

# Diseño e Implementación de un Pipeline Bioinformático para el Análisis Genómico en la Asociación VACTERL



Universitat  
Oberta  
de Catalunya



UNIVERSITAT DE  
BARCELONA

**Bernat Moreno Batlle**

MU Bioinf. i Bioest.

TFM Bioinformàtica Estadística y  
Aprendizaje Automático

**Tutor/a de TF**

Romina Astrid Rebrij

**Profesor/a responsable de la  
asignatura**

Carles Ventura Royo

16/01/2024



Universitat Oberta  
de Catalunya

uoc.edu



Aquesta obra està subjecta a una llicència de  
[Reconeixement-NoComercial-SenseObraDerivada 3.0  
Espanya de Creative Commons](https://creativecommons.org/licenses/by-nc-nd/3.0/es/)

<b>Título del trabajo</b>	<i>“Diseño e Implementación de un Pipeline Bioinformático para el Análisis Genómico en la Asociación VACTERL”</i>
<b>Nombre del autor:</b>	<i>Bernat Moreno Batlle</i>
<b>Nombre del consultor/a:</b>	<i>Romina Astrid Rebrij</i>
<b>Nombre del PRA:</b>	<i>Carles Ventura Royo</i>
<b>Fecha de entrega (mm/aaaa):</b>	<i>01/2024</i>
<b>Titulación o programa:</b>	<i>Máster en Bioinformática y Bioestadística</i>
<b>Área del Trabajo Final:</b>	<i>TFM Bioinformática Estadística y Aprendizaje Automático</i>
<b>Idioma del trabajo:</b>	<i>Castellano</i>
<b>Palabras clave</b>	<i>Pipeline bioinformático, Asociación VACTERL y secuenciación NGS</i>

### Resumen del Trabajo

En este proyecto, se ha creado un pipeline bioinformático en cuatro etapas esenciales: preprocesamiento, alineamiento, detección de variantes y anotación. Durante el preprocesamiento, se utilizaron herramientas como *FASTQC* y *Trimmomatic* para garantizar la calidad y limpieza de los datos genómicos. El alineamiento se llevó a cabo con *BWA* para mapear las secuencias a la referencia genómica. La detección de variantes se realizó mediante *FreeBayes*, identificando de manera precisa las variaciones genéticas. Se utilizó *SNPeff* en la anotación, proporcionando información detallada sobre el impacto de las variantes en genes específicos.

Los resultados revelaron genes con un impacto funcional significativo, como *EPHX2*, *FDFT1*, *TLR8* e *ILK*, sugiriendo su posible asociación con la condición VACTERL. Sin embargo, las conclusiones deben ser interpretadas con cautela. A pesar de los genes identificados, la naturaleza heterogénea y multifactorial de la condición VACTERL presenta un desafío considerable en la identificación de sus bases genéticas. Además, al basarse en datos de un solo paciente, los resultados no permiten afirmar de manera concluyente la validez en términos de causalidad y determinación de la enfermedad. La falta de replicación en un conjunto más amplio de pacientes limita la generalización de los resultados. Se destaca la necesidad de futuros estudios con una muestra más amplia y replicaciones adicionales para obtener conclusiones más robustas y representativas de la diversidad genética de esta compleja entidad clínica.

### Abstract

In this project, we have developed a comprehensive bioinformatics pipeline consisting of four key processes: preprocessing, alignment, variant detection, and annotation. The

methodology involved the use of tools such as FASTQC, Trimmomatic, BWA, FreeBayes, and *SNPeff*. Preprocessing ensured data quality using FASTQC and Trimmomatic. BWA facilitated alignment, mapping sequences to the reference genome. FreeBayes was employed for variant detection, accurately identifying genetic variations. For functional insight, SNPeff was utilized for annotation, providing detailed information on the impact of variants on specific genes.

The results unveiled genes with significant functional impact, including *EPHX2*, *FDFT1*, *TLR8*, and *ILK*, hinting at their potential association with the VACTERL condition. However, conclusions must be approached with caution. Despite the identification of specific genes, the heterogeneous and multifactorial nature of the VACTERL condition poses a substantial challenge in uncovering its genetic foundations. Moreover, relying on data from a single patient prevents definitive assertions regarding the causality and determination of the disease. The lack of replication in a broader patient cohort hinders the generalizability of the findings. Emphasizing the need for future studies with a larger sample size and additional replications, this work underscores the complexity of the genetic landscape underlying this clinical entity.

# ÍNDICE

1. Introducción .....	1
1.1 Contexto y Justificación .....	1
1.2 Descripción general .....	2
1.3 Objetivos .....	2
1.4 Impacto en sostenibilidad, ética social y diversidad .....	3
1.5 Enfoque y método que seguir .....	4
1.5.1 Preprocesamiento de datos .....	5
1.5.2 Alineamiento de secuencias .....	5
1.5.3 Detección de Variantes .....	6
1.5.4 Anotación de Variantes .....	6
1.6 Planificación .....	7
1.6.1 Tareas .....	7
1.6.2 Calendario .....	9
1.6.3 Hitos .....	10
1.6.4 Análisis de riesgos .....	10
1.7 Resultados esperados .....	10
2. La asociación VACTERL y la utilidad bioinformática .....	12
2.1 La asociación VACTERL .....	12
2.1.2 Epidemiología y etiología .....	12
2.2 Next Generation Sequencing (NGS) .....	16
2.2.1 Aplicaciones y tipos: .....	16
3. Materiales y métodos .....	18
3.1 Preprocesamiento .....	18
3.1.1 FASTQC .....	18
3.1.2 Trimmomatic .....	24
3.2 Alineamiento de secuencias .....	26
3.2.1 BWA .....	26
3.2.2 Samtools .....	27
3.3 Detección de variantes .....	27
3.3.1 <i>FreeBayes</i> .....	27
3.4 Anotación de variantes .....	29
3.4.1 <i>SnpEff</i> .....	29
3.5 Aplicación web .....	30
3.5.1 <i>Streamlit</i> .....	30
4. Resultados .....	34
5. Discusión .....	46

6.	Conclusiones .....	49
7.	Glosario .....	51
8.	Bibliografía.....	52
9.	ANNEXO 1: Informe <i>FASTQC</i> .....	61
10.	ANNEXO 2: Informe <i>FASTQC (trimmed)</i> .....	64
11.	ANNEXO 3: <i>SnpEff report</i> .....	67

## LISTA DE FIGURAS

Figura 1. Informe resumen del Fastqc Report. Izquierda "good data" y derecha "bad data". .....	19
Figura 2. Información básica sobre los datos. ....	20
Figura 3. Representación de la calidad de los nucleótidos (eje Y) para cada posición de la muestra (eje X). ....	20
Figura 4. Representación del número total de lecturas (eje Y) versus el puntaje promedio de calidad (eje X) de unos buenos datos (izquierda) y unos malos datos (derecha). ....	22
Figura 5. Porcentaje de bases nucleotídicas (eje Y) en cada posición de las lecturas (eje X) de unos buenos datos (izquierda) y unos malos datos (derecha). ....	22
Figura 6. Nombre de lecturas (eje Y) versus el porcentaje de GC (eje X) por lectura de unos buenos datos (izquierda) y unos malos datos (derecha). ....	23
Figura 7. Distribución de la longitud de las secuencias. ....	24
Figura 8. Porcentaje de lecturas repetidas de unos buenos datos (izquierda) y unos malos datos (derecha). ....	24
Figura 9. Diferencia entre el tipo de variantes y su significado. ....	28
Figura 10. Visualización de los resultados en la aplicación creada con información de cada variante, su impacto y la región donde actúa. ....	32
Figura 11. Visualización del archivo VCF de las variantes finales mediante la aplicación creada. ....	33
Figura 12. Visualización de ESEMBL para el gen EPHX2 obtenido con el enlace de la aplicación de la tabla 1. ....	33
Figura 13. Visualización de ESEMBL para el cromosoma 1 en la posición 1045707 obtenido con el enlace de la aplicación de la tabla 2. ....	34
Figura 14. Distribución de la calidad de los datos del proyecto. ....	35
Figura 15. Porcentaje de la composición de nucleótidos en los datos del proyecto. ...	36
Figura 16. Porcentaje de la composición de nucleótidos en los datos del proyecto después de <i>Trimmomatic</i> . ....	36
Figura 17. Niveles de duplicación de los datos del proyecto. ....	37
Figura 18. Información básica de los datos originales (izquierda) y los datos filtrados (derecha). ....	38
Figura 19. representación del porcentaje de variantes en las regiones del genoma. ..	44
Figura 20. Resultados para el gen 5S_rRNA para su impacto funcional y region. ....	45

# 1. Introducción

## 1.1 Contexto y Justificación

La asociación o síndrome VACTERL se define por la presencia de al menos 3 malformaciones, que se describen a continuación: "V: anomalías vertebrales, A: atresia anal, C: anomalías cardíacas, TE: fístula traqueoesofágica, R: anomalías renales, L: defectos en las extremidades" (1). Se trata de una asociación relativamente poco común en la que su descripción siempre ha sido problemática debido al limitado conocimiento y la diversidad de criterios al utilizar diferentes números y tipos de fenotipos para definirla (2). Es en este contexto, que el campo de la genómica y la secuenciación de última generación (*Next-Generation Sequencing* o NGS) ofrecen una oportunidad para mejorar la comprensión de esta condición y su detección.

El propósito del presente Trabajo Final de Máster es el diseño e implementación de un pipeline bioinformático, para la transformación de datos obtenidos mediante la secuenciación NGS de un paciente afectado por VACTERL en información genética significativa y relevante (3). Esta investigación se fundamenta en la necesidad de profundizar en la exploración genética de esta asociación y comprender el papel crucial que la genómica y la bioinformática pueden desempeñar en la identificación de variantes genéticas asociadas a enfermedades congénitas.

La bioinformática desempeña un papel crucial en esta investigación, ya que facilita el análisis de grandes volúmenes de datos genómicos generados a partir de la secuenciación NGS (4). Por lo tanto, el desarrollo de un pipeline bioinformático específico es fundamental para lograr los objetivos de este estudio. Esta herramienta, permitirá el procesamiento de datos de secuenciación NGS y la extracción de información relevante sobre las variantes genéticas que padece el paciente con la asociación VACTERL.



## 1.2 Descripción general

Este Trabajo Final de Máster implica la creación de un pipeline bioinformático personalizado, basado en la investigación bibliográfica, y consecuentemente la evaluación de las mejores herramientas disponibles para la realización de cada paso de este. El pipeline será usado con el objetivo de poder seleccionar e identificar las variantes genéticas que presenta un paciente con asociación VACTERL.

En añadidura, se desarrollará una aplicación web interactiva para el análisis de los resultados y su visualización. Este enfoque integral, mejorará la comprensión de los resultados y la aplicación clínica de los hallazgos genéticos en pacientes con asociación VACTERL.

## 1.3 Objetivos

Los objetivos de este trabajo titulado “*Diseño e Implementación de un Pipeline Bioinformático para el Análisis Genómico en la Asociación VACTERL*” se basan todos en torno a la realización del pipeline bioinformático inicialmente, por lo que, serían los siguientes:

Objetivo general:

- Desarrollar un pipeline bioinformático integral para analizar datos genómicos en pacientes con asociación VACTERL.

Objetivos específicos:

- Identificar los procesos necesarios, las herramientas y métodos más apropiados para cada paso del pipeline.
- Procesar datos de secuenciación NGS de un paciente con VACTERL.
- Detectar variantes genéticas relevantes asociadas a la asociación VACTERL.
- Desarrollar una aplicación web interactiva para visualizar los resultados del pipeline.

## 1.4 Impacto en sostenibilidad, ética social y diversidad

Para realizar este Trabajo de Fin de Máster, solo se han usado datos ya computarizados por lo que no ha habido un impacto significativo referente a la sostenibilidad. Si bien pues, no se han usado residuos como los que se podrían generar en laboratorios o en otro tipo de estudios, se podría llegar a pensar cómo se generaron los datos iniciales provenientes de la paciente o el consumo energético que ha supuesto el desarrollo del proyecto.

Para llegar a obtener la secuenciación de la paciente se tuvo que recoger una muestra inicial de esta, y posteriormente, en el laboratorio prepararla y proceder con la NGS. Este proceso inicial de todo el proyecto supondría un coste negativo en cuanto a la sostenibilidad referente a los residuos posibles generados y a la afectación ecológica, si estos no fueron tratados de forma adecuada tal como manda la normativa que regula la gestión de los residuos en laboratorios es la Ley 7/2022, de 8 de abril (5). En cuanto al proyecto en sí de final de máster, podría considerarse un impacto negativo de consumo de energía al contar que se ha trabajado prácticamente cada día una media de 7 horas en un ordenador. Tal como se comenta en el ODS 7 (Objetivo de Desarrollo Sostenible 2023, ONU), el consumo de energía es el factor que más contribuye al cambio climático, ya que representa alrededor del 60 por ciento del total de las emisiones globales de gases de efecto invernadero (6), pero aun así, hoy en día se puede contar con un acceso energético sostenible y bien restablecido.

Por ello, en este trabajo, en cuanto a sostenibilidad podría haber un posible impacto negativo en el inicio de todo el estudio en la fase de laboratorio, pero, en la que concierne al trabajo realizado para este proyecto, no se han considerado relevantes para tenerlos en cuenta ya que no está en el diseño de este. Sí pero, que se ha realizado un buen uso de la energía y electricidad en cuando al uso del ordenador y la energía que este supone.

Relacionado con el aspecto ético y responsabilidad social, no se considera significativo ningún impacto negativo en este proyecto al ser tan técnico. Aun así, cabe destacar la ley de la propiedad intelectual regulada por el real Decreto Legislativo 1/1996 del 12 de abril (7), la cual se ha respetado y tenido en cuenta en este proyecto en todo momento, por ende, en todo momento, se ha citado

toda la bibliografía buscada y utilizada para evitar el plagio. En añadidura, como todo proyecto científico, podría tener un impacto positivo relacionado con el aporte de nueva información y datos, con el objetivo de investigación, en este caso de una enfermedad poco conocida y con poca información de ella.

Por último, no olvidar de dónde salió el archivo inicial resultado de la NGS con el que se ha trabajado. Este, proviene de la secuenciación, por lo que su origen es de una muestra la cual esta proviene de una paciente. A pesar de que en este Trabajo de Fin de Máster en concreto, se ha basado a partir del archivo inicial, y por lo tanto en su diseño no se contempla el uso de ninguna muestra ni su procesamiento, es importante destacar que en todo momento se ha respetado la privacidad de la paciente y el uso de sus datos.

En conclusión, en el diseño de este trabajo no se tuvieron en cuenta impactos negativos relacionados con la ética, la diversidad y la sostenibilidad debido a ser tan técnico. A pesar de ello, se han considerado aspectos como el gasto de energía, el uso adecuado de las fuentes de información y que en todo momento se estaba trabajando bajo unos datos provenientes de una paciente, por lo que se ha mantenido la confidencialidad y respeto hacia ella.

### 1.5 Enfoque y método que seguir

Un pipeline bioinformático es un conjunto de pasos y herramientas utilizadas para procesar datos biológicos, como los obtenidos de la secuenciación genómica, con el objetivo de convertir datos crudos en información significativa e interpretable (8). En este contexto, este Trabajo Final de Máster se centra en el desarrollo de un pipeline bioinformático para el análisis de los datos de secuenciación NGS de una paciente de seis meses diagnosticada con asociación VACTERL que provienen del estudio público: *"Discovering Genotype Variants in an Infant with VACTERL through Clinical Exome Sequencing: A Support for Personalized Risk Assessment and Disease Prevention"* publicado en la revista *Pediatric Reports* en 2021 [SRX9057024] (3).

Este pipeline, implica los siguientes pasos, cada uno de los cuales, requiere herramientas específicas:

### 1.5.1 Preprocesamiento de datos

El primer paso en esta investigación implica una evaluación crítica de la calidad de los datos en bruto (*raw data*) obtenidos durante el proceso de secuenciación. Esta fase tiene como objetivo principal eliminar secuencias de baja calidad o secuencias que podrían ser contaminantes, como los adaptadores. Para llevar a cabo esta tarea, se utilizan un conjunto de herramientas bioinformáticas ampliamente reconocidas y establecidas (9).

En primer lugar, se utiliza la herramienta *FASTQC*, conocida por su capacidad para generar informes visuales detallados y de fácil comprensión que proporcionan una visión completa de la calidad de los datos iniciales. Este análisis es fundamental para identificar cualquier problema de calidad en las secuencias que podría afectar a la precisión de los resultados posteriores (10).

Posteriormente, se emplea *Trimmomatic*, una herramienta especializada en la identificación y eliminación de adaptadores presentes en las secuencias, así como en la mejora de la calidad global del conjunto de datos. Esto garantiza que los datos que se pasan a las etapas siguientes del *pipeline* sean de la más alta calidad posible (11). Otras herramientas posibles serían, por ejemplo, *fastP*, como se menciona en Chen et al., 2018.

### 1.5.2 Alineamiento de secuencias

Para comparar el ADN de la muestra secuenciada con su secuencia de referencia, es necesario para cada lectura identificar la parte correspondiente de esta en los datos de secuenciación. Este proceso se conoce como alineación o emparejamiento de las lecturas con la secuencia de referencia del genoma humano (GRCh38/hg38) (13). Una vez que se haya completado esta operación, es posible identificar las variaciones presentes en la muestra.

La herramienta utilizada para la identificación es el paquete de alineamiento de *Burrows-Wheeler* (BWA), un paquete de alineación de lecturas que permite alinear secuencias cortas frente al genoma de referencia permitiendo producir huecos (gaps). BWA se basa en la búsqueda de la inversa transformada de

*Burrows-Wheeler* (BWT) para alinear eficazmente las lecturas de secuenciación contra una gran secuencia de referencia, permitiendo variaciones e inserciones (14). Con el resultado del alineamiento, se obtiene un archivo BAM que puede ser manipulado a través del paquete *SAMtools* para su posterior análisis. Otras herramientas posibles son, por ejemplo, *Bowtie*, como se menciona en Langmead, 2010).

### 1.5.3 Detección de Variantes

La detección de variantes, o "*variant calling*", es la fase que implica la identificación y caracterización de las variantes genéticas, como los polimorfismos de nucleótido único (SNPs) o las inserciones y deleciones (indels) (16).

La herramienta bioinformática *FreeBayes* se utilizará para realizar la detección de variantes en este proyecto, ya que es una herramienta ampliamente empleada en la detección de variantes en secuencias genómicas. Es conocida por su alta precisión y su capacidad para detectar una variedad de variantes genéticas y es ampliamente utilizada en la comunidad científica y ofrece la flexibilidad de personalizar los parámetros de acuerdo con las necesidades del estudio (17,18).

### 1.5.4 Anotación de Variantes

La anotación de variantes, que implica asignar información funcional a las variantes del ADN basada en la nomenclatura estandarizada de la Sociedad de Variación del Genoma Humano (HGVS), es un desafío fundamental en el análisis de datos de NGS que ha llevado al desarrollo de numerosos algoritmos bioinformáticos (19) .

Las posibles herramientas que permiten esta anotación de variantes son por ejemplo *Annovar* y *Variant Effect Predictor* (VEP) o SnpEff. En este caso, se ha optado por utilizar SnpEff, que es una herramienta eficiente y ágil para anotar las consecuencias funcionales de las variaciones genéticas a partir de datos de secuenciación de alta capacidad. Además, SnpEff proporciona informe complementario que ayuda con la comprensión de las variantes descritas (20,21).

Estas fases son cruciales para el análisis de los datos genéticos del paciente con síndrome VACTERL, y las herramientas seleccionadas se han elegido para garantizar resultados precisos y significativos en cada etapa del pipeline.

Finalmente, con el fin de presentar los resultados y facilitar la visualización y el análisis de los resultados generados después de la ejecución del pipeline, se desarrollará una pequeña aplicación web interactiva, ubicada en el servidor GitHub, a través de la herramienta de trabajo *Streamlit*. Esta herramienta utiliza el lenguaje de programación Python, lo que proporciona comodidad y experiencia en este lenguaje.

## 1.6 Planificación

### 1.6.1 Tareas

#### Etapas 1: Identificación de las Herramientas y Métodos Óptimos:

- Tarea 1: Investigar las mejores prácticas en el campo de la bioinformática para el análisis de secuencias NGS.
- Tarea 2: Examinar las herramientas disponibles para cada fase del pipeline e identificar las más adecuadas.

#### Etapas 2: Documentación e Investigación Inicial sobre la asociación VACTERL:

- Tarea 1: Realizar una revisión exhaustiva de la bibliografía científica existente relacionada con la asociación VACTERL para comprender sus características genéticas y clínicas.
- Tarea 2: Identificar las publicaciones más relevantes sobre el tema y recopilar la información necesaria para contextualizar la investigación.

#### Etapas 3: Preprocesamiento de Datos:

- Tarea 1: Utilizar la herramienta *FASTQC* para evaluar la calidad de los datos en bruto.

- Tarea 2: Utilizar *Trimmomatic* para eliminar adaptadores y mejorar la calidad de los datos.

