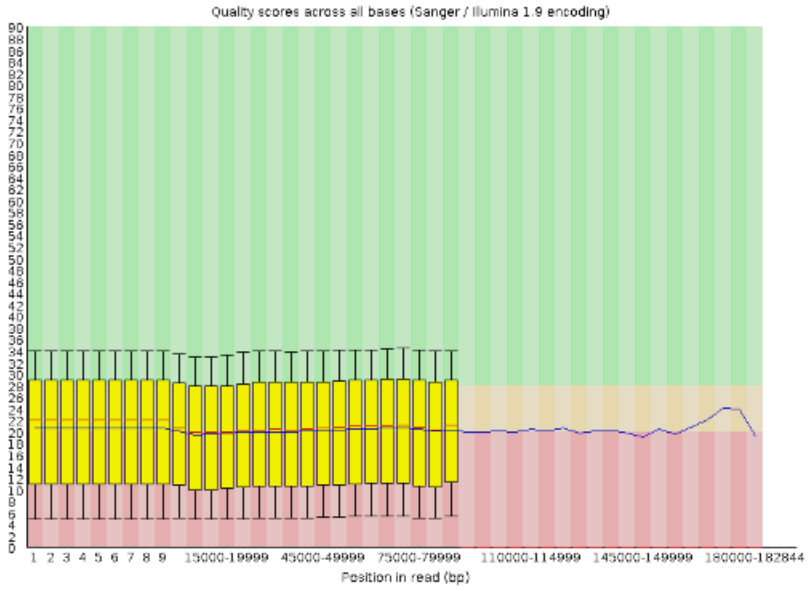


SRX8683642

Estadísticas Básicas

Medida	Valor
Nombre del archivo	SRX8683642_cleanpn.fastq
Tipo de archivo	Llamadas de base convencionales
Codificación	Sanger/Illumina 1.9
Secuencias totales	515703
Bases Totales	1,4 GBP
Secuencias marcadas como de mala calidad	0
Longitud de la secuencia	1-182844
%GC	38

Calidad por secuencia de bases



Contenido del adaptador