#### Etapas 4: Alineamiento de Secuencias:

- Tarea 1: Utilizar BWA para el alineamiento de las lecturas de secuencias con la secuencia de referencia.
- Tarea 2: Elaborar la PAC 2: Desarrollo y seguimiento del proyecto. Fase1 (20/11/2023)

#### Etapas 5: Detección de Variantes:

- Tarea 1: Utilizar *FreeBayes* para detectar y caracterizar las variantes genéticas presentes en la muestra.

#### Etapas 6: Anotación de Variantes:

- Tarea 1: Utilizar *SnpEff* para asignar información funcional a las variantes.
- Tarea 2: Elaborar la PAC 3: Desarrollo y seguimiento del proyecto. Fase1 (23/12/2023)

#### Etapas 7: Desarrollo de la Aplicación Web Interactiva:

- Tarea 1: Utilizar *Streamlit* en *Python* para desarrollar la aplicación web interactiva.
- Tarea 2: Optimizar la experiencia del usuario y la visualización de los resultados.
- Tarea 3: Implementar la aplicación en un servidor web.

#### Elaboración de la Memoria y Presentación Virtual:

- Tarea 1: Redacción de la memoria y presentación.
- Tarea 2: Preparación y ensayo de la presentación.

## 1.6.2 Calendario

	PAC 1: Plan de trabajo		PAC 2: Desarrollo y seguimiento del proyecto. Fase 1					PAC 3: Desarrollo y seguimiento del proyecto. Fase 2					PAC 4: Cierre de la memoria y presentación			PAC 5: Defensa pública del trabajo	
	27-16/10		17-23/10	24-30/10	31-06/11	7-13/11	14-20/11	21-27/11	28/11-04/12	05/12-11/12	12-18/12	19-23/12	25-31/12	01/01-07-01	07-14/01	14-21/01	22/01-02/02
Plan de Trabajo																	
Etap 1. Investigación de Mejores Prácticas en Bioinformática																	
Etap 1. Selección de Herramientas Óptimas																	
Etap 2. Recopilación de Información Clave VACTERL																	
Etap 3. Evaluación de Calidad de Datos con FastQC																	
Etap 3. Preprocesamiento de Datos con Trimmomatic																	
Etap 4. Alineamiento de Secuencias con BWA																	
Etap 5. Detección de Variantes con FreeBayes																	
Etap 6. Anotación de Variantes con SnpEff																	
Etap 7. Desarrollo de Aplicación Web con Streamlit																	
Etap 7. Optimización de la Experiencia del Usuario																	
Redacción de la memoria y presentación																	
Elaboración y aprendizaje de la presentación																	
Defensa pública del TFM (Trabajo de Fin de Máster)																	



### 1.6.3 Hitos

Descripción	Fecha
PAC 1: Plan de trabajo	16/10/2023
PAC 2: Desarrollo y seguimiento del proyecto. Fase 1	20/11/2023
PAC 3: Desarrollo y seguimiento del proyecto. Fase 2	23/12/2023
PAC 4: Cierre de la memoria y presentación	14/01/2024
PAC 5: Defensa pública del trabajo ante una Comisión Evaluadora	02/02/2024

### 1.6.4 Análisis de riesgos

Riesgos	Consecuencias
Complejidad Técnica	El uso de métodos y herramientas bioinformáticas complejas puede causar retrasos inesperados en las etapas de alineación y detección de variantes, e incluso podría dar lugar a errores técnicos que requerirían tiempo adicional para su resolución.
Problemas de Compatibilidad de Herramientas	Si las herramientas y software utilizados no son completamente compatibles y no pueden integrarse de manera eficaz, esto podría alargar los tiempos de implementación y dificultar la cohesión del pipeline.
Limitaciones en la Aplicación Web	Si surgen dificultades en el desarrollo de la aplicación web, esto podría resultar en problemas de visualización y en la comunicación efectiva de los resultados, lo que podría afectar negativamente la experiencia del usuario.

## 1.7 Resultados esperados

### A. Plan de Trabajo:

- Documento inicial en formato PDF que contenga las pautas y los tiempos estimados de ejecución de todas las tareas necesarias para el desarrollo del trabajo y el cumplimiento de los objetivos.

### B. Memoria:

- Documento en formato PDF que detalle toda la investigación, desarrollo, resultados y conclusiones obtenidas a lo largo del trabajo de fin de máster, así como el código implementado.

C. Producto:

- Enlace de acceso a la aplicación web desarrollada en *Streamlit*.
- Acceso a un repositorio público que permita acceder al código desarrollado.

D. Presentación Virtual:

- Presentación en formato PPT para exponer el trabajo realizado.

## 2. La asociación VACTERL y la utilidad bioinformática

### 2.1 La asociación VACTERL

En 1970 se denominó por primera vez la asociación VATER, la cual incluía la aparición no aleatoria de un grupo de malformaciones congénitas: defectos vertebrales (V), atresia anal (A), fístula traqueo-esofágica (TE) con o sin atresia esofágica y displasia renal (R) (22,23). Debido a presentarse estas anomalías juntas con más frecuencia de lo que cabría esperar por azar, se empezó a nombrar esta manifestación conjunta como asociación. Más tarde, por observaciones de múltiples pacientes; investigadores decidieron añadir la malformación cardíaca (C) y las anomalías de las extremidades (L) a la descripción de este conjunto (24,25). Por lo que, la patología finalmente pasó a ser la Asociación VACTERL (AV), definida en términos generales como un conjunto de anomalías congénitas que afectan a múltiples sistemas corporales.

Sin embargo, es importante destacar que el término “asociación”, se refiere a la coocurrencia no aleatoria de un grupo de múltiples características clínicas, pero no implica una única causa subyacente o unificadora, por lo que, a falta de esta causa, la afección todavía no se considera un síndrome (23,26).

Los afectados por la AV pueden presentar, a de más, otras anomalías congénitas, por lo que incrementa la causa de muerte fetal, mortalidad infantil o afecta en la morbilidad o en el incremento del riesgo de otras patologías como cáncer, diabetes o defectos cardiovasculares (27). Por ello, se está dando mera importancia a entender esta patología, sus procesos patológicos y la causa, avanzando en la secuenciación del material genético de los afectados.

#### 2.1.2 Epidemiología y etiología

La prevalencia de la AV, basada en estudios epidemiológicos en Europa y Estados Unidos, oscila entre 1/10000 y 1/40000 neonatos/niños aproximadamente (28–30). La mortalidad, comparada con los años antes del 1996 (20 %) se ha reducido, siendo así aproximadamente un 5.6% hoy en día, gracias a los avances quirúrgicos, de diagnóstico y el cuidado médico (31).

Una explicación de la agrupación de rasgos es la idea de que las malformaciones surgen tempranamente durante la blastogénesis. La blastogénesis se

caracteriza por ser el proceso de determinación, por lo que mediante gradientes de moléculas de señalización las células se agrupan y dependiendo de dónde se coloquen, se determinará qué serán y su función futura(32,33) (34,35). En este proceso pues, se tienden a dar lugar a anomalías politópicas o defectos congénitos que afectan a múltiples sistemas orgánicos, nombrando así un defecto del campo del desarrollo (36).

Se puede estipular sobre que la AV tiene una base genética gracias a estudios etiológicos que siguen encontrando causas genéticas (37–39), pero todavía no está determinada una base general e/o igualitaria para todos o la mayoría de afectados (23). Por este motivo, se asume que existe una heterogeneidad etiológica, sugiriendo clústeres de subgrupos con diferentes causas dentro de la AV (40–42) (43–45). Sin embargo, otros estudios, identifican que, además, puede haber causas relacionadas con factores de riesgo maternos como la diabetes, las técnicas de reproducción asistida, enfermedades crónicas, exceso de progesterona-estrógenos, etc.(46,47) (25,39), demostrando así, la falta de una única causa para la AV y la heterogeneidad que presenta.

En resumen, a modo de explicación etiológica, la AV incluye varias malformaciones relativamente graves. Estas malformaciones se asociaban a una alta tasa de letalidad en el período neonatal hasta hace relativamente poco tiempo. Así pues, para que el trastorno se repita en la población, las causas pueden incluir eventos de *novo*, mutaciones recesivas no comunes o una etiología multifactorial con múltiples interacciones genéticas y ambientales (48).

Para la AV, no hay un acuerdo estandarizado para su diagnóstico debido a la falta de su causalidad, por este motivo, solo se basa en las pruebas prenatales para observar si el feto presenta anomalías físicas (Tabla 1) y en añadidura, un estudio genético.

Tabla 1. Frecuencia estimada de las anomalías de VACTERL en % y pruebas sugeridas para los pacientes (además de un cuidadoso examen físico) en los que se sospeche una asociación VACTERL (26)(49).

Característica	Frecuencia	Prueba diagnóstica
Anomalías vertebrales	60-80 %	Radiografía, ecografía o resonancia magnética de la columna vertebral.
Atresia anal	55-90 %	Exploración física, ecografía abdominal para anomalías genitourinarias
Malformaciones cardíacas	40-80%	Ecocardiograma
Fístula traqueo-esofágica	50-80 %	Examen físico/observación
Anomalías renales	50-80 %	Ecografía renal
Anomalías de las extremidades	40-50 %	Examen físico, radiografías

Investigadores, defienden que, para diagnosticar, debe ser mandatorio que se presenten mínimo tres de las anomalías físicas VACTERL (23,48,50). Sin embargo, por falta de pruebas concluyentes sobre los genes implicados, el cuadro genético de diagnóstico no está definido, aun así, estudios genéticos han encontrado mutaciones y genes afectados similares entre pacientes(51-53) (48,54,55) (Tabla 2):

- ZIC3: Gen del factor de transcripción localizado en Xq26 que participa en el desarrollo temprano. Suele dar lugar a un espectro de defectos de asimetría izquierda-derecha, anomalías cardíacas complejas, anomalías esplénicas y hepatobiliares y mal posición intestinal (56). Por lo tanto, diferentes mutaciones en este gen ZIC3, pueden conllevar a defectos cardíacos, heterotaxia ligada al cromosoma X o atresia anal. Estudios genéticos han comprobado que las mutaciones en ZIC3 suelen estar presentes en la AV(51,57-59) (23,55,56,60).
- Vía hedgehog (Hh): Esta vía de señalización, está implicada en la inducción de la señal en la organogénesis, en el patrón del tubo neural ventral y el eje anterior-posterior de las extremidades y la organización cerebral. La vía, inicia con sus tres genes principales (Sonic (Shh), Indian y Desert) que se unen al receptor de membrana Patched (Ptch), este libera la proteína Smoothened (Smo) para que migre hacia el cilio

primario, y posteriormente se desencadenan las reacciones que culminan en la activación de sus efectores Gli (Gli 1, 2 y 3)(61) (62).

En humanos, las mutaciones en los genes de la vía Hh se han asociado con un espectro de malformaciones congénitas. En particular, las mutaciones en Shh (7q36) y GLI3 (7p14.1) se asocian con holoprosencefalia tipo 3, síndromes orales, faciales y de extremidades, ano imperforado... A de más, mutaciones en los mediadores de la transducción de la Hh también suelen ser patogénicas como en KIF7 o en otros genes implicados como HOXD13 (2q31.1), la familia FOX como FOXF1 (16q24.1) ...(23,55,62,63).

Tabla 2. Genes candidatos afectados en la asociación VACTERL. Estos genes son algunos de los más comentados en la literatura de estudios genéticos (23,28,55,62).

GEN	REGIÓN CROMOSÓMICA	FENOTIPO RELACIONADO CON VACTERL	REFERENCIA
ZIC3	Xq26	Todas las anomalías VACTERL	(23,55,56,60)
SHH	7q36	Estenosis esofágica, agenesia parcial del sacro, anomalías de las vías urinarias	(23,55,62,63)
HOXD13	2q31.1	Atresia anal, malformaciones cardíacas, reflujo vesico-ureteral	(23,55,64)
HOXA13	7p15.2	Malformaciones genitourinarias, anomalías de las extremidades	(23,65,66)
FOXF1	16q24.1-q24.2	Anomalías vertebrales, cardiovasculares, renales, defectos pulmonares y atresias gastrointestinales	(55,67,68)
GLI3	7p14.1	Malformaciones ano-rectales, anomalías renales y de las extremidades	(63,69)
----	Deleción 22q11.2	Malformaciones cardíacas, anomalías renales, otras numerosas anomalías de tipo VACTERL	(23,31)

## 2.2 Next Generation Sequencing (NGS)

La revolución en Genética Humana llegó en los años 70 con la introducción de la secuenciación Sanger, marcando un hito en el comienzo de la era genómica (69). La identificación de la secuencia de nucleótidos que compone la molécula de ADN constituye un análisis detallado de su estructura y sirve como una herramienta eficaz para descubrir variantes en el material genético. No obstante, en años recientes, han surgido nuevas plataformas de secuenciación conocidas como *Next-Generation Sequencing* (NGS) o secuenciación de nueva generación de alto rendimiento, las cuales tienen la capacidad de generar de manera simultánea y masiva millones de fragmentos de ADN en un solo proceso de secuenciación (70). Esta combinación, permite una secuenciación paralela masiva de secuencias de ADN o ARN de distintas longitudes o incluso de todo el genoma (71), y por lo tanto, un aumento significativo en la eficiencia a un coste menor, ofreciendo beneficios sustanciales en comparación con los métodos convencionales.

La NGS implica varios pasos principales en la secuenciación. Por ejemplo, la NGS de ADN conlleva la fragmentación del ADN, la preparación de bibliotecas, la secuenciación paralela masiva generando millones archivos *FASTQ* con información de la secuenciación y alineamiento de las lecturas comparándolo con un genoma de referencia, el análisis bioinformático y la anotación e interpretación de variantes/mutaciones (72).

### 2.2.1 Aplicaciones y tipos:

Gracias a su gran rendimiento, este tipo de secuenciación es idóneo para una gran escala de estudios como la clonación molecular, la búsqueda de genes patógenos, los diagnósticos prenatales, los estudios comparativos y de evolución, etc (73). Por ello, la NGS se utiliza para interrogar genomas o exomas completos con el fin de descubrir mutaciones totalmente nuevas y genes causantes de enfermedades. En pediatría, podría aprovecharse para desentrañar la base genética de síndromes inexplicables de los que todavía no se sabe su etiología (71), en este caso, la VA.

Para la secuenciación, se pueden usar tres tipos de método para analizar:

- Secuenciación dirigida (paneles): La técnica más frecuentemente empleada, comprende el proceso de aislar, enriquecer y secuenciar segmentos específicos del genoma, focalizándose especialmente en las regiones codificantes de interés y las zonas intrónicas adyacentes, es decir, se centra en los genes implicados en una misma patología. Se crean paneles genéticos diseñados para respaldar el diagnóstico, permitiendo la identificación de variantes, polimorfismos, reordenamientos y variantes somáticas de baja frecuencia en numerosas muestras de manera concurrente (74). Se utilizan cuando el fenotipo del paciente es claro y su causa puede tener origen en múltiples genes. Si no se detecta ninguna alteración importante que explique el fenotipo reportado, se puede ampliar el estudio con la secuenciación del exoma o del genoma completo.
- Secuenciación del exoma: El exoma se ha definido tradicionalmente como la secuencia que abarca todos los genes codificadores de proteínas del genoma y cubre entre el 1 y el 2% del genoma, según las especies (75). La secuenciación del exoma posibilita la detección de mutaciones genéticas y factores de riesgo en familias y muestras que anteriormente se consideraban poco informativas para los estudios genéticos (75,76). Hasta el momento, la aplicación de la secuenciación del exoma ha demostrado ser exitosa en el análisis de muchas enfermedades raras, con la identificación de genes causantes de la patología cuando se desconoce la causa genética o para aquellas enfermedades con alta heterogeneidad fenotípica y genética (74).
- Secuenciación del genoma: Este método llamado metagenómica, es usado para obtener la secuenciación total del genoma, y es comúnmente usado para la identificación y clasificación taxonómica de familias y especies de bacterias, virus u otros patógenos o hasta especies ya extinguidas como el mamut (77,78). En la actualidad, la metagenómica también se aplica a investigaciones médicas o forenses, en 2009 se fundó el Proyecto del Microbioma Humano, esta iniciativa pretende cartografiar las comunidades microbianas asociadas al intestino, la boca, la piel o la vagina humanos (77,79).



### 3. Materiales y métodos

En esta sección, se plantea la transición desde la teoría descrita anteriormente hacia la aplicación práctica de las metodologías y herramientas y el motivo de este uso, en el análisis bioinformático de la AV. Se detallan las herramientas específicas utilizadas en cada paso de la construcción del pipeline bioinformático.

En primer lugar, el archivo de datos secuenciales identificado como "SRR12568924.fastq.gz" fue adquirido para este estudio mediante un proceso automatizado de descarga. Este archivo, al no estar presente en el entorno local, se descargó usando el comando "*wget*", y seguidamente, el *script* verificó la disponibilidad del recurso en el repositorio oficial del Servicio de Archivos del Centro Nacional para la Información Bioteconológica (NCBI) (80).

Los datos provienen de una paciente de seis meses, nacida de padres chinos no consanguíneos. No se registraron antecedentes de diabetes ni exposición a otros factores ambientales durante el embarazo ni hubo embarazos previos con malformaciones congénitas, y la historia familiar fue poco relevante. La paciente fue diagnosticada con la asociación VACTERL debido a la presencia de ano imperforado con fístula rectovestibular, anomalías vertebrales sacras y agenesia del cóccix, reflujo vesicoureteral y complejas anomalías cardiovasculares, como ventrículo derecho de doble salida y estenosis subaórtica asociada con defecto del tabique ventricular (81).

#### 3.1 Preprocesamiento

El preprocesamiento de datos representa la etapa inicial y crítica en el análisis de datos genómicos, enfocada en garantizar la calidad y la integridad de los datos brutos obtenidos de la secuenciación NGS. En esta fase se emplearon herramientas especializadas como *FastQC* y *Trimmomatic* para asegurar la calidad y preparar los datos para análisis posteriores.

##### 3.1.1 FASTQC

*FASTQC* (82), es una herramienta esencial que permite una evaluación exhaustiva de calidad de los datos brutos de secuenciación. Este programa, proporciona informes visuales detallados sobre parámetros fundamentales como

la distribución de calidad de las bases, la presencia de secuencias duplicadas, la presencia de adaptadores, entre otros. Así, permite un análisis que facilita la identificación de posibles problemas que podrían afectar a la precisión y confiabilidad de resultados posteriores.

El código implementado comprueba la existencia de la herramienta *FASTQC* y en este caso, se descargó la versión 0.12.1 desde la URL oficial. Posteriormente, descomprime el archivo descargado y establece los permisos necesarios para la ejecución del software. Una vez completada la instalación o verificado su previo establecimiento, se procede a ejecutar *FASTQC* en el archivo *FASTQ* designado (SRR12568924.fastq.gz) obteniendo el informe.

A continuación, se explica detalladamente todos los parámetros que genera teóricamente y se obtuvieron en el informe *FASTQC* y su aprobación o fallo de las pruebas:

### 1. Summary:

En primer lugar, se generó un informe resumen de los parámetros donde a través de los colores verde “GOOD”, amarillo “WARNING” y rojo “FAIL” se puede visualizar la calidad de los datos, tal como se ve en la *Figura 1*:























Summary	Summary
 <a href="#">Basic Statistics</a>	 <a href="#">Basic Statistics</a>
 <a href="#">Per base sequence quality</a>	 <a href="#">Per base sequence quality</a>
 <a href="#">Per tile sequence quality</a>	 <a href="#">Per tile sequence quality</a>
 <a href="#">Per sequence quality scores</a>	 <a href="#">Per sequence quality scores</a>
 <a href="#">Per base sequence content</a>	 <a href="#">Per base sequence content</a>
 <a href="#">Per sequence GC content</a>	 <a href="#">Per sequence GC content</a>
 <a href="#">Per base N content</a>	 <a href="#">Per base N content</a>
 <a href="#">Sequence Length Distribution</a>	 <a href="#">Sequence Length Distribution</a>
 <a href="#">Sequence Duplication Levels</a>	 <a href="#">Sequence Duplication Levels</a>
 <a href="#">Overrepresented sequences</a>	 <a href="#">Overrepresented sequences</a>
 <a href="#">Adapter Content</a>	 <a href="#">Adapter Content</a>

Figura 1. Informe resumen del Fastqc Report. Izquierda "good data" y derecha "bad data".

### 2. Basic Statistics:

Se formó una pequeña tabla con información clave de los datos como son el nombre del archivo, nombre total de secuencias, la longitud y el % de GC:

### ✓ Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Total Bases	10 Mbp
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Figura 2. Información básica sobre los datos.

### 3. Per base sequence quality:

Este parámetro muestra la calidad promedio de las bases en cada posición a lo largo de las secuencias. El gráfico generado reveló la distribución de la calidad de las bases a lo largo de la secuencia. Es fundamental para identificar áreas específicas de baja calidad que pueden afectar la fiabilidad de los resultados.

### ✗ Per base sequence quality

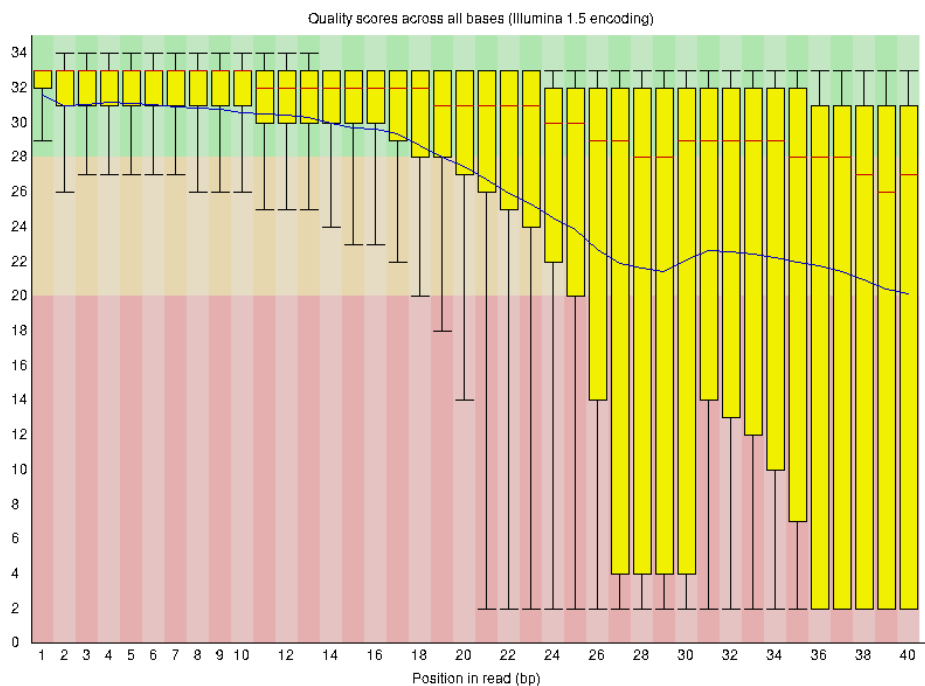


Figura 3. Representación de la calidad de los nucleótidos (eje Y) para cada posición de la muestra (eje X).

Viendo la figura de ejemplo 3, el gráfico de caja en el nucleótido 1 muestra la distribución de puntajes de calidad para el primer nucleótido de todas las lecturas de la muestra. El cuadro amarillo representa el percentil 25 y 75, con la línea roja

como la mediana. Las líneas exteriores representan el percentil 10 y 90. La línea azul representa el puntaje promedio de calidad para el nucleótido (83).

Según estas métricas, los puntajes de calidad para el primer nucleótido son bastante altos, con casi todas las lecturas teniendo puntajes por encima de 28. No obstante, se observa una tendencia a la disminución de los puntajes de calidad a medida que se avanza desde el inicio hacia el final de las lecturas.

En el contexto de lecturas generadas por secuenciación Illumina, esta reducción en la calidad no es inesperada, ya que puede estar relacionada con la intensidad de la señal fluorescente y la pureza de esta (84). La baja intensidad de la fluorescencia o la presencia de múltiples señales fluorescentes pueden impactar negativamente en la calidad asignada al nucleótido. Además, debido a la naturaleza del método de secuenciación por síntesis, es común observar algunas caídas en la calidad; sin embargo, anomalías significativas en la calidad pueden indicar posibles problemas en la instalación o procesos de secuenciación (85).

#### 4. *Per sequence quality scores:*

En este parámetro, se encuentran otros gráficos que representan el número total de lecturas versus el puntaje promedio de calidad a lo largo de toda la longitud de esa lectura. Lo que se debe observar es que la distribución del puntaje promedio de calidad debería ser bastante consistente en el rango superior del gráfico. Se aspira a que la mayoría de las lecturas tengan un puntaje promedio de calidad alto y que esta distribución se mantenga uniforme en la parte superior del gráfico.

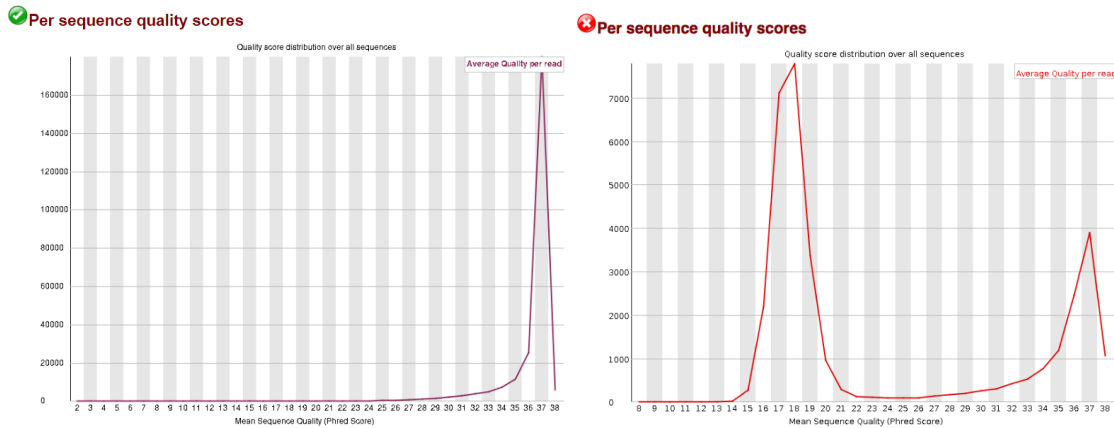


Figura 4. Representación del número total de lecturas (eje Y) versus el puntaje promedio de calidad (eje X) de unos buenos datos (izquierda) y unos malos datos (derecha).

### 5. Per base sequence content:

Este gráfico en este parámetro representa el porcentaje de bases nucleotídicas identificadas para cada uno de los cuatro nucleótidos en cada posición a lo largo de todas las lecturas en el archivo.

En la secuenciación de ADN de genoma completo, la proporción de cada uno de los cuatro nucleótidos debería mantenerse relativamente constante a lo largo de la longitud de la lectura, con  $\%A=\%T$  y  $\%G=\%C$ . En la mayoría de los protocolos de preparación de bibliotecas para RNA-Seq, se observa una distribución no uniforme de bases en los primeros 10-15 nucleótidos; este fenómeno es normal y siempre serán clasificados como “*FAIL*” por *FASTQC*, aunque la secuencia sea de calidad aceptable (86).

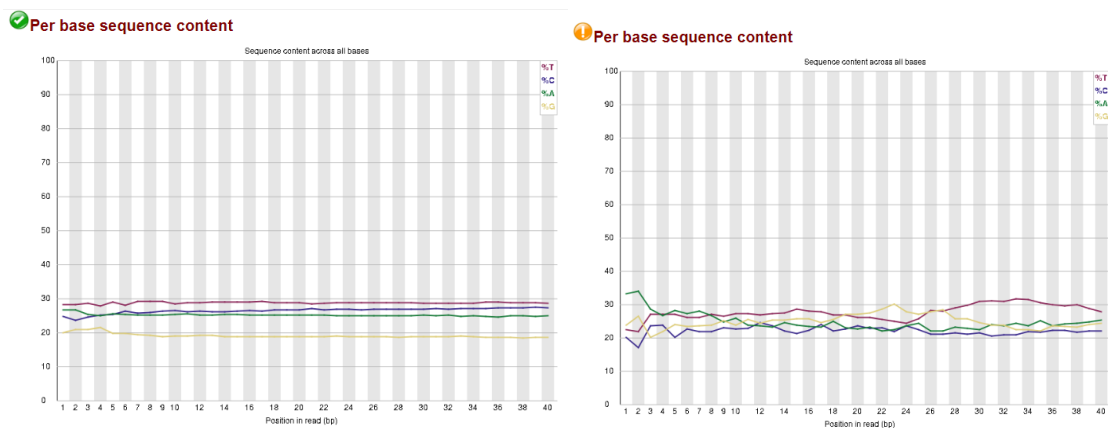


Figura 5. Porcentaje de bases nucleotídicas (eje Y) en cada posición de las lecturas (eje X) de unos buenos datos (izquierda) y unos malos datos (derecha).

## 6. Per sequence GC content:

En este apartado, el gráfico muestra el número de lecturas versus el %GC por lectura, considerando una 'Distribución Teórica' que asume un contenido uniforme de GC para todas las lecturas. En la secuenciación completa del genoma, se espera que el %GC de todas las lecturas forme una distribución normal, con el punto más alto en el contenido medio de GC para el organismo secuenciado. *FASTQC* identifica desviaciones entre la distribución observada y la teórica, marcándolas como “*FAIL*”, aunque este patrón puede variar según el tipo de secuenciación (87).

Una distribución atípica podría señalar una biblioteca contaminada o un subconjunto sesgado. La presencia de una distribución normal desplazada indica un sesgo sistemático independiente de la posición de la base. *FASTQC* advierte si la suma de las desviaciones de la distribución normal supera el 15% de las lecturas, marcando como error si supera el 30%. Las líneas roja y azul representan el %GC por lectura y la distribución teórica respectivamente. Este análisis de controles de calidad puede resultar muy inusual evidenciando variaciones significativas (88).

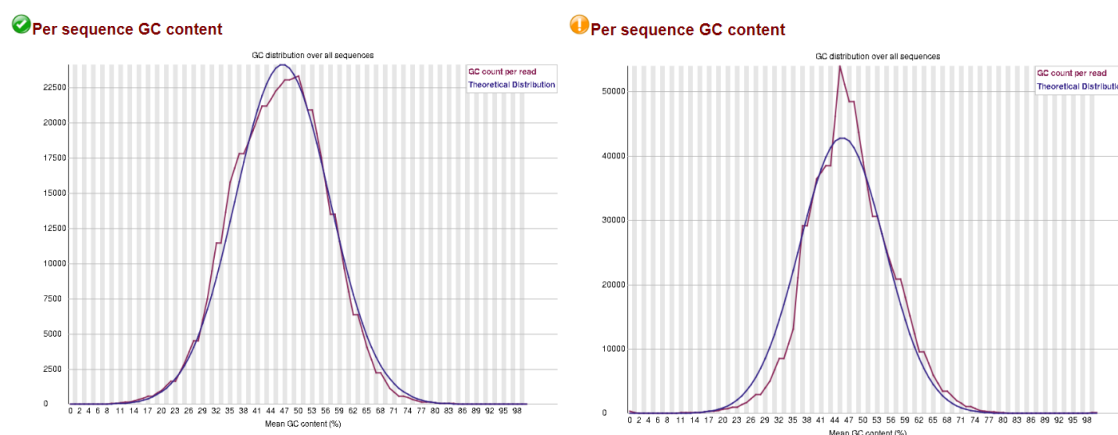


Figura 6. Nombre de lecturas (eje Y) versus el porcentaje de GC (eje X) por lectura de unos buenos datos (izquierda) y unos malos datos (derecha).

## 7. Sequence length distribution:

Muestra la longitud de las secuencias y su distribución en el conjunto de datos. Este gráfico es esencial para identificar si existen secuencias anormalmente

cortas o largas, lo que podría indicar problemas con la preparación de la biblioteca o errores de secuenciación.

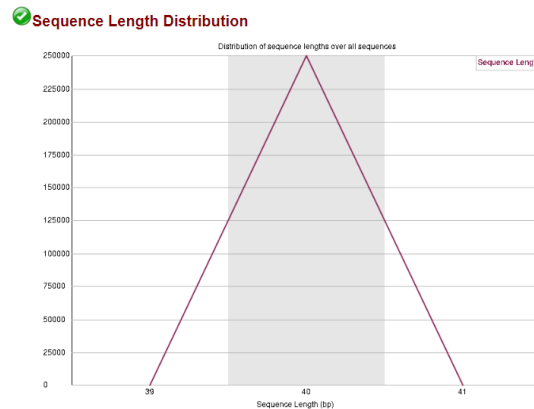


Figura 7. Distribución de la longitud de las secuencias.

### 8. Sequence Duplication Levels:

Siendo este el último parámetro relevante que se consigue, en este aparece una gráfica que muestra el porcentaje de lecturas repetidas en un archivo y señala posibles duplicados por PCR o secuencias sobrerrepresentadas. En secuenciación completa de genomas, se espera una diversidad alta con casi el 100% de lecturas únicas. En RNA-Seq, la duplicación es común en transcritos abundantes, aunque *FASTQC* a veces marca esto como error, aun siendo normal (89).

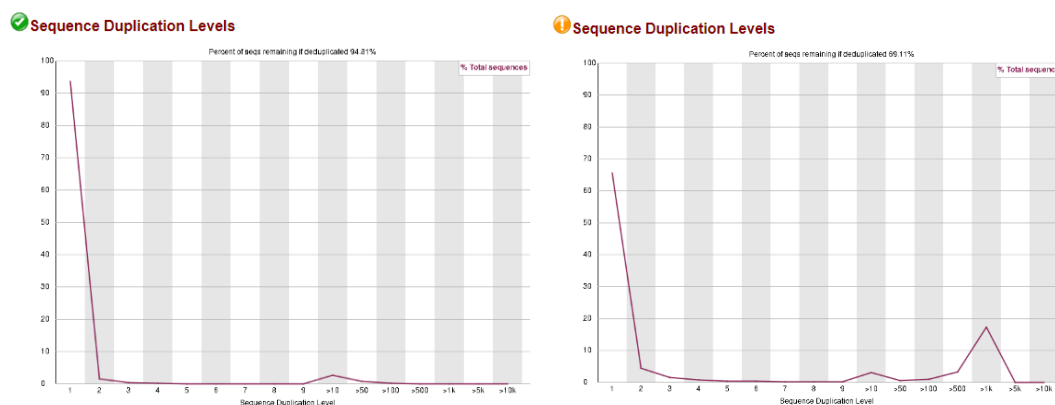


Figura 8. Porcentaje de lecturas repetidas de unos buenos datos (izquierda) y unos malos datos (derecha).

### 3.1.2 Trimmomatic

La evolución de la secuenciación genómica ha impulsado la necesidad de herramientas eficaces para el procesamiento de datos, y en este contexto,

*Trimmomatic* se destaca como una pieza fundamental en el flujo de trabajo de la bioinformática. Esta herramienta desarrollada por el equipo de Anthony Bolger de *Bjorn Usadel Lab*, “[usadellab.org](http://usadellab.org)”, aborda la crucial tarea de filtrar y manipular secuencias de ADN o ARN, mejorando sustancialmente la calidad de los datos antes de los análisis posteriores (90).

*Trimmomatic*, es esencial para garantizar la calidad de los datos, ya que, gracias a su capacidad para recortar bases no deseadas, eliminar adaptadores y filtrar las secuencias o datos de baja calidad, mejora significativamente la fiabilidad de los resultados obtenidos en estudios genómicos. Su flexibilidad es clave porque proporciona una amplia gama de parámetros ajustables, lo que permite personalizar el proceso de filtrado según las necesidades específicas de su análisis genómico (91).

En este proyecto, se empleó el comando “`java -jar Trimmomatic-0.39/trimmomatic-0.39.jar`” para invocar la versión 0.39 de *Trimmomatic* a través de Java. Se utilizó la opción “SE” para indicar que el archivo de entrada contenía *reads* de secuenciación genómica en formato *Single-End*, representando un único extremo de la secuencia.

Para definir el archivo que contenía la información de los adaptadores a recortar y sus respectivos parámetros, se usó “ILLUMINACLIP:TruSeq3-SE.fa:2:30:10” (92). Este, contiene los números que representan el umbral de coincidencia, el número de bases palindrómicas y el número de bases simples para una coincidencia exacta.

Debido a la falta de uniformidad detectada en el gráfico “*Per base sequence content*” de la *Figura 5* del proyecto, con el fin de mejorar la calidad y esta uniformidad de las secuencias obtenidas, se aplicaron dos ajustes importantes: “HEADCROP:15”, que elimina las primeras 15 bases de cada *read* para desechar posibles secuencias no deseadas o regiones de baja calidad, y “CROP:130”, que recorta todas las *reads* a una longitud estándar de 130 bases. Además, a través de “MINLEN:36”, se descartan aquellas secuencias las cuales tienen una longitud inferior a 36.



## 3.2 Alineamiento de secuencias

### 3.2.1 BWA

La evolución de las técnicas de secuenciación ha generado una demanda constante de herramientas precisas para manejar la gran diversidad de datos en genómica. En este escenario, BWA emerge como una piedra angular en el flujo de trabajo de la bioinformática. Desarrollada por Heng Li, del Centro de Genómica de Harvard y el Instituto Broad, BWA se enfoca en una tarea fundamental: mapear secuencias de lecturas cortas alineándolas con un genoma de referencia (93).

Este paquete de programas software está formado por tres algoritmos: BWA-backtrack, BWA-SW y BWA-MEM. El primero, BWA-backtrack, está diseñado para las lecturas de secuencias de hasta 100 pares de base, mientras que los otros dos algoritmos, están diseñados para secuencias más largas (94). La característica de BWA, es su capacidad para realizar alineamientos precisos de estas secuencias contra el genoma de referencia. Este proceso, no solo identifica regiones de similitud y divergencia, sino que también se adapta a distintos tipos de datos y configuraciones.

En el flujo típico de análisis genómico, BWA toma posición tras las fases iniciales de preprocesamiento de datos brutos, como la filtración por calidad y eliminación de adaptadores, como es el caso en este proyecto. El alineamiento es fundamental para lograr un mapeo preciso de las secuencias con respecto al genoma de referencia, lo que garantiza la precisión de los resultados.

En este estudio, se usó el comando BWA-MEM para iniciar BWA y realizar el mapeo de las secuencias de lecturas cortas sobre el genoma de referencia humano hg38 (95). La elección de hg38, se fundamenta en su excepcional calidad y precisión en comparación con versiones anteriores del genoma humano, como hg19 o GRCh37. Esta mejora en la precisión y corrección de variaciones genómicas en hg38 facilita alineamientos más precisos y una interpretación más fiable de variaciones genéticas, mutaciones y estructuras genómicas esenciales para la investigación genómica (96,97).

### 3.2.2 Samtools

SAMtools es una colección de programas de código abierto ampliamente utilizada en el campo de la bioinformática para manipular archivos SAM (*Sequence Alignment/Map*) y BAM (*Binary Alignment/Map*) (98). Heng Li, de *Sanger Institute*, escribió el código original de la primera versión del programa. Seguidamente, otros autores como Bob Handsaker hicieron aportaciones para mejorarlo, y así, poder ofrecer una gama diversa de herramientas para el procesamiento, visualización y análisis de datos de secuenciación genómica (99).

Una de las funciones principales de *SAMtools* es la conversión eficiente entre los formatos SAM y BAM. Mientras que SAM representa los alineamientos de las secuencias en un formato de texto legible por humanos, BAM es su contraparte binaria, más compacta y eficiente en el almacenamiento de grandes conjuntos de datos de alineamiento genómico.

En el presente proyecto, una de las funciones que permite realizar, es la capacidad de calcular estadísticas detalladas sobre el alineamiento formado. El comando “*samtools flagstat*”, proporciona información útil sobre el porcentaje de lecturas alineadas, las alineaciones múltiples, las lecturas emparejadas correctamente y otros datos estadísticos relevantes para evaluar la calidad del alineamiento. Con los resultados obtenidos de este comando, se puede observar en la *Figura 8* como no existen duplicados en la muestra. Aun así, se decidió ejecutar “*samtools rmdup*” para eliminar posibles duplicados con el objetivo de limpiar y obtener unos resultados más precisos.

## 3.3 Detección de variantes

### 3.3.1 FreeBayes

*FreeBayes* es una herramienta diseñada para la detección precisa de variantes genéticas a partir de datos de la NGS. Este software, emplea un algoritmo probabilístico basado en el modelo de Bayes para identificar variantes, incluyendo *SNPs*, *indels* (Inserciones y Deleciones) y complejas variaciones genómicas (100).

Este algoritmo, aprovecha la información de alineamiento de lecturas de secuencias de ADN a un genoma de referencia, para realizar llamadas de variantes, considerando aspectos como la calidad de la base, la cobertura y la información de mapeo de las lecturas (101). Esto permite una identificación precisa y sensible de variantes, incluso en regiones con baja cobertura o en muestras con diversidad genética.

*FreeBayes*, con su versatilidad, se adapta a múltiples contextos genómicos, abarcando desde el análisis exhaustivo de genomas completos hasta la identificación precisa de variantes en exomas y regiones específicas del genoma.

Type	What is means	Example
SNP	Single-Nucleotide Polymorphism	Reference = 'A', Sample = 'C'
Ins	Insertion	Reference = 'A', Sample = 'AGT'
Del	Deletion	Reference = 'AC', Sample = 'C'
MNP	Multiple-nucleotide polymorphism	Reference = 'ATA', Sample = 'GTC'
MIXED	Multiple-nucleotide and an InDel	Reference = 'ATA', Sample = 'GTCAGT'

Figura 9. Diferencia entre el tipo de variantes y su significado.

En este estudio, se usó *FreeBayes* como parte del análisis bioinformático para realizar el proceso de "*variant calling*" a partir de nuestros datos de secuenciación. Esto permitió identificar y caracterizar las variantes genéticas relevantes para los objetivos marcados de investigación. Dentro de *FreeBayes* se utiliza un enfoque llamado "*diploide simple*" con filtrado de calidad y cobertura (`--min-mapping-quality 30 --min-base-quality 20 --min-supporting-allele-qsum 0 -genotype-variant-threshold 0`) and `--min-coverage`) (102), para asegurar la precisión y fiabilidad de la detección de variantes genéticas.

Una vez obtenido el archivo resultante en el formato específico VCF (*Variant Call Format*), se utiliza la herramienta "*vcffilter*", para aplicar los siguientes filtros:

- **-f 'DP > 10'**: Este filtro retiene las variantes en las que la profundidad de la lectura (DP) es mayor que 10. La profundidad de lectura se refiere al número total de veces que una posición del genoma ha sido secuenciada.

- **-f 'QUAL > 30'**: Conserva las variantes con una calidad (QUAL) superior a 30. La calidad de una variante es un indicador de la confiabilidad de la llamada de la variante, siendo valores más altos indicativos de mayor confianza.
- **-f 'AF > 0.05'**: Filtra las variantes basándose en su frecuencia alélica (AF), manteniendo solo aquellas con una frecuencia alélica superior a 0.05. La frecuencia alélica representa la proporción de veces que se observa una variante en una población de muestras.

### 3.4 Anotación de variantes

#### 3.4.1 *SnpEff*

*SnpEff* es una herramienta bioinformática especializada en la predicción y anotación de los efectos de las variantes de un solo nucleótido (SNPs) y pequeñas inserciones/deleciones (INDELs) en genomas. Esta herramienta es ampliamente utilizada para interpretar las consecuencias funcionales de las variantes genéticas identificadas a partir de datos de secuenciación (103).

Este software, realiza anotaciones detalladas de las variantes identificadas, clasificándolas en distintos niveles de impacto funcional, regiones reguladoras, intrónicas, exónicas y promotoras entre otras. Esto permite una comprensión más profunda de cómo estas variantes pueden influir en la estructura de proteínas, la regulación génica o su impacto genómico (104). Una de las fortalezas de *SnpEff* es su capacidad para manejar grandes volúmenes de datos genómicos de manera eficiente, proporcionando salidas estructuradas y fáciles de interpretar que ayudan en la identificación y priorización de variantes relevantes para estudios posteriores (105).

En este estudio, *SnpEff* fue utilizado como parte integral del análisis bioinformático para interpretar las implicaciones funcionales de las variantes genéticas identificadas a partir de los datos de secuenciación, permitiendo así una comprensión más completa de su impacto en el contexto de nuestra investigación.

El archivo en formato VCF, incluye las variantes identificadas junto con anotaciones detalladas (106); permite almacenar información sobre cada

variante, como su ubicación precisa en el genoma, el tipo de variante y su posible impacto funcional en términos de cambios en proteínas, regiones reguladoras y otras áreas genómicas relevantes (107).

Además, *SnpEff* produce un archivo de resumen en formato *html* (ANNEXO 3), que condensa las características funcionales de las variantes identificadas. Este resumen detalla la naturaleza de cada variante, proporcionando una visión general de su clasificación, como variantes sinónimas, no sinónimas, inserciones, deleciones, entre otras, permitiendo así una comprensión más completa del conjunto de variantes identificadas.

### 3.5 Aplicación web

#### 3.5.1 *Streamlit*

Para poder visualizar las variantes obtenidas después de ejecutar todo el pipeline y todos los pasos, se obtiene un archivo VCF con las anotaciones. Este, es transformado a Excel para poder analizar y para que sea compatible con la herramienta de código abierto *Streamlit*. *Streamlit*, es un paquete de Python, que permite crear aplicaciones web interactivas para el análisis de datos a través de la creación de interfaces y de widgets interactivos que permiten al usuario manipular y explorar los datos de manera dinámica.

En este proyecto, permite crear dos tablas interactivas con información esencial sobre las variantes. La primera, está compuesta con información sobre el impacto funcional que han tenido las variantes respecto al gen original y la región a la cual han causado dicho impacto. Por ello, está dividida con las siguientes columnas:

- **GeneName:** El nombre del gen.
- **Genelid:** El identificador único del gen.
- **TranscriptId:** El identificador único del transcrito asociado al gen.
- **BioType:** El tipo biológico o funcional del transcrito (como "rRNA" para ácido ribonucleico ribosomal o "*protein\_coding*" para genes que codifican proteínas).

- ***HIGH, LOW, MODERATE, MODIFIER***: Categorías que indican el impacto o la gravedad de las variantes en el gen o transcrito, clasificadas en diferentes niveles de efecto, desde "*HIGH*" (alto) hasta "*MODIFIER*" (modificador).
- ***EXON, INTRON, UPSTREAM, DOWNSTREAM, UTR\_3\_PRIME, UTR\_5\_PRIME***: Estas columnas representan diferentes regiones del genoma donde se han identificado o anotado las variantes. Cada número en estas columnas indica la cantidad de variantes encontradas en una región específica del gen o transcrito, como exones, intrones, regiones aguas arriba (*UPSTREAM*) o aguas abajo (*DOWNSTREAM*) del gen, y regiones no traducidas en el ARN mensajero (*UTR\_3\_PRIME* y *UTR\_5\_PRIME*).

Esta misma, está diseñada con una serie de filtros que permiten acotar la búsqueda a los genes de interés. En primer lugar, hay un buscador que permite seleccionar por el nombre del gen, la información de dicho gen y todas sus variantes. Seguidamente, se puede filtrar por el tipo biológico o funcional del transcrito como "*protein\_coding*" si nos interesan solo los genes que codifican por proteínas. Finalmente, se puede realizar un filtrado por el grado de impacto o gravedad de las variantes, es decir, por *HIGH, LOW, MODERATE* o *MODIFIER*, y solo se mostrarán las variantes que acoten el grado escogido.

**Number of effects by impact and region:**

Choose GeneName  
e.g. A2M

Choose BioType  
e.g. rRNA

**HIGH** **LOW** **MODERATE** **MODIFIER**

#GeneName	Genelid	TranscriptId	BioType	↓ HIGH	EXON	INTRON	UPSTREAM	DOWNSTREAM	UTR_3_PRIME	UTR_5_PRIME	ENSEMBL
EPHX2	ENSG00000120915	ENST00000521400	protein_coding	263	267	2	0	0	1	0	ENSG00000120915
FDFT1	ENSG00000079459	ENST00000618539	protein_coding	89	92	9	0	0	0	0	ENSG00000079459
FDFT1	ENSG00000079459	ENST00000615631	protein_coding	89	92	9	0	0	0	0	ENSG00000079459
FDFT1	ENSG00000079459	ENST00000220584	protein_coding	89	92	9	0	0	0	0	ENSG00000079459
PNP	ENSG00000198805	ENST00000361505	protein_coding	60	63	0	0	0	0	0	ENSG00000198805
TLR8	ENSG00000101916	ENST00000218032	protein_coding	59	65	0	1	0	0	0	ENSG00000101916
CYP17A1	ENSG00000148795	ENST00000369887	protein_coding	35	38	2	0	0	0	0	ENSG00000148795
PTPRB	ENSG00000127329	ENST00000261266	protein_coding	34	44	16	1	0	0	0	ENSG00000127329
OAT	ENSG00000065154	ENST00000368845	protein_coding	33	34	3	0	0	0	0	ENSG00000065154
ALDH4A1	ENSG00000159423	ENST00000290597	protein_coding	30	35	3	0	0	0	0	ENSG00000159423

Figura 10. Visualización de los resultados en la aplicación creada con información de cada variante, su impacto y la región donde actúa.

En la segunda tabla, se presenta información esencial del archivo VCF que contiene diversas métricas para cada variante. Esta tabla incluye las siguientes columnas:

- **#CHROM:** Nombre del cromosoma (1 al 22 para autosomas, 'X' e 'Y' para cromosomas sexuales).
- **POS:** Posición exacta de la variante en el cromosoma.
- **REF:** Secuencia de referencia en esa posición.
- **ALT:** Secuencia alternativa encontrada en esa posición.
- **QUAL:** Calidad de la llamada de variante, indicando su confiabilidad.
- **DP:** Profundidad de cobertura, muestra cuántas veces se leyó esa posición.
- **AC:** Número de copias del alelo alternativo observado.
- **AF:** Frecuencia del alelo alternativo en relación con el total de copias de alelos observados.
- **AO:** Número de observaciones que contienen alelos alternativos.

- **MQM:** Calidad promedio de las bases que mapean en esa posición.
- **TYPE:** Tipo de variante, como delección ("del") o polimorfismo de nucleótido único ("snp").
- **ANN:** Información detallada sobre la anotación de la variante, incluyendo el gen afectado, tipo de transcrito, impacto de la variante (MODIFIER, LOW, etc.), posición específica en el transcrito y naturaleza de la variante.

**Variant Details (VCF file):**

Choose CHROM  
e.g. chr15

#CHROM	POS	REF	ALT	QUAL	DP	AC	AF	AO	MQM	TYPE	ANN	ENSEMBL
chr1	1,035,169	TGGGGGGGGGGGGTGGGCAGGGGTGCC	TGGGGGGGGGGGGTGGGCAGGGGTGCC	151.357	15	1,1	0,5,0,5	5,5	60,60	del,del	TGGGGGGGGGGGGTGGGCAGGGGTGCC intron_v	<a href="#">Ubicación</a>
chr1	1,035,169	TGGGGGGGGGGGGTGGGCAGGGGTGCC	TGGGGGGGGGGGGTGGGCAGGGGTGCC	151.357	15	1,1	0,5,0,5	5,5	60,60	del,del	TGGGGGGGGGGGGTGGGCAGGGGTGCC intron_v	<a href="#">Ubicación</a>
chr1	1,041,950	T	C	568.389	23	2	1	23	60	snp	C splice_region_variant&intron_variant LOW	<a href="#">Ubicación</a>
chr1	1,043,223	CTG	CG	273.086	14	2	1	11	60	del	CG upstream_gene_variant MODIFIER AGRN E	<a href="#">Ubicación</a>
chr1	1,045,707	A	G	1,456.57	57	2	1	57	60	snp	G upstream_gene_variant MODIFIER AGRN E	<a href="#">Ubicación</a>
chr1	1,046,551	A	G	1,884.43	75	2	1	71	60	snp	G upstream_gene_variant MODIFIER AGRN E	<a href="#">Ubicación</a>
chr1	1,047,561	T	C	1,277.79	49	2	1	49	60	snp	C upstream_gene_variant MODIFIER AGRN E	<a href="#">Ubicación</a>
chr1	1,047,614	T	C	1,611.68	61	2	1	61	60	snp	C upstream_gene_variant MODIFIER AGRN E	<a href="#">Ubicación</a>
chr1	1,048,922	T	C	718.526	28	2	1	28	60	snp	C upstream_gene_variant MODIFIER AGRN E	<a href="#">Ubicación</a>
chr1	1,049,886	C	T	1,289.97	49	2	1	49	60	snp	T upstream_gene_variant MODIFIER AGRN E	<a href="#">Ubicación</a>

Figura 11. Visualización del archivo VCF de las variantes finales mediante la aplicación creada.

En esta tabla, la información detallada proporciona una visión completa de las variantes, desde su ubicación en el cromosoma hasta la calidad de la llamada de la variante y su impacto potencial en los genes asociados. La incorporación de enlaces, en las dos tablas, a la base de datos ENSEMBL permite una exploración más profunda del gen y de la región, enriqueciendo así la comprensión de las variantes y su contexto genómico.

**Gene: EPHX2** ENSG00000120915

Description: epoxide hydrolase 2 [Source:HGNC Symbol;Acc:HGNC:3402]

Gene Synonyms: ABHD20, SEH

Location: Chromosome 8: 27,490,781-27,545,564 forward strand. GRCh38: CM000670.2

About this gene: This gene has 12 transcripts (splice variants), 203 orthologues, 12 paralogues and is associated with 1 phenotype.

Transcripts: [Hide transcript table](#)

Transcript ID	Name	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ENST00000521400.6	EPHX2-207	2776	555aa	Protein coding	<a href="#">CCDS6060</a>	<a href="#">P34913-1</a>	<a href="#">NM_001979.6</a>	MANE Select Ensembl Canonical GENCODE basic APPRIS P1 TSL1
ENST00000380476.7	EPHX2-201	2064	502aa	Protein coding	<a href="#">CCDS59097</a>	<a href="#">P34913-2</a>	-	GENCODE basic TSL2
ENST00000521780.5	EPHX2-209	2038	489aa	Protein coding	<a href="#">CCDS59098</a>	<a href="#">P34913-3</a>	-	GENCODE basic TSL2
ENST00000518379.5	EPHX2-204	1959	523aa	Protein coding	-	<a href="#">E5RCU2</a>	-	GENCODE basic TSL5
ENST00000517536.5	EPHX2-202	1600	372aa	Protein coding	-	<a href="#">E5REH6</a>	-	GENCODE basic TSL5
ENST00000521684.1	EPHX2-208	935	312aa	Protein coding	-	<a href="#">H0YAW7</a>	-	TSL3 CDS 5' and 3' incomplete GENCODE basic
ENST00000518328.5	EPHX2-203	574	158aa	Protein coding	-	<a href="#">E5R53</a>	-	TSL4 CDS 3' incomplete
ENST00000520666.1	EPHX2-206	582	No protein	Protein coding CDS not defined	-	-	-	TSL4
ENST00000523827.1	EPHX2-212	565	No protein	Protein coding CDS not defined	-	-	-	TSL4

Figura 12. Visualización de ESEMBL para el gen EPHX2 obtenido con el enlace de la aplicación de la tabla 1.



## Chromosome 1: 1,045,707-1,045,707



### Region in detail



Figura 13. Visualización de ESEMBL para el cromosoma 1 en la posición 1045707 obtenido con el enlace de la aplicación de la tabla 2.

La plataforma interactiva está disponible en el siguiente enlace: <https://bernatmorenotfm2024.streamlit.app/>. Para acceder al código en Python utilizado para crear esta interfaz, se puede encontrar en el repositorio GitHub: <https://github.com/Bernatmorenobatlle/bernatmorenotfm2024>. En este repositorio, se encuentran las tablas que contienen la información para la web, llamadas: "DATA\_VCF.csv" y "SNPEFF\_VCF.csv". El código de programación implementado está dentro del archivo llamado "Hello.py", y los otros son archivos propiamente autogenerados por a la herramienta *Streamlit*.

## 4. Resultados

En este apartado, se ofrece un análisis detallado de los datos genómicos obtenidos mediante las herramientas descritas en el apartado anterior y los archivos resultantes.

En primer lugar, durante la fase de evaluación inicial de calidad, se generaron los informes mediante *FASTQC* (ANNEXO 1) para analizar la integridad de los datos genómicos. Antes de la aplicación de *Trimmomatic*, se observó una distribución uniforme de la calidad de las secuencias, con prácticamente todas las lecturas presentando niveles aceptables de calidad ( $Q\text{-score} > 30$ ).

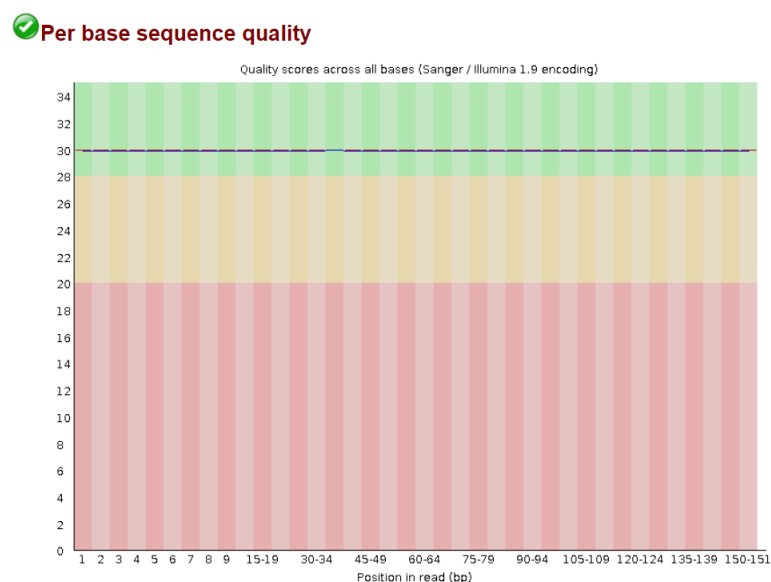


Figura 14. Distribución de la calidad de los datos del proyecto.

Sin embargo, un aspecto de relevancia residió en el parámetro de "*per base sequence content*" donde se identificó una discrepancia específica. Aproximadamente, las primeras y últimas 15 bases de las secuencias presentaron una distribución no uniforme en la composición de nucleótidos. Estos hallazgos sugieren la existencia de variaciones en la secuencia, lo que posiblemente indique peculiaridades biológicas o la presencia de elementos no habituales en estas áreas del genoma analizado.

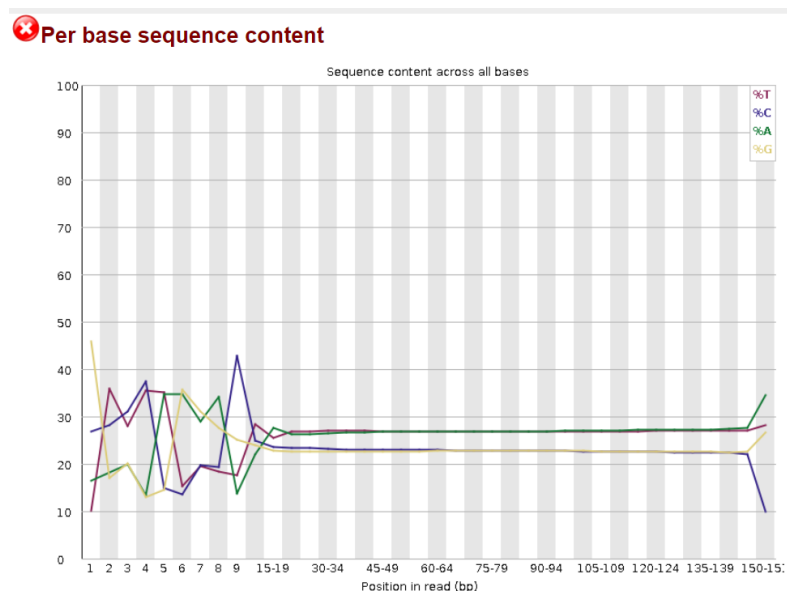


Figura 15. Porcentaje de la composición de nucleótidos en los datos del proyecto.

Debido a este resultado, el proceso de filtrado con *Trimmomatic* se enfoca en recortar tanto las primeras como las últimas 15 bases de las secuencias. Esta estrategia está diseñada para abordar específicamente este problema, buscando mejorar la fiabilidad de los datos resultantes, obteniendo así el gráfico siguiente donde se puede observar esta uniformidad de las bases que se buscaba:

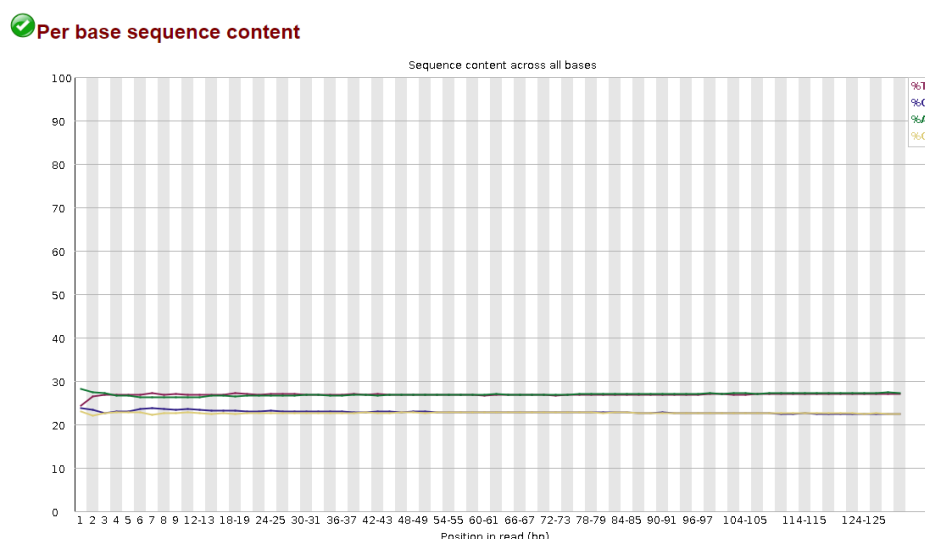


Figura 16. Porcentaje de la composición de nucleótidos en los datos del proyecto después de *Trimmomatic*.

Finalmente, en el informe de calidad (*FASTQC*) (ANNEXO 2), se han identificado algunos avisos (!) en parámetros específicos como el contenido de GC por secuencia, la distribución de la longitud y los niveles de duplicación. Es

importante destacar que estos hallazgos no representan una alarma significativa en la integridad general de los datos genómicos.

En el caso de la longitud de las secuencias, se observa un recorte tras la aplicación de *Trimmomatic*, como medida correctiva para abordar la problemática previamente identificada. Esta acción ha ayudado a mitigar los desafíos encontrados en las etapas iniciales del análisis, permitiendo una mejora gradual en la uniformidad de la longitud.

Por otro lado, los niveles de duplicación registrados no muestran una relevancia sustancial, ya que se mantienen en rangos muy bajos. Este fenómeno no afecta significativamente la calidad general de los datos, lo que sugiere que estos niveles de duplicación no representan una preocupación sustancial para el análisis posterior.

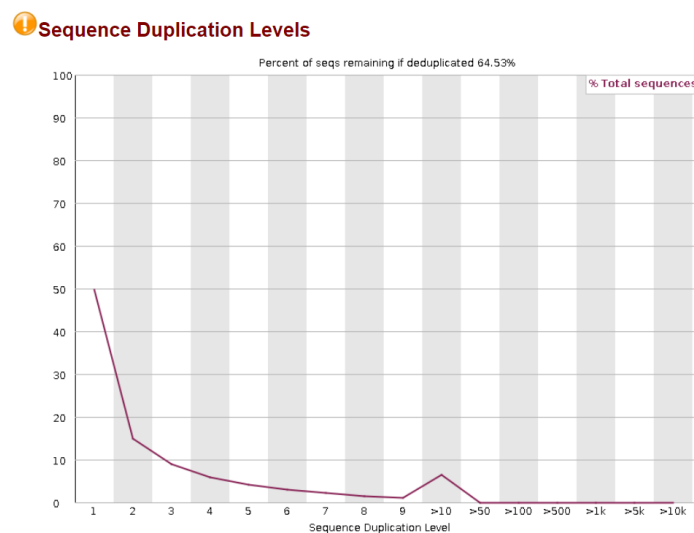


Figura 17. Niveles de duplicación de los datos del proyecto.

Los avisos señalados en el informe *FASTQC* se consideran indicativos, más que críticas anomalías en la calidad de los datos genómicos. Tras la aplicación del filtrado con *Trimmomatic*, se generó un nuevo archivo *FASTQ*, manteniendo un total de 17,198,228 secuencias de las 17,367,368 del conjunto original. En otras palabras, tras el filtrado, el 99% de las secuencias superaron el proceso, eliminándose únicamente el 1% restante.

Basic Statistics		Basic Statistics	
Measure	Value	Measure	Value
Filename	SRR12568924.fastq.gz	Filename	SRR12568924_trimmed.fq.gz
File type	Conventional base calls	File type	Conventional base calls
Encoding	Sanger / Illumina 1.9	Encoding	Sanger / Illumina 1.9
Total Sequences	17367368	Total Sequences	17198228
Total Bases	2.4 Gbp	Total Bases	2 Gbp
Sequences flagged as poor quality	0	Sequences flagged as poor quality	0
Sequence length	32-151	Sequence length	36-130
%GC	46	%GC	45

Figura 18. Información básica de los datos originales (izquierda) y los datos filtrados (derecha).

En segundo lugar, una vez obtenido el archivo con los datos filtrados, se procedió a realizar el alineamiento con BWA-MEM, obteniendo así un archivo BAM.

Utilizando la herramienta “*Samtools flagstat*” (108), se llevó a cabo un análisis de las métricas del proceso de alineamiento al genoma de referencia. Estas métricas contienen valiosa información que ilustra la eficacia y la precisión del alineamiento, proporcionando datos detallados sobre el proceso.

Métricas del Archivo BAM:

- **Lecturas Totales y Calidad:** Se han identificado un total de 17,214,435 lecturas en el archivo BAM. De estas, 17,198,228 se clasifican como lecturas primarias, indicando que estas lecturas cumplen con altos estándares de calidad y son adecuadas para análisis detallados.
- **Tipo de Lecturas y Alineamientos:** Se observa la presencia de 16,207 lecturas suplementarias en el archivo BAM, las cuales se alinean a múltiples regiones del genoma. Destacando que el 99.89% de las lecturas totales (17,194,900) se han alineado correctamente, lo que resalta la eficiencia y precisión del proceso de alineación.
- **Alineamientos Primarios y Duplicados:** Entre las lecturas alineadas correctamente, se identifican 17,178,693 alineamientos primarios, sin embargo, no se han detectado duplicados en este conjunto de datos, resaltando la singularidad de las lecturas mapeadas.

Estos resultados, ofrecen una visión detallada del proceso de alineación y la calidad de las lecturas en el archivo BAM generado. La alta proporción de lecturas primarias y su alineación efectiva refuerzan la confiabilidad de los

análisis genómicos posteriores y la integridad de las lecturas con respecto al genoma de referencia.

Una vez se ha comprobado la eficacia del proceso de alineamiento se procede con el llamado de variantes con la herramienta *FreeBayes*. Para poder analizar la calidad del llamado de variantes, se utiliza la herramienta “*bcftools stats*” (109) para ver la calidad del archivo producido por *FreeBayes*. Esta, proporciona un análisis exhaustivo de las diferentes categorías de variantes presentes en el archivo VCF.

- Resumen de Variantes:
  - Se registraron 224,671 registros de variantes en el archivo VCF analizado
  - La mayoría de las variantes corresponden a *SNPs* (190,917), seguidas por *MNPs* (3,385) e *indels* (29,099).
  - Además, se identificaron 2,197 eventos con otros tipos de variantes, como alelos simbólicos o sustituciones complejas.
  - Se detectaron 2,366 sitios multialélicos, de los cuales 354 presentan múltiples alelos alternativos, todos catalogados como *SNPs*.
  - Se identificaron 43,322 variantes genéticas únicas (*singletons*), observadas una sola vez en el conjunto de datos. Estas abarcan principalmente transiciones (23,989) y transversiones (19,333), con un total de 8,640 *indels* en este conjunto único.
- Transiciones/Transversiones (TSTV):
  - La relación entre transiciones y transversiones se estima en 1.26, sugiriendo una ligera predominancia de transiciones sobre transversiones en las variantes *SNPs* identificadas.

El proceso de anotación de variantes culmina en la generación de un archivo que alberga diversas métricas y gráficos explicativos que profundizan en la naturaleza y relaciones entre las variantes identificadas. Este archivo ha sido sometido a un filtrado, reduciendo el conjunto de variantes para identificar

aquellas con mayor probabilidad de ser genuinas, destacando por su alta calidad de anotación. Los criterios aplicados en este proceso de filtrado son:

- Calidad de Variante (*QUAL*): Se ha establecido un umbral mínimo de calidad para las variantes, seleccionando únicamente aquellas con *QUAL* > 30.
- Profundidad de Cobertura (*DP*): Esta métrica indica la frecuencia con la que una posición específica ha sido secuenciada. Una alta cobertura incrementa la confiabilidad de las variantes, reduciendo la probabilidad de errores de secuenciación o llamadas erróneas. Se han conservado las variantes con *DP* > 10.
- Frecuencia de Alelo (*AF*): Se han mantenido las variantes con una frecuencia de alelo superior al 5%, con un umbral de *AF* > 0.05.

Los resultados del análisis de variantes muestran una diversidad significativa en cuanto a la naturaleza y el impacto de las alteraciones genómicas. Las variantes se distribuyen en diversas categorías, proporcionando una versión detallada de su composición y efecto (ANNEXO 3).

En total, se registraron 28.340 variantes, de estas, se identificaron como *SNPs*, 23,938 variantes que implican cambios en un solo nucleótido en comparación con la secuencia de referencia. Los *MNPs*, con 494 casos, involucran cambios en múltiples nucleótidos de manera simultánea. Las inserciones (*INS* - 1,570) y deleciones (*DEL* - 2,105) señalan adiciones o eliminaciones de fragmentos de ADN, respectivamente, y las variantes mixtas (*MIXED* - 233) implican combinaciones de diversos tipos de variantes.

Tabla 3. Tipos de variantes y su presencia en los datos.

<b><i>TYPE</i></b>	<b><i>TOTAL</i></b>
<i>SNP</i>	23,938
<i>MNP</i>	494
<i>INS</i>	1.570
<i>DEL</i>	2.105
<i>MIXED</i>	233

<b><i>TOTAL</i></b>	28,340
---------------------	--------

Al considerar el impacto de estas variantes, las clasificaciones en categorías como alto (*HIGH*), bajo (*LOW*), moderado (*MODERATE*) y modificadores (*MODIFIER*) indican la gama de efectos potenciales sobre las proteínas y su función biológica:

- ***HIGH***: Estas variantes suelen tener un impacto directo en la función de la proteína codificada por el gen. Pueden causar cambios drásticos o interrumpir la función normal de la proteína. En los resultados, las variantes con este tipo de impacto representan el 1.18%, 2,134 del total, lo que sugiere que hay un número limitado de variantes con efectos altamente significativos en la función de las proteínas.
- ***LOW***: Las variantes clasificadas como de impacto bajo pueden tener ciertos efectos en la función de la proteína, pero generalmente son menos disruptivas o tienen efectos más sutiles. Estas pueden incluir cambios en regiones no críticas o cambios que no alteran significativamente la estructura o función de la proteína. Constituyen el 8.528% de las variantes identificadas, 15,416 variantes.
- ***MODERATE***: Las variantes con impacto moderado podrían afectar la función de la proteína, pero sin provocar cambios tan significativos como las variantes de alto impacto. Estos cambios pueden influir en la estructura o la actividad de la proteína de manera moderada. En los resultados, hay 8,671, que representa un 4.797% del total.
- ***MODIFIER***: Estas variantes tienen un impacto mínimo o nulo en la función de la proteína. Suelen encontrarse en regiones no codificantes del genoma o en partes de los genes que no afectan directamente la síntesis de proteínas. Constituyen una gran mayoría, representando el 85.495%. A menudo, las variantes en esta categoría se consideran neutrales o no afectan significativamente la función biológica.

Estos resultados, sugieren una proporción considerablemente alta de variantes clasificadas como modificadores, lo que indica que la mayoría de las variantes



identificadas en este conjunto de datos podrían tener un impacto mínimo en la función de las proteínas o en el fenotipo.

Tabla 4. Cantidad de los diferentes tipos de impacto de las variantes.

<b>TYPE</b>	<b>COUNT</b>	<b>PERCENT</b>
<i>HIGH</i>	2,134	1.18%
<i>LOW</i>	15,416	8.528%
<i>MODERETE</i>	8,671	4.797%
<i>MODIFIER</i>	154,553	85.495%

Si nos centramos en la clasificación de las variantes por su clase funcional, se puede observar cómo existen 3 tipos diferentes de mutaciones en el genoma:

- **MISSENSE:** Representa la mayoría de las variantes identificadas, con un total de 8,016. Estas mutaciones son cambios en un solo par de bases que resultan en un aminoácido diferente en la proteína codificada, lo que podría afectar su función.
- **NONSENSE:** Se refiere a un tipo de mutación que lleva a la formación de un codón de parada prematuro, lo que resulta en una proteína truncada y no funcional. Son menos frecuentes, con un total de 51.
- **SILENT:** Estas variantes no causan cambios en los aminoácidos codificados por los genes y, por lo tanto, no alteran la función de la proteína. Son las más comunes, con un total de 10,887.

Tabla 5. Cantidad de los diferentes tipos de clase funcional de variantes.

<b>TYPE</b>	<b>COUNT</b>	<b>PERCENT</b>
<i>MISSENSE</i>	8,016	42.292%
<i>NONSENSE</i>	51	0.269%
<i>SILENT</i>	10,887	57.439%

La ratio entre las variantes MISSENSE y SILENT, calculado en 0.7363, proporciona una percepción sobre la proporción relativa entre dos tipos de mutaciones en el genoma.

Este valor sugiere que hay una mayor presencia de cambios en la secuencia de ADN que pueden alterar la función de las proteínas (variantes MISSENSE) en comparación con aquellas que no producen cambio alguno en los aminoácidos codificados (variantes SILENT). Una ratio menor a 1, como en este caso, indica que hay más variantes que pueden tener un efecto sobre la función proteica en relación con aquellas que no lo hacen.

Finalmente, las variantes pueden ser identificadas dependiendo de la región y del tipo. Estos distintos tipos de variantes podrían tener diversas implicaciones funcionales en la expresión génica y la función de las proteínas. En nuestro proyecto, las principales regiones donde se han detectado variantes genéticas son las siguientes:

- **DOWNSTREAM:** Estas variantes, representando el 15.379%, se encuentran justo después de los genes y podrían tener implicaciones en la regulación génica posterior a la transcripción.
- **EXON:** Con el 16.737%, estas variantes están en las regiones codificantes de los genes, afectando directamente la estructura y función de las proteínas. Dentro de estas, hay principalmente 3 tipos de variantes encontradas; las *missense\_variant*, *non\_coding\_transcript\_exon\_variant* i las *synonymous\_variant* que, a pesar de no cambiar el aminoácido codificado, estas ocurren dentro los exones.
- **INTERGENIC:** Estas variantes, representando el 0.868%, se encuentran entre genes, y su presencia puede influir en la organización o regulación global del genoma.
- **INTRON:** Representando el 49.884%, estas variantes residen en las regiones no codificantes de los genes, desempeñando roles regulatorios, estructurales o de procesamiento de ARN.
- **MOTIF:** Estas variantes (0.097%) podrían afectar secuencias de unión a proteínas reguladoras o factores de transcripción, lo que influye en la expresión génica.
- **SPLICE\_SITE\_ACCEPTOR y SPLICE\_SITE\_DONOR:** Representan regiones esenciales para el proceso de empalme del ARN. Variantes aquí

(0.041% y 0.032%, respectivamente) podrían afectar la precisión del empalme.

- **SPLICE\_SITE\_REGION:** Estas variantes (1.953%) se encuentran en regiones más amplias que incluyen los sitios de empalme, potencialmente influenciando el procesamiento del ARN.
- **TRANSCRIPT:** Variantes en esta región (0.656%) podrían influir en la estructura o función del ARN mensajero.
- **UPSTREAM:** Estas variantes (11.195%) están localizadas arriba de los genes y podrían afectar la regulación de la transcripción y la expresión génica.
- **UTR\_3\_PRIME y UTR\_5\_PRIME:** Representan las regiones no traducidas del ARN mensajero. Variantes aquí (2.144% y 1.014%, respectivamente) pueden influir en la estabilidad del ARN o la regulación de la traducción.

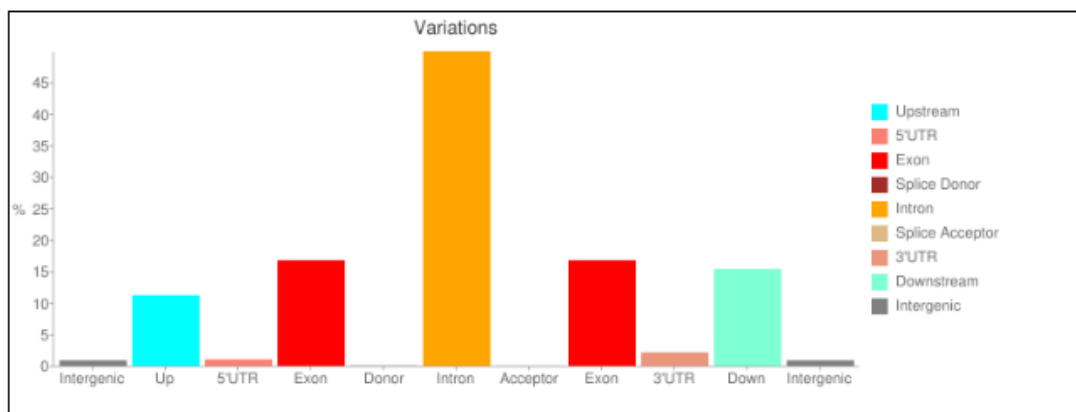


Figura 19. representación del porcentaje de variantes en las regiones del genoma.

Para poder ver los resultados comentados anteriormente e interactuar con ellos, la aplicación realizada en *Streamlit*, permite filtrar por el impacto o la gravedad de las variantes en el gen y poder observar en que región del genoma se han identificado o anotada dicha variante. Cada número de las columnas indica la cantidad de variantes encontradas en la región específica del gen. Por ejemplo, en la primera fila, el gen "5S\_rRNA" tiene dos transcritos diferentes ("ENST00000618635" y "ENST00000614916"). Ambos están asociados con el

tipo biológico "rRNA" y tienen variantes anotadas en la región "MODIFIER" (4 y 2 variantes respectivamente). Además, se observan variantes en regiones específicas del gen para cada transcrito, como intrones, UPSTREAM y DOWNSTREAM, pero no en exones, *Figura 20*.

Number of effects by impact and region:

Choose GeneName

5S\_rRNA

Choose BioType

e.g. rRNA

HIGH LOW MODERATE MODIFIER

#GeneName	GeneId	TranscriptId	BioType	↓ HIGH	LOW	MODERATE	MODIFIER	EXON	INTRON	UPSTREAM	DOWNSTREAM	UTR_3_PRIME	UTR_5_PRIME	ENSEMBL
5S_rRNA	ENSG00000277411	ENST00000614916	rRNA	0	0	0	2	0	0	2	0	0	0	<a href="#">ENSG000002774</a>
5S_rRNA	ENSG00000274059	ENST00000618635	rRNA	0	0	0	4	0	0	2	2	0	0	<a href="#">ENSG000002740</a>

Figura 20. Resultados para el gen 5S\_rRNA para su impacto funcional y region.

Además, en la otra tabla presente en la aplicación interactiva, proporciona información adicional sobre las variantes registradas donde cada fila representa una variante genética específica con sus detalles asociados como su ubicación precisa en el cromosoma, las secuencias de referencia y alternativas, la calidad de la identificación de la variante y su profundidad de cobertura y finalmente una última columna con la información descrita en la otra tabla de la aplicación, *Figura 11*.

## 5. Discusión

Uno de los objetivos claves de este proyecto era poder relacionar las variantes relevantes obtenidas de los datos con la asociación VACTERL. Para ello, se decide escoger las variantes las cuales pueden tener un impacto más grande dentro del genoma, es decir, aquellas cuyo impacto funcional sea muy grande (HIGH) y la región de afectación sea el exón, la parte codificadora del genoma.

Como se puede observar (*Figura 10*), hay un gen el cual se han encontrado un gran número de variantes cuyas las cuales tienen un gran impacto. Este gen es *EPHX2*, un gen que codifica por una enzima llamada epóxido hidrolasa 2 que participa en procesos metabólicos que implican la metabolización de sustancias extrañas para convertirlas en formas más solubles i fáciles de eliminar (110). Este gen, participa en la regulación de funciones cardiovasculares a través de la actividad del epóxido hidrolasa soluble con actividad de fosfatasa de lípidos-fosfato (81). Si bien no existe una conexión directa conocida con la AV, su función en la regulación cardiovascular podría ser relevante. Dado que *EPHX2* está relacionado con este tipo de funciones, es posible que las vías en las que participa el gen tengan algún grado de influencia en aspectos específicos de la patología.

El gen *FDFT1* codifica una enzima llamada sintasa de escualeno, que desempeña un papel crucial en la biosíntesis del colesterol y los isoprenoides, componentes esenciales para la formación de membranas celulares y la regulación de procesos biológicos (111). Otros estudios, han destacado la relación entre la expresión de *FDFT1* y varios tipos de cáncer, como el colorrectal, de próstata y otros, sugiriendo su participación en la proliferación celular y en fenotipos más agresivos (112,113).

A pesar de estas asociaciones con cánceres, no se ha establecido un vínculo directo entre *FDFT1* y la AV. Sin embargo, dado el papel crucial de *FDFT1* en la regulación de vías metabólicas y su influencia en la proliferación celular, podría plantearse la hipótesis de que mutaciones o variantes en este gen podrían tener un impacto en procesos embrionarios y de desarrollo, lo que potencialmente podría contribuir a anomalías congénitas como las asociadas la AV.

Otro gen que se puede observar que sus variantes tienen un impacto funcional muy elevado es el gen TLR8. *Toll -Like Receptor 8* es una proteína que forma parte de la familia de receptores de reconocimiento de patrones y que desempeñan un papel crucial en la respuesta inmunitaria (114). Aun así, tampoco hay evidencia directa que vincule variantes de este gen con la AV, se podría dar la posibilidad que una modificación significativa de su función y alteración interfiere con los procesos cruciales para el desarrollo embrionario y su posterior aparición de anomalías características del AV (115).

El gen ILK (*Integrin-Linked Kinase*) codifica una proteína que desempeña un papel fundamental en la señalización celular y la interacción entre las células y la matriz extracelular. Esta proteína está involucrada en múltiples procesos biológicos, como la adhesión celular, la migración, la supervivencia celular y la regulación del ciclo celular (116). Además, ILK es una proteína que interviene en vías de señalización cruciales durante el desarrollo embrionario y en la formación de vasos (117). Las alteraciones en estas vías de señalización podrían teóricamente influir en procesos relevantes para el desarrollo normal, como la morfogénesis, la diferenciación celular o la formación de estructuras corporales, que podrían estar relacionadas con la AV.

Si tratamos de confirmar la presencia de los genes mencionados en la *Tabla 2*, los cuales, según la literatura bibliográfica, se cree que pueden estar relacionados con AV, como *ZIC3*, *SHH*, *HOXD13*, *HOXA13*, *FOXF1* y *GLI3*, observamos algunas discrepancias. Algunos de estos genes no aparecen en el archivo final de variantes, la cual cosa no implica necesariamente que no estén presentes en los datos iniciales, pero es posible que no hayan superado los filtros específicos implementados en este trabajo o no estén presentes en las muestras del paciente. Un ejemplo de esto es la ausencia de los genes *HOXD13*, *HOXA13* y *FOXF1* en el archivo final de variantes. En cambio, se hallaron variantes con implicaciones modificadoras en exones e intrones de *ZIC3*, *SHH* y *GLI3*. *GLI3* mostró variantes *MODERATE* en diferentes transcritos, mientras que *SHH* presentó variantes *MODIFIER* en intrones, incluso en transcritos asociados con degradación mediada por *nonsense*. En *ZIC3* se encontraron variantes *MODIFIER* en intrones y una variante *UPSTREAM* en un transcrito procesado.

Finalmente, cabe mencionar que, en el trabajo original, se identificaron diversas variantes genéticas asociadas con malformaciones complejas en diferentes órganos. Seis de estas variantes están vinculadas a disfunciones cardíacas específicas como son la variante del gen *OLR1* y la variante de *PSMA6* que se relacionan con infarto de miocardio, la variante de *AKAP10* está asociada con defectos en la conducción cardíaca, la variante del gen *PON1* con enfermedad arterial, y las variantes del gen *EPHX2*, descrito anteriormente en los resultados de este trabajo, y *GHRL* con síndrome metabólico. Todas esas variantes están presentes en el archivo final de variantes de este trabajo.

Además, se encontraron variantes en protooncogenes y supresores tumorales que no habían sido previamente asociados con VACTERL. Estos incluyen variantes en los protooncogenes *CCND1*, *AURKA*, *MERTK*, *CSF1R*, *MYB*, *ROS1*, *PCM1*, *FGFR2*, *MYH11*, *BRCC3*, y variantes en los supresores tumorales *SDHA*, *RB1CC1*, *PTCH1*, *DMBT1* y *BCR*. Estas asociaciones abren nuevas perspectivas en la comprensión genética de la enfermedad VACTERL.

En este contexto, es importante destacar que, aunque se han identificado variantes genéticas en varios genes, que podrían ser asociados con VACTERL, estas deben considerarse como hipótesis preliminares y no como conclusiones definitivas dado que la presencia de variantes en estos genes no establece una relación causal directa con el desarrollo de la enfermedad VACTERL. La complejidad de la AV y su etiología aún no completamente comprendida, así como la falta de estudios funcionales específicos para estas variantes, plantean desafíos significativos en la interpretación de las variantes genéticas identificadas.

## 6. Conclusiones

Las conclusiones de este estudio reflejan el cumplimiento satisfactorio de los objetivos planteados en el contexto de la asociación VACTERL, una condición que, hasta la fecha, carece de una comprensión clara de sus causas genéticas subyacentes. Se ha conseguido desarrollar un pipeline bioinformático integral funcional con el objetivo de analizar datos genómicos en una paciente afectada por la AV, obteniendo así un conjunto de variantes de genes que podrían estar relacionados con dicha asociación.

A pesar de los logros obtenidos en este estudio, es crucial destacar que la patología estudiada, presenta un desafío considerable en cuanto a la identificación de sus bases genéticas dada la naturaleza heterogénea y multifactorial. Los resultados obtenidos a partir de datos de una sola paciente, aunque valiosos, no pueden afirmar ni refutar de manera concluyente su validez en el contexto de causalidad y determinación de la enfermedad. Por ello, debido a la falta de replicación en un conjunto más amplio de pacientes, se limita la generalización de los resultados obtenidos.

La necesidad de realizar pseudorréplicas o replicaciones adicionales se plantea como un paso fundamental para validar y contextualizar de manera más precisa la información genómica adquirida en este estudio inicial.

En resumen, a pesar de los avances significativos logrados con el pipeline desarrollado, los resultados no conllevan una conclusión determinante en cuanto a la causalidad y base genética de la AV. El camino hacia la comprensión completa de las bases genéticas de este conjunto de malformaciones implica la realización de estudios adicionales, con un enfoque en la replicación y la ampliación de la muestra, para obtener conclusiones más robustas y representativas de la diversidad genética de esta compleja entidad clínica.

Referente al apartado de ética y sostenibilidad, los impactos previstos como el buen uso de la bibliografía y de la energía sostenible y la responsabilidad del manejo de datos de una paciente, en todo momento se han respetado, por lo que se han mitigado y no han aparecido de nuevos en su desarrollo. Y en relación al impacto positivo que se esperaba obtener como aporte científico, en este



trabajo se ha conseguido cumplir este hito, ya que aún que pueda ser un aporte ínfimo de información, todo nuevo dato sobre esta patología poco conocida será un avance cada vez mayor.

## 7. Glosario

- **AV:** Asociación VACTERL.
- **VACTERL:** V: anomalías vertebrales, A: atresia anal, C: anomalías cardíacas, TE: fístula traqueoesofágica, R: anomalías renales, L: defectos en las extremidades.
- **NGS:** *Next-Generation Sequencing*.
- **BWA:** Alineamiento de *Burrows-Wheeler*.
- **BWT:** Inversa transformada de *Burrows-Wheeler*.
- **SNPs:** Polimorfismos de nucleótido único.
- **Indels:** Inserciones y deleciones.
- **NCBI:** Centro Nacional para la Información Biotecnológica.
- **SAM:** *Sequence Alignment/Map*.
- **BAM:** *Binary Alignment/Map*.
- **VCF:** *Variant Call Format*.

## 8. Bibliografía

1. Botto LD, Khoury MJ, Mastroiacovo P, Castilla EE, Moore CA, Skjaerven R, et al. The spectrum of congenital anomalies of the VATER association: An international study. *Am J Med Genet*. 1997 Jul 11;71(1):8–15.
2. Solomon BD, Pineda-Alvarez DE, Raam MS, Bous SM, Keaton AA, Vélez JI, et al. Analysis of Component Findings in 79 Patients Diagnosed with VACTERL Association. 2010;
3. Pelizzo G, Chiricosta L, Mazzon E, Zuccotti GV, Avanzini MA, Croce S, et al. Discovering Genotype Variants in an Infant with VACTERL through Clinical Exome Sequencing: A Support for Personalized Risk Assessment and Disease Prevention. *Pediatric Reports* 2021, Vol 13, Pages 45-56 [Internet]. 2021 Jan 5 [cited 2023 Oct 13];13(1):45–56. Available from: <https://www.mdpi.com/2036-7503/13/1/6/htm>
4. Magi A, Benelli M, Gozzini A, Girolami F, Torricelli F, Brandi ML. Bioinformatics for Next Generation Sequencing Data. *Genes (Basel)* [Internet]. 2010 [cited 2023 Oct 13];1:294–307. Available from: [www.mdpi.com/journal/genes](http://www.mdpi.com/journal/genes)
5. BOE-A-2022-5809 Ley 7/2022, de 8 de abril, de residuos y suelos contaminados para una economía circular. [Internet]. [cited 2024 Jan 16]. Available from: <https://www.boe.es/buscar/act.php?id=BOE-A-2022-5809>
6. Energy - United Nations Sustainable Development [Internet]. [cited 2024 Jan 16]. Available from: <https://www.un.org/sustainabledevelopment/energy/>
7. BOE-A-1996-8930 Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia. [Internet]. [cited 2024 Jan 16]. Available from: <https://www.boe.es/buscar/act.php?id=BOE-A-1996-8930>
8. Bioinformatics Pipeline - MATLAB & Simulink - MathWorks España [Internet]. [cited 2023 Oct 15]. Available from: <https://es.mathworks.com/help/bioinfo/bioinformatics-pipeline.html>
9. Frías López C. Desarrollo de técnicas bioinformáticas para el análisis de datos de secuenciación masiva en sistemática y genómica evolutiva: Aplicación en el análisis del sistema quimiosensorial en artrópodos. TDX (Tesis Doctorals en Xarxa) [Internet]. 2019 Nov 29 [cited 2023 Oct 16]; Available from: <https://www.tdx.cat/handle/10803/668283>
10. Smith AD, de Sena Brandine G. Falco: high-speed FastQC emulation for quality control of sequencing data. *F1000Res* [Internet]. 2019 [cited 2023 Oct 16];8. Available from: [/pmc/articles/PMC7845152/](https://pmc/articles/PMC7845152/)
11. Bolger AM, Lohse M, Usadel B. Genome analysis Trimmomatic: a flexible trimmer for Illumina sequence data. 2014 [cited 2023 Oct 16];30(15):2114–20. Available from: <http://www.usadellab.org/cms/index>
12. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* [Internet]. 2018 Sep 9 [cited 2023 Oct 16];34(17):i884. Available from: [/pmc/articles/PMC6129281/](https://pmc/articles/PMC6129281/)

13. Next Generation Sequencing (NGS)/Alignment - Wikibooks, open books for an open world [Internet]. [cited 2023 Oct 16]. Available from: [https://en.wikibooks.org/wiki/Next\\_Generation\\_Sequencing\\_\(NGS\)/Alignment](https://en.wikibooks.org/wiki/Next_Generation_Sequencing_(NGS)/Alignment)
14. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* [Internet]. 2009 Jul 7 [cited 2023 Oct 16];25(14):1754. Available from: [/pmc/articles/PMC2705234/](https://pmc/articles/PMC2705234/)
15. Langmead B. Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics / editorial board, Andreas D Baxevanis . [et al]* [Internet]. 2010 [cited 2023 Oct 16];CHAPTER(SUPP.32):Unit. Available from: [/pmc/articles/PMC3010897/](https://pmc/articles/PMC3010897/)
16. Variant Calling – NGS Analysis [Internet]. [cited 2023 Oct 16]. Available from: <https://learn.gencore.bio.nyu.edu/variant-calling/>
17. Variant Calling using Freebayes [Internet]. [cited 2024 Jan 9]. Available from: <https://manual.omicsbox.biobam.com/user-manual/omicsbox-modules/module-genetic-variation/variant-calling/variant-calling-using-freebayes/>
18. Variant calling with Freebayes | In-depth-NGS-Data-Analysis-Course [Internet]. [cited 2024 Jan 9]. Available from: [https://hbctraining.github.io/In-depth-NGS-Data-Analysis-Course/sessionVI/lessons/02\\_variant-calling.html](https://hbctraining.github.io/In-depth-NGS-Data-Analysis-Course/sessionVI/lessons/02_variant-calling.html)
19. Tuteja S, Kadri S, Yap KL. A performance evaluation study: Variant annotation tools - the enigma of clinical next generation sequencing (NGS) based genetic testing. *J Pathol Inform* [Internet]. 2022 Jan 1 [cited 2023 Oct 16];13:100130. Available from: [/pmc/articles/PMC9577137/](https://pmc/articles/PMC9577137/)
20. Variation Annotation with SnpEff - Unipro UGENE User Manual v. 38 - WIKI [Internet]. [cited 2024 Jan 9]. Available from: <https://doc.ugene.net/wiki/display/UM38/Variation+Annotation+with+SnpEff>
21. Variant annotation with snpEff | In-depth-NGS-Data-Analysis-Course [Internet]. [cited 2024 Jan 9]. Available from: [https://hbctraining.github.io/In-depth-NGS-Data-Analysis-Course/sessionVI/lessons/03\\_annotation-snpEff.html](https://hbctraining.github.io/In-depth-NGS-Data-Analysis-Course/sessionVI/lessons/03_annotation-snpEff.html)
22. Medina-Escobedo G, Ridaaura-Sanz C. The VATER association. Vertebral defects, Anal atresia, T-E fistula with esophageal atresia, Radial and Renal dysplasia: a spectrum of associated defects. *J Pediatr* [Internet]. 1973;82(1):231–40. Available from: <https://pubmed.ncbi.nlm.nih.gov/4681850/>
23. Solomon BD. The etiology of VACTERL association: Current knowledge and hypotheses. *Am J Med Genet C Semin Med Genet* [Internet]. 2018 Dec;178(4):440–6. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/ajmg.c.31664>
24. Lautz TB, Mandelia A, Radhakrishnan J. VACTERL associations in children undergoing surgery for esophageal atresia and anorectal malformations: Implications for pediatric surgeons. *J Pediatr Surg*. 2014 Aug;50(8):1245–50.
25. Nora AH, Nora JJ. A syndrome of multiple congenital anomalies associated with teratogenic exposure. *Arch Environ Health* [Internet]. 1975;30(1):17–21. Available from: <https://pubmed.ncbi.nlm.nih.gov/1109267/>

26. Solomon BD. VACTERL/VATER association. *Orphanet J Rare Dis* [Internet]. 2011 Aug;6(1):1–12. Available from: <https://link.springer.com/articles/10.1186/1750-1172-6-56>
27. Pelizzo G, Chiricosta L, Mazzon E, Zuccotti GV, Avanzini MA, Croce S, et al. Discovering Genotype Variants in an Infant with VACTERL through Clinical Exome Sequencing: A Support for Personalized Risk Assessment and Disease Prevention. *Pediatric Reports* 2021, Vol 13, Pages 45-56 [Internet]. 2021 Jan;13(1):45–56. Available from: <https://www.mdpi.com/2036-7503/13/1/6/htm>
28. Reutter H, Hilger AC, Hildebrandt F, Ludwig M. Underlying genetic factors of the VATER/VACTERL association with special emphasis on the “Renal” phenotype. *Pediatric Nephrology*. 2016 Nov;31(11):2025–33.
29. The spectrum of congenital anomalies of the VATER association: an international study - PubMed [Internet]. Available from: <https://pubmed.ncbi.nlm.nih.gov/9215761/>
30. Czeizel A, Ludányi I. An aetiological study of the VACTERL-association. *Eur J Pediatr* [Internet]. 1985 Nov;144(4):331–7. Available from: <https://pubmed.ncbi.nlm.nih.gov/4076249/>
31. Schramm C, Draaken M, Bartels E, Boemers TM, Aretz S, Brockschmidt FF, et al. De novo microduplication at 22q11.21 in a patient with VACTERL association. *Eur J Med Genet*. 2011 Jan;54(1):9–13.
32. Lubinsky M. The VACTERL association: mosaic mitotic aneuploidy as a cause and a model. *Journal of Assisted Reproduction and Genetics*. 2019 Aug 15;36(8):1549.
33. Bryant S V., Gardiner DM. The relationship between growth and pattern formation. *Regeneration*. 2016 Apr 1;3(2):103–22.
34. Lubinsky M. The VACTERL association: mosaic mitotic aneuploidy as a cause and a model. *J Assist Reprod Genet* [Internet]. 2019 Aug;36(8):1549. Available from: </pmc/articles/PMC6708033/>
35. Bryant S V, Gardiner DM. The relationship between growth and pattern formation. *Regeneration* [Internet]. 2016 Apr;3(2):103–22. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/reg2.55>
36. Opitz JM. The developmental field concept. *Am J Med Genet*. 1985;21(1):1–11.
37. Li Y, Liu P, Wang W, Jia H, Bai Y, Yuan Z, et al. A novel genotype-phenotype between persistent-cloaca-related VACTERL and mutations of 8p23 and 12q23.1. *Pediatr Res* [Internet]. 2023 Dec; Available from: <https://pubmed.ncbi.nlm.nih.gov/38135728/>
38. van de Putte R, van Rooij IALM, Marcelis CLM, Guo M, Brunner HG, Addor MC, et al. Spectrum of congenital anomalies among VACTERL cases: a EUROCAT population-based study. *Pediatr Res* [Internet]. 2020 Feb;87(3):541–9. Available from: <https://pubmed.ncbi.nlm.nih.gov/31499513/>
39. van de Putte R, van Rooij IALM, Haanappel CP, Marcelis CLM, Brunner HG, Addor MC, et al. Maternal risk factors for the VACTERL association: A EUROCAT case-control study. *Birth Defects Res* [Internet]. 2020 May;112(9):688–98. Available from: <https://pubmed.ncbi.nlm.nih.gov/32319733/>

40. Van Rooij IALM, Wijers CHW, Rieu PNMA, Hendriks HS, Brouwers MM, Knoers N V., et al. Maternal and paternal risk factors for anorectal malformations: A Dutch case-control study. *Birth Defects Research Part A: Clinical and Molecular Teratology*. 2010 Mar 1;88(3):152–8.
41. Barbujani G, Russo A, Farabegoli A, Calzolari E, Rao DC. Inferences on the inheritance of congenital anomalies from temporal and spatial patterns of occurrence. *Genetic Epidemiology*. 1989;6(4):537–52.
42. Solomon BD, Pineda-Alvarez DE, Raam MS, Cummings DAT. Evidence for inheritance in patients with VACTERL association. *Human Genetics*. 2010 Jun 6;127(6):731–3.
43. Van Rooij IALM, Wijers CHW, Rieu PNMA, Hendriks HS, Brouwers MM, Knoers N V, et al. Maternal and paternal risk factors for anorectal malformations: A Dutch case-control study. *Birth Defects Res A Clin Mol Teratol* [Internet]. 2010 Mar;88(3):152–8. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/bdra.20649>
44. Barbujani G, Russo A, Farabegoli A, Calzolari E, Rao DC. Inferences on the inheritance of congenital anomalies from temporal and spatial patterns of occurrence. *Genet Epidemiol*. 1989;6(4):537–52.
45. Solomon BD, Pineda-Alvarez DE, Raam MS, Cummings DAT. Evidence for inheritance in patients with VACTERL association. *Hum Genet* [Internet]. 2010 Jun;127(6):731–3. Available from: <https://link.springer.com/article/10.1007/s00439-010-0814-7>
46. van de Putte R, van Rooij IALM, Haanappel CP, Marcelis CLM, Brunner HG, Addor MC, et al. Maternal risk factors for the VACTERL association: A EUROCAT case-control study. *Birth defects research*. 2020 May 15;112(9):688–98.
47. Nora AH, Nora JJ. A syndrome of multiple congenital anomalies associated with teratogenic exposure. *Archives of environmental health*. 1975;30(1):17–21.
48. Solomon BD, Bear KA, Kimonis V, de Klein A, Scott DA, Shaw-Smith C, et al. Clinical geneticists' views of VACTERL/VATER association. *Am J Med Genet A* [Internet]. 2012 Dec;158A(12):3087–100. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/ajmg.a.35638>
49. Solomon BD. VACTERL/VATER association. *Orphanet Journal of Rare Diseases*. 2011 Aug 16;6(1):1–12.
50. Gedikbasi A, Yazarbas K, Yildirim G, Yildirim D, Arslan O, Gul A, et al. Prenatal diagnosis of VACTERL syndrome and partial caudal regression syndrome: A previously unreported association. *Journal of Clinical Ultrasound* [Internet]. 2009 Oct;37(8):464–6. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/jcu.20590>
51. Chen Y, Liu Z, Chen J, Zuo Y, Liu S, Chen W, et al. The genetic landscape and clinical implications of vertebral anomalies in VACTERL association. *Journal of medical genetics*. 2016 Jul 1;53(7):431–7.
52. Solomon BD, Bear KA, Kimonis V, de Klein A, Scott DA, Shaw-Smith C, et al. Clinical geneticists' views of VACTERL/VATER association. *American Journal of Medical Genetics Part A*. 2012 Dec 1;158A(12):3087–100.

53. Temtamy SA, Miller JD. Extending the scope of the VATER association: Definition of the VATER syndrome. *The Journal of Pediatrics*. 1974 Sep 1;85(3):345–9.
54. Temtamy SA, Miller JD. Extending the scope of the VATER association: Definition of the VATER syndrome. *J Pediatr*. 1974 Sep;85(3):345–9.
55. Chen Y, Liu Z, Chen J, Zuo Y, Liu S, Chen W, et al. The genetic landscape and clinical implications of vertebral anomalies in VACTERL association. *J Med Genet* [Internet]. 2016 Jul;53(7):431–7. Available from: <https://pubmed.ncbi.nlm.nih.gov/27084730/>
56. Wessels MW, Kuchinka B, Heydanus R, Smit BJ, Dooijes D, De Krijger RR, et al. Polyalanine expansion in the ZIC3 gene leading to X-linked heterotaxy with VACTERL association: a new polyalanine disorder? *J Med Genet* [Internet]. 2010 May;47(5):351–5. Available from: <https://jmg.bmj.com/content/47/5/351>
57. Wessels MW, Kuchinka B, Heydanus R, Smit BJ, Dooijes D, De Krijger RR, et al. Polyalanine expansion in the ZIC3 gene leading to X-linked heterotaxy with VACTERL association: a new polyalanine disorder? *Journal of Medical Genetics*. 2010 May 1;47(5):351–5.
58. Li S, Liu S, Chen W, Yuan Y, Gu R, Song Y, et al. A novel ZIC3 gene mutation identified in patients with heterotaxy and congenital heart disease. *Scientific Reports* 2018 8:1. 2018 Aug 17;8(1):1–12.
59. Solomon BD. The etiology of VACTERL association: Current knowledge and hypotheses. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*. 2018 Dec 1;178(4):440–6.
60. Li S, Liu S, Chen W, Yuan Y, Gu R, Song Y, et al. A novel ZIC3 gene mutation identified in patients with heterotaxy and congenital heart disease. *Scientific Reports* 2018 8:1 [Internet]. 2018 Aug;8(1):1–12. Available from: <https://www.nature.com/articles/s41598-018-30204-3>
61. Ngan ESW, Kim KH, Hui CC. Sonic Hedgehog Signaling and VACTERL Association. *Molecular Syndromology*. 2013 Feb;4(1–2):32.
62. Ngan ESW, Kim KH, Hui CC. Sonic Hedgehog Signaling and VACTERL Association. *Mol Syndromol* [Internet]. 2013 Feb;4(1–2):32. Available from: [/pmc/articles/PMC3638778/](https://pubmed.ncbi.nlm.nih.gov/24363877/)
63. Kim PCW, Mo R, Hui CC. Murine models of VACTERL syndrome: Role of sonic hedgehog signaling pathway. *J Pediatr Surg*. 2001 Feb;36(2):381–4.
64. Garcia-Barceló MM, Wong KKY, Lui VCH, Yuan ZW, So MT, Ngan ESW, et al. Identification of a HOXD13 mutation in a VACTERL patient. *Am J Med Genet A* [Internet]. 2008 Dec;146A(24):3181–5. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/ajmg.a.32426>
65. Quinonez SC, Innis JW. Human HOX gene disorders. *Mol Genet Metab*. 2014 Jan;111(1):4–15.
66. Mundt E, Bates MD. Genetics of Hirschsprung disease and anorectal malformations. *Semin Pediatr Surg*. 2010 May;19(2):107–17.

67. Agochukwu NB, Pineda-Alvarez DE, Keaton AA, Warren-Mora N, Raam MS, Kamat A, et al. Analysis of FOXF1 and the FOX gene cluster in patients with VACTERL association. *Eur J Med Genet*. 2011 May;54(3):323–8.
68. Hilger AC, Halbritter J, Pennimpede T, van der Ven A, Sarma G, Braun DA, et al. Targeted Resequencing of 29 Candidate Genes and Mouse Expression Studies Implicate ZIC3 and FOXF1 in Human VATER/VACTERL Association. *Hum Mutat* [Internet]. 2015 Dec;36(12):1150–4. Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/humu.22859>
69. Matissek SJ, ElSawa SF. GLI3: a mediator of genetic diseases, development and cancer. *Cell Communication and Signaling* 2020 18:1 [Internet]. 2020 Apr;18(1):1–20. Available from: <https://link.springer.com/articles/10.1186/s12964-020-00540-x>
70. Qin D. Next-generation sequencing and its clinical application. *Cancer Biol Med* [Internet]. 2019;16(1):4. Available from: [/pmc/articles/PMC6528456/](https://pubmed.ncbi.nlm.nih.gov/3084567/)
71. Behjati S, Tarpey PS. What is next generation sequencing? *Arch Dis Child Educ Pract Ed* [Internet]. 2013 Dec;98(6):236. Available from: [/pmc/articles/PMC3841808/](https://pubmed.ncbi.nlm.nih.gov/241808/)
72. Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, et al. Guidelines for Validation of Next-Generation Sequencing–Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. *J Mol Diagn* [Internet]. 2017 May;19(3):341. Available from: [/pmc/articles/PMC6941185/](https://pubmed.ncbi.nlm.nih.gov/2741185/)
73. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. *J Biomed Biotechnol*. 2012;2012.
74. Santamaría González M, Miguel Lezana Rosales J. Aplicaciones Clínicas De Las Técnicas Actuales De Biología Molecular Técnicas De Secuenciación Masiva (Ngs). *Cont Lab Clin* [Internet]. 2017;37:33–40. Available from: [https://www](https://www.sciencedirect.com/science/article/pii/S0959631117300000).
75. Singleton AB. Exome sequencing: A transformative technology. *Lancet Neurol* [Internet]. 2011 Oct;10(10):942–6. Available from: <http://www.thelancet.com/article/S147444221170196X/fulltext>
76. Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. Exome Sequencing: Current and Future Perspectives. *G3 Genes|Genomes|Genetics* [Internet]. 2015 Aug;5(8):1543–50. Available from: <https://dx.doi.org/10.1534/g3.115.018564>
77. Simon C, Daniel R. Metagenomic analyses: Past and future trends. *Appl Environ Microbiol* [Internet]. 2011 Feb;77(4):1153–61. Available from: <https://journals.asm.org/doi/10.1128/AEM.02345-10>
78. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. *Elife*. 2019 Sep;8.
79. Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, et al. The NIH Human Microbiome Project. *Genome Res* [Internet]. 2009 Dec;19(12):2317–23. Available from: <https://pubmed.ncbi.nlm.nih.gov/19819907/>
80. National Center for Biotechnology Information [Internet]. [cited 2023 Dec 24]. Available from: <https://www.ncbi.nlm.nih.gov/>



81. Pelizzo G, Chiricosta L, Mazzon E, Zuccotti GV, Avanzini MA, Croce S, et al. Discovering Genotype Variants in an Infant with VACTERL through Clinical Exome Sequencing: A Support for Personalized Risk Assessment and Disease Prevention. *Pediatr Rep* [Internet]. 2021 Jan 5 [cited 2024 Jan 10];13(1):45. Available from: [/pmc/articles/PMC7838983/](https://pubmed.ncbi.nlm.nih.gov/38888883/)
82. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. [cited 2023 Dec 18]. Available from: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
83. University of Missouri. FASTQC Report.
84. Quality control: Assessing FASTQC results | Introduction to RNA-Seq using high-performance computing - ARCHIVED [Internet]. [cited 2023 Dec 19]. Available from: [https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc\\_fastqc\\_assessment.html](https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html)
85. Why does the per base sequence quality decrease over the read in Illumina? [Internet]. [cited 2023 Dec 19]. Available from: <https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina>
86. Per Base Sequence Content [Internet]. [cited 2023 Dec 19]. Available from: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html>
87. Romiguier J, Ranwez V, Douzery EJP, Galtier N. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Res* [Internet]. 2010 Aug [cited 2023 Dec 19];20(8):1001. Available from: [/pmc/articles/PMC2909565/](https://pubmed.ncbi.nlm.nih.gov/209565/)
88. Shi H, Xu X. Learning the Sequences Quality Control of Bioinformatics Analysis Method. 2016.
89. Duplicate Sequences [Internet]. [cited 2023 Dec 24]. Available from: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/8%20Duplicate%20Sequences.html>
90. USADELLAB.org - Trimmomatic: A flexible read trimming tool for Illumina NGS data [Internet]. [cited 2023 Dec 25]. Available from: <http://www.usadellab.org/cms/index.php?page=trimmomatic>
91. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* [Internet]. 2014 Aug 8 [cited 2023 Dec 25];30(15):2114. Available from: [/pmc/articles/PMC4103590/](https://pubmed.ncbi.nlm.nih.gov/24768208/)
92. GEO Accession viewer [Internet]. [cited 2024 Jan 9]. Available from: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2076238>
93. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* [Internet]. 2009 Jul 7 [cited 2023 Dec 27];25(14):1754. Available from: [/pmc/articles/PMC2705234/](https://pubmed.ncbi.nlm.nih.gov/19261171/)
94. Lorente P. Automatización de los procesos de alineamiento y búsqueda de variantes en secuencias de ADN. 2012;

95. Homo sapiens genome assembly GRCh38 - NCBI - NLM [Internet]. [cited 2023 Dec 27]. Available from: [https://www.ncbi.nlm.nih.gov/datasets/genome/GCF\\_000001405.26/](https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/)
96. Pan B, Kusko R, Xiao W, Zheng Y, Liu Z, Xiao C, et al. Similarities and differences between variants called with human reference genome HG19 or HG38. BMC Bioinformatics [Internet]. 2019 Mar 14 [cited 2023 Dec 27];20(Suppl 2). Available from: </pmc/articles/PMC6419332/>
97. Li H. Exploring single-sample snp and indel calling with whole-genome de novo assembly. Bioinformatics. 2012 Jul;28(14):1838–44.
98. Samtools [Internet]. [cited 2023 Dec 28]. Available from: <https://www.htslib.org/>
99. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet]. 2009 Aug 8 [cited 2023 Dec 28];25(16):2078. Available from: </pmc/articles/PMC2723002/>
100. Variant Calling using Freebayes [Internet]. [cited 2023 Dec 30]. Available from: <https://manual.omicsbox.biobam.com/user-manual/omicsbox-modules/module-genetic-variation/variant-calling/variant-calling-using-freebayes/>
101. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012;
102. Variant Calling part 2 (Galaxy) - Bioinformatics Documentation [Internet]. [cited 2024 Jan 9]. Available from: [https://melbournebioinformatics.github.io/MelBioInf\\_docs/tutorials/var\\_detect\\_advanced/var\\_detect\\_advanced/](https://melbournebioinformatics.github.io/MelBioInf_docs/tutorials/var_detect_advanced/var_detect_advanced/)
103. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) [Internet]. 2012 Apr 4 [cited 2023 Dec 30];6(2):80. Available from: </pmc/articles/PMC3679285/>
104. Introduction - SnpEff & SnpSift [Internet]. [cited 2023 Dec 30]. Available from: <https://pcingola.github.io/SnpEff/snpeff/introduction/>
105. Yen JL, Garcia S, Montana A, Harris J, Chervitz S, Morra M, et al. A variant by any name: quantifying annotation discordance across tools and clinical databases. Genome Med [Internet]. 2017 Jan 26 [cited 2024 Jan 3];9(1). Available from: </pmc/articles/PMC5267466/>
106. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics [Internet]. 2011 Aug 8 [cited 2024 Jan 3];27(15):2156. Available from: </pmc/articles/PMC3137218/>
107. Lyon MS, Andrews SJ, Elsworth B, Gaunt TR, Hemani G, Marcora E. The variant call format provides efficient and robust storage of GWAS summary statistics. Genome Biol [Internet]. 2021 Dec 1 [cited 2024 Jan 3];22(1). Available from: </pmc/articles/PMC7805039/>
108. samtools-flagstat(1) manual page [Internet]. [cited 2024 Jan 5]. Available from: <https://www.htslib.org/doc/samtools-flagstat.html>

109. bcftools [Internet]. [cited 2024 Jan 5]. Available from:  
<https://www.htslib.org/doc/1.0/bcftools.html>
110. Homo sapiens epoxide hydrolase 2 (EPHX2), transcript variant 1, mRNA - Nucleotide - NCBI [Internet]. [cited 2024 Jan 10]. Available from:  
[https://www.ncbi.nlm.nih.gov/nucore/NM\\_001979](https://www.ncbi.nlm.nih.gov/nucore/NM_001979)
111. Ha NT, Lee CH. Roles of Farnesyl-Diphosphate Farnesyltransferase 1 in Tumour and Tumour Microenvironments. *Cells* [Internet]. 2020 Nov 1 [cited 2024 Jan 10];9(11):1–33. Available from: [/pmc/articles/PMC7693003/](https://pmc/articles/PMC7693003/)
112. Jiang H, Tang E, Chen Y, Liu H, Zhao Y, Lin M, et al. Squalene synthase predicts poor prognosis in stage I-III colon adenocarcinoma and synergizes squalene epoxidase to promote tumor progression. *Cancer Sci*. 2022 Mar 1;113(3):971–85.
113. Fukuma Y, Matsui H, Koike H, Sekine Y, Shechter I, Ohtake N, et al. Role of squalene synthase in prostate cancer risk and the biological aggressiveness of human prostate cancer. *Prostate Cancer Prostatic Dis* [Internet]. 2012 Dec [cited 2024 Jan 10];15(4):339–45. Available from: <https://pubmed.ncbi.nlm.nih.gov/22546838/>
114. Homo sapiens toll like receptor 8 (TLR8), transcript variant 1, mRNA - Nucleotide - NCBI [Internet]. [cited 2024 Jan 10]. Available from:  
[https://www.ncbi.nlm.nih.gov/nucore/NM\\_016610](https://www.ncbi.nlm.nih.gov/nucore/NM_016610)
115. Veneziani I, Alicata C, Moretta L, Maggi E. Human toll-like receptor 8 (TLR8) in NK cells: Implication for cancer immunotherapy. *Immunol Lett* [Internet]. 2023 [cited 2024 Jan 10];261:13–6. Available from: <https://doi.org/10.1016/j.imlet.2023.07.003>
116. Homo sapiens integrin linked kinase (ILK), transcript variant 1, mRNA - Nucleotide - NCBI [Internet]. [cited 2024 Jan 10]. Available from:  
[https://www.ncbi.nlm.nih.gov/nucore/NM\\_004517](https://www.ncbi.nlm.nih.gov/nucore/NM_004517)
117. Isabel Serrano Martínez. CONSECUENCIAS ESTRUCTURALES Y FUNCIONALES DE LA DELECCIÓN CONDICIONAL DE LA QUINASA LIGADA A INTEGRINAS (ILK) EN ANIMALES ADULTOS. 2010;

# 9. ANNEXO 1: Informe FASTQC

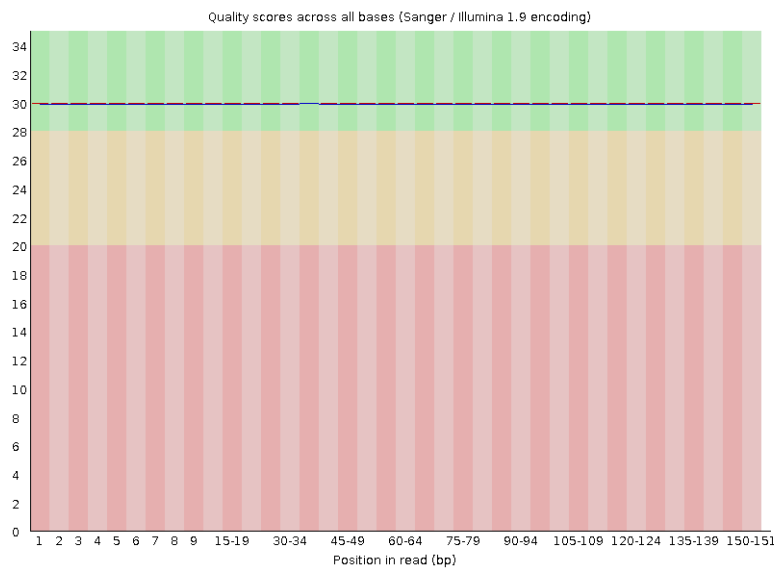
## Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

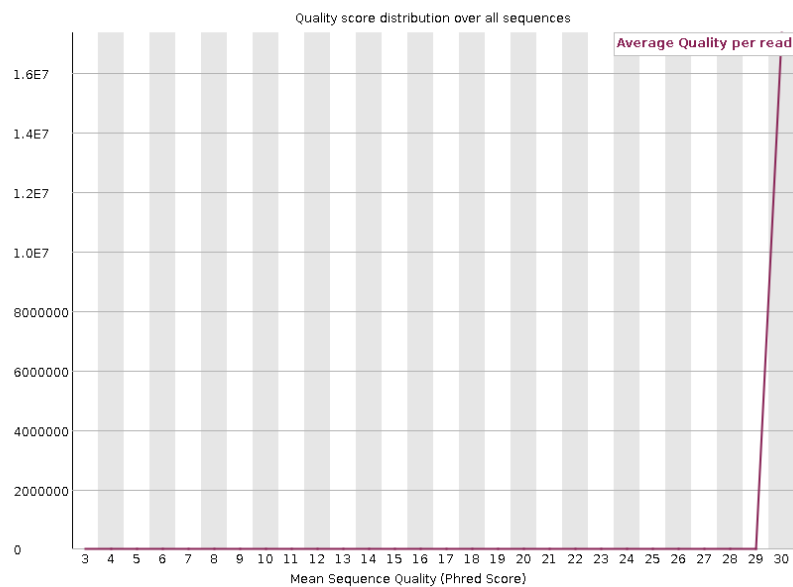
## Basic Statistics

Measure	Value
Filename	SRR12568924.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	17367368
Total Bases	2.4 Gbp
Sequences flagged as poor quality	0
Sequence length	32-151
%GC	46

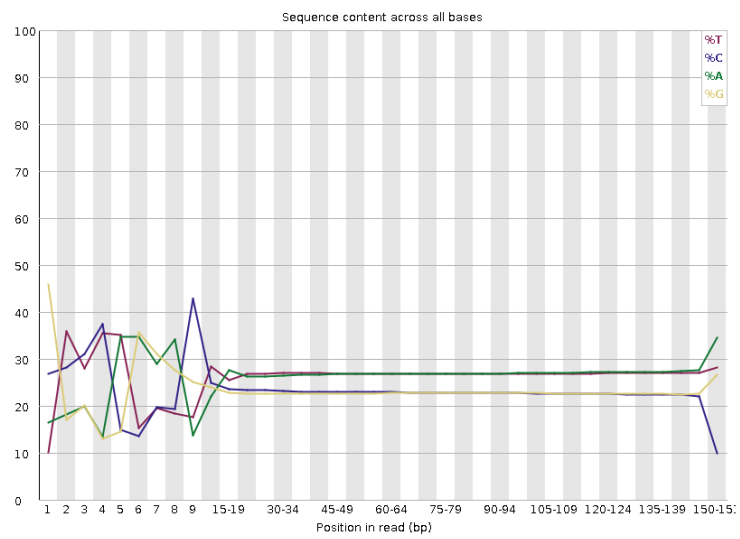
## Per base sequence quality



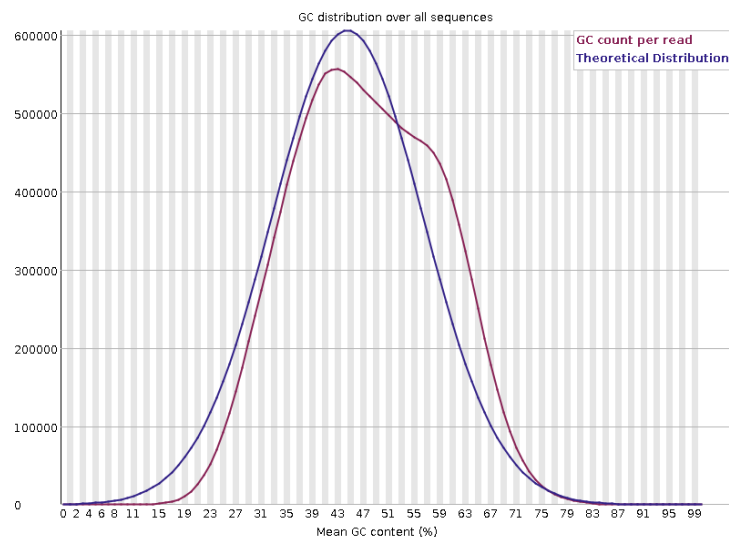
## Per sequence quality scores



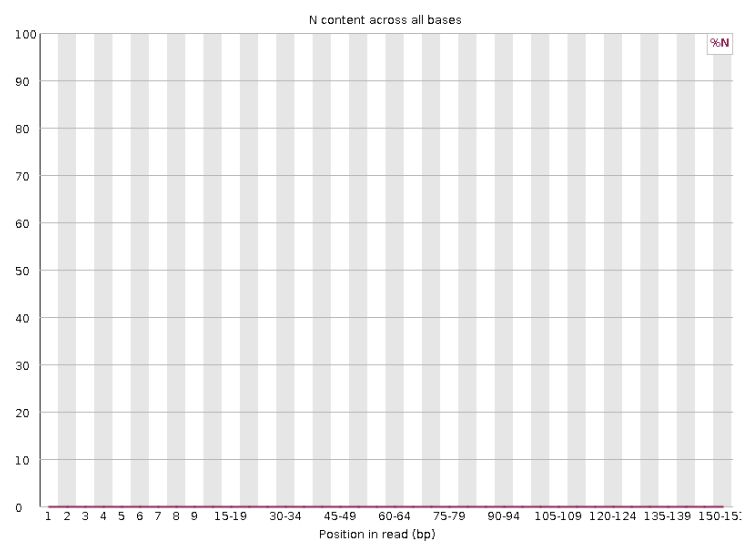
### ✖ Per base sequence content



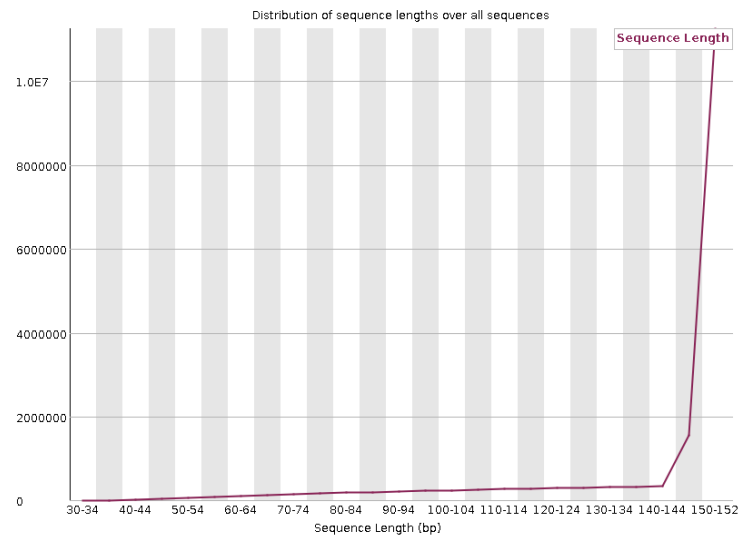
### ⚠ Per sequence GC content



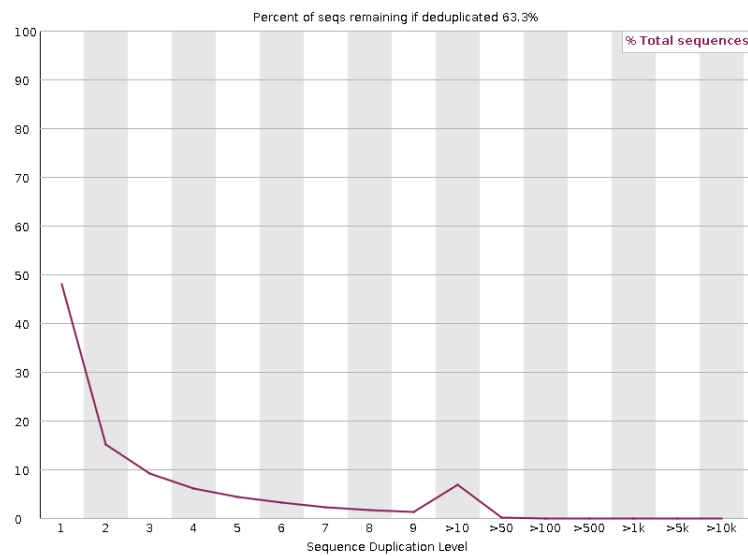
### ✔ Per base N content



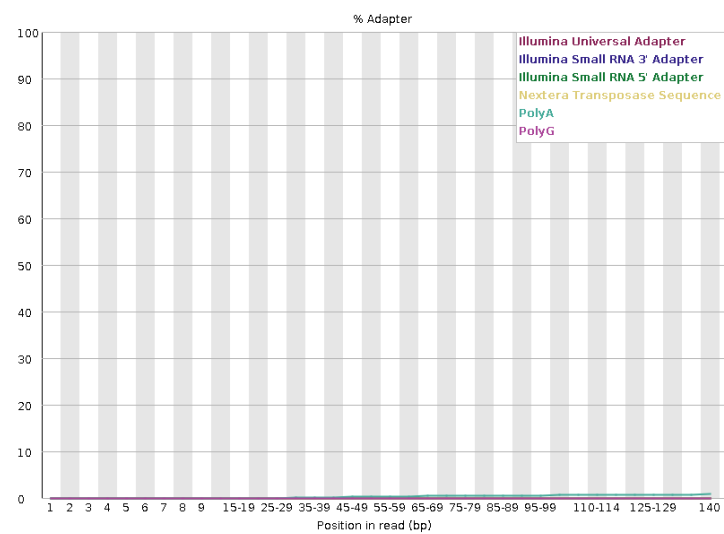
## ⚠ Sequence Length Distribution



## ⚠ Sequence Duplication Levels



## ✅ Adapter Content



# 10. ANNEXO 2: Informe FASTQC (trimmed)

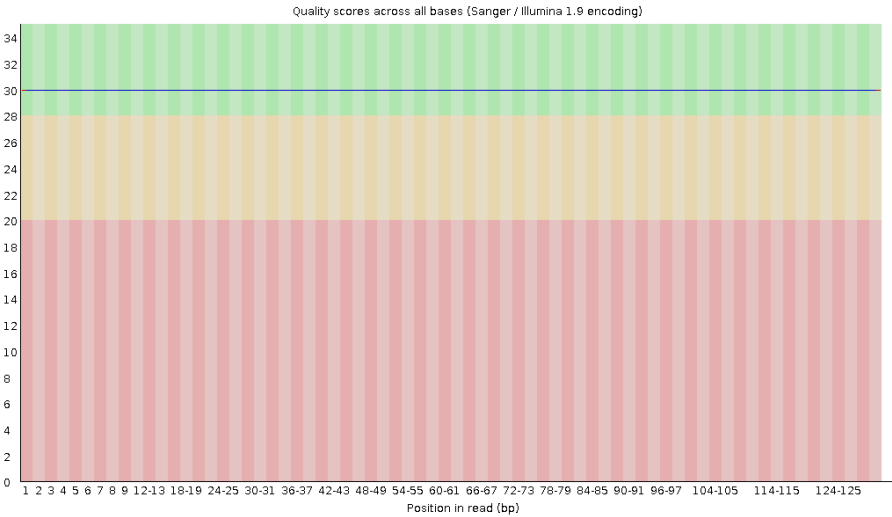
## Summary

- ✔ [Basic Statistics](#)
- ✔ [Per base sequence quality](#)
- ✔ [Per sequence quality scores](#)
- ✔ [Per base sequence content](#)
- ⚠ [Per sequence GC content](#)
- ✔ [Per base N content](#)
- ⚠ [Sequence Length Distribution](#)
- ⚠ [Sequence Duplication Levels](#)
- ✔ [Overrepresented sequences](#)
- ✔ [Adapter Content](#)

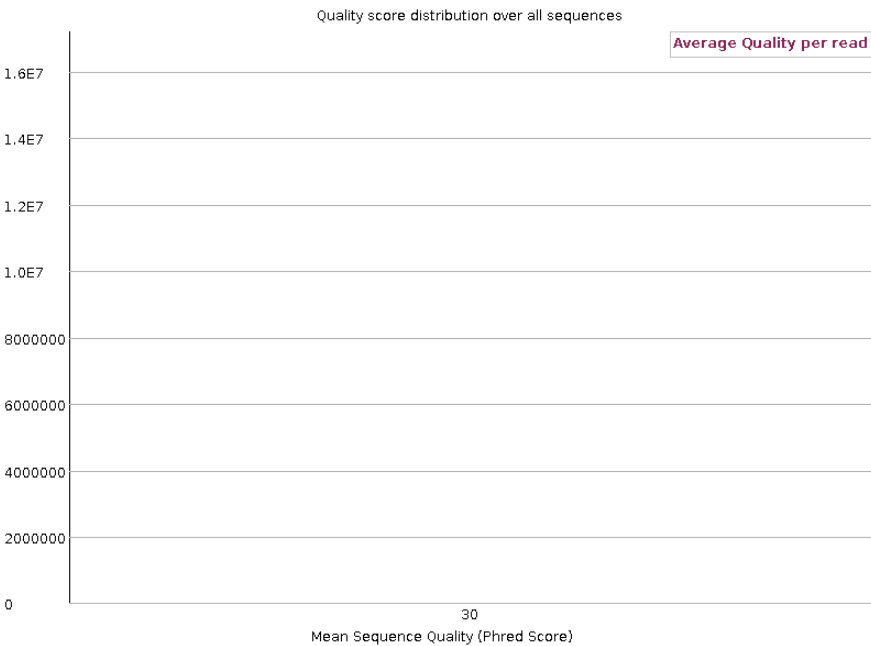
## ✔ Basic Statistics

Measure	Value
Filename	SRR12568924_trimmed.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	17198228
Total Bases	2 Gbp
Sequences flagged as poor quality	0
Sequence length	36-130
%GC	45

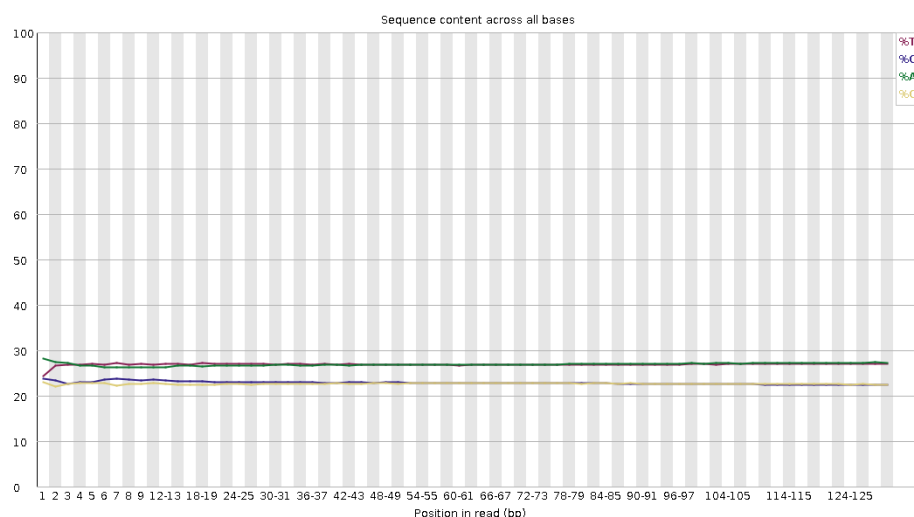
## ✔ Per base sequence quality



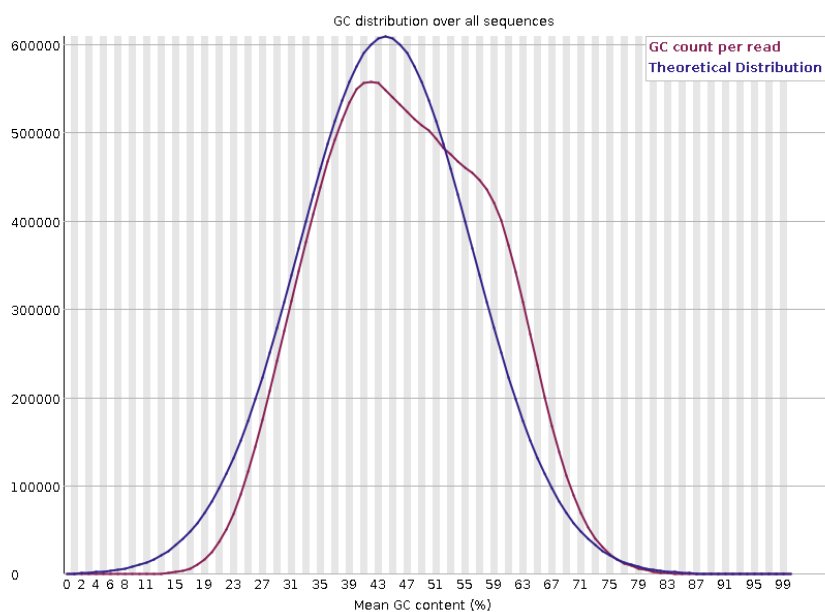
## ✔ Per sequence quality scores



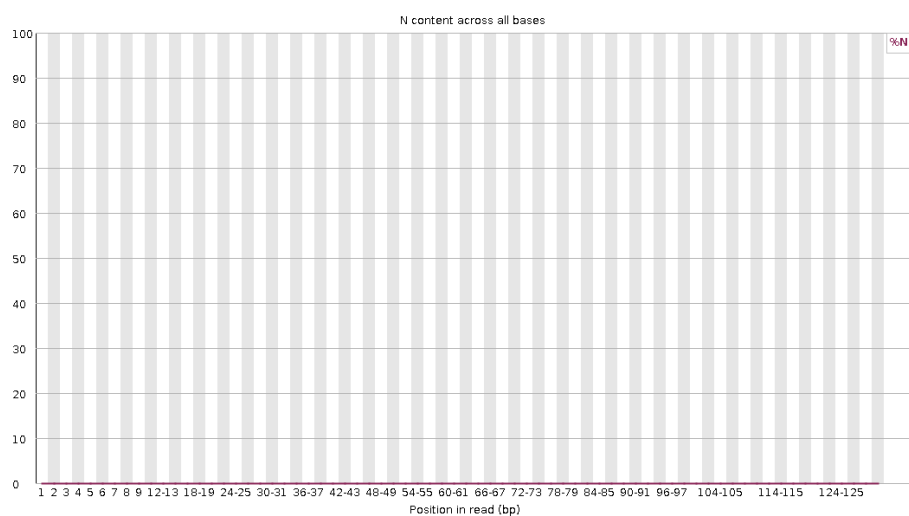
### ✔ Per base sequence content



### ⚠ Per sequence GC content

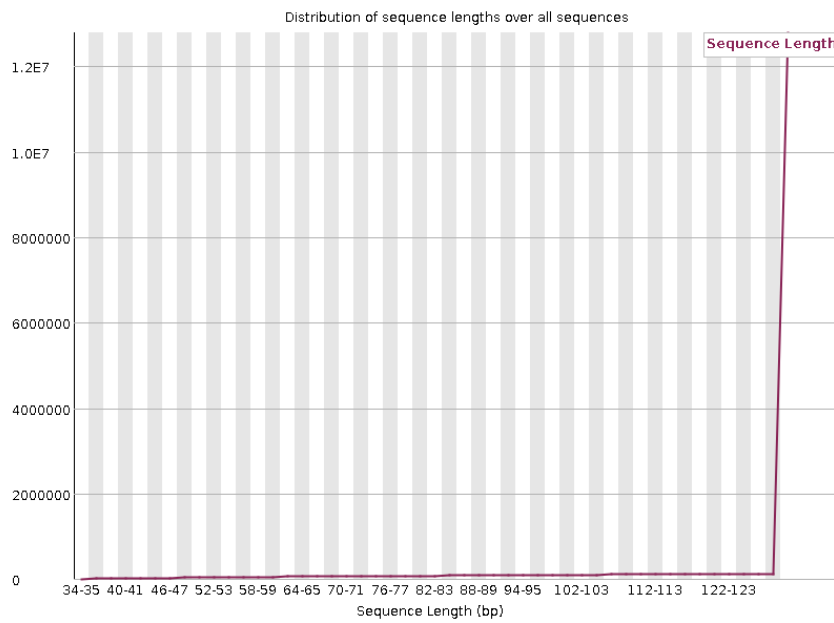


### ✔ Per base N content

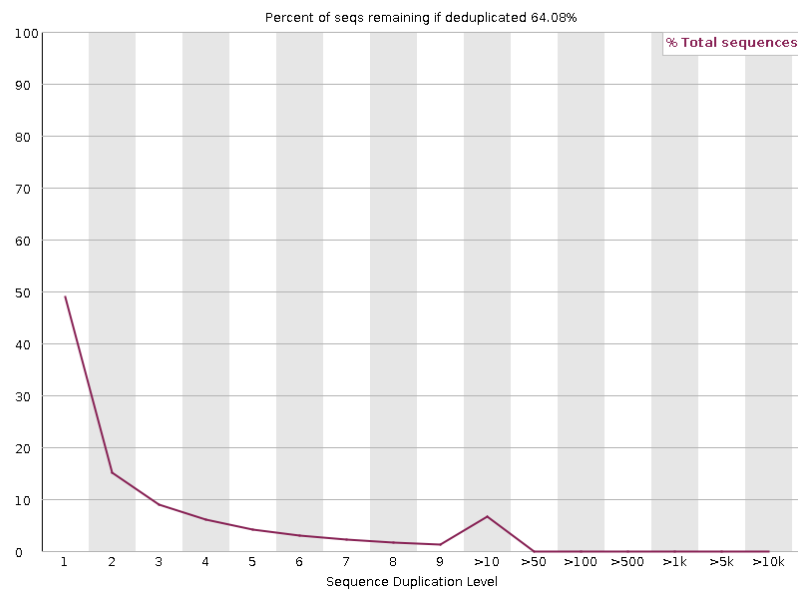




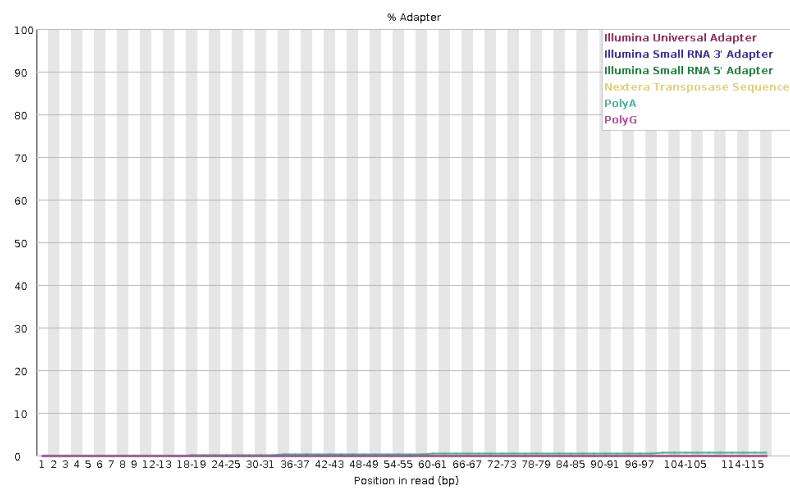
## Sequence Length Distribution



## Sequence Duplication Levels



## Adapter Content



11. ANNEXO 3: SnpEff report

Summary	
Genome	GRCh38.86
Date	2024-01-06 18:48
SnpEff version	SnpEff 4.3t (build 2017-11-24 10:18), by Pablo Cingolani
Command line arguments	SnpEff -i vcf -o vcf -no SYNONYMOUS_START -no SYNONYMOUS_CODING -no SYNONYMOUS_STOP -stats /corral4/main/jobs/054/731/54731167/outputs/dataset_0236c68d-4328-49d7-8a35-3b9ae094ab33.dat GRCh38.86 /corral4/main/objects/2/d/d/dataset_2dd102bc-188b-429b-bcd3-ee717c75ad94.dat
Warnings	26,297
Errors	581
Number of lines (input file)	27,832
Number of variants (before filter)	28,340
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	28,340
Number of known variants (i.e. non-empty ID)	0 ( 0% )
Number of multi-allelic VCF entries (i.e. more than two alleles)	508
Number of effects	181,355
Genome total length	47,871,129,571
Genome effective length	3,088,269,860
Variant rate	1 variant every 108,972 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
1	248,956,422	2,590	96,122
2	242,193,529	2,257	107,307
3	198,295,559	1,678	118,173
4	190,214,555	1,440	132,093
5	181,538,259	1,460	124,341
6	170,805,979	1,504	113,567
7	159,345,973	1,481	107,593
8	145,138,636	1,076	134,887
9	138,394,717	1,193	116,005
10	133,797,422	1,319	101,438
11	135,086,622	1,588	85,067
12	133,275,309	1,490	89,446
13	114,364,328	757	151,075
14	107,043,718	736	145,439
15	101,991,189	1,001	101,889
16	90,338,345	1,025	88,134
17	83,257,441	1,198	69,497
18	80,373,285	537	149,670
19	58,617,616	1,141	51,373
20	64,444,167	543	118,681
21	46,709,983	357	130,840
22	50,818,468	530	95,883

Number variants by type

Type	Total
SNP	23,938
MNP	494
INS	1,570
DEL	2,105
MIXED	233
INV	0
DUP	0
BND	0
INTERVAL	0
Total	28,340

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	2,134	1.18%
LOW	15,416	8.528%
MODERATE	8,671	4.797%
MODIFIER	154,553	85.495%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	8,016	42.292%
NONSENSE	51	0.269%
SILENT	10,887	57.439%

Missense / Silent ratio: 0.7363

Number of effects by type and region

Type			Region		
Type (alphabetical order)	Count	Percent			
3_prime_UTR_variant	3,877	2.099%			
5_prime_UTR_premature_start_codon_gain_variant	210	0.114%			
5_prime_UTR_variant	1,624	0.879%			
TFBS_ablation	1	0.001%			
TF_binding_site_variant	174	0.094%			
conservative_inframe_deletion	76	0.041%			
conservative_inframe_insertion	67	0.036%			
disruptive_inframe_deletion	107	0.058%			
disruptive_inframe_insertion	52	0.028%			
downstream_gene_variant	27,802	15.054%			
frameshift_variant	70	0.038%			
initiator_codon_variant	1	0.001%			
intergenic_region	1,570	0.85%			
intragenic_variant	9	0.005%			
intron_variant	93,336	50.537%			
missense_variant	8,187	4.433%			
non_coding_transcript_exon_variant	9,284	5.027%			
non_coding_transcript_variant	6	0.003%			
protein_protein_contact	92	0.05%			
sequence_feature	1,170	0.634%			
splice_acceptor_variant	76	0.041%			
splice_donor_variant	62	0.034%			
splice_region_variant	3,828	2.073%			
start_lost	13	0.007%			
stop_gained	56	0.03%			
stop_lost	10	0.005%			
stop_retained_variant	6	0.003%			
structural_interaction_variant	1,768	0.957%			
synonymous_variant	10,916	5.911%			
upstream_gene_variant	20,237	10.957%			

Type (alphabetical order)	Count	Percent
DOWNSTREAM	27,802	15.379%
EXON	30,257	16.737%
INTERGENIC	1,570	0.868%
INTRON	90,178	49.884%
MOTIF	175	0.097%
SPLICE_SITE_ACCEPTOR	74	0.041%
SPLICE_SITE_DONOR	57	0.032%
SPLICE_SITE_REGION	3,531	1.953%
TRANSCRIPT	1,185	0.656%
UPSTREAM	20,237	11.195%
UTR_3_PRIME	3,875	2.144%
UTR_5_PRIME	1,833	1.014%

Base changes (SNPs)

	A	C	G	T
A	0	817	4,086	630
C	963	0	1,055	4,370
G	4,482	1,055	0	903
T	627	4,065	885	0

Ts/Tv (transitions / transversions)

Transitions	23,723
Transversions	9,450
Ts/Tv ratio	2.5104

Amino acid changes

How to read this table:

- Rows are reference amino acids and columns are changed amino acids. E.g. Row 'A' column 'E' indicates how many 'A' amino acids have been replaced by 'E' amino acids.
- Red background colors indicate that more changes happened (heat-map).
- Diagonals are indicated using grey background color
- WARNING: This table may include different translation codon tables (e.g. mamalian DNA and mitochondrial DNA).

	*	-	?	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
*	6	2											1					1	1			1	4
-	1		19	16	5	2	1		19	2	8	2	3			7	63	4	40	4	1		1
?																							
A		17		1,164		10	28		33							60			59	227	160		
C		5			177				4	2								83	35			4	31
D		16		34		688	83		45	31					103				2		15		15
E		6	53		34		76	423		72		103	2		1		55				19		
F			1			13			289	6	29		57						32		11		9
G		28	2	48	13	82	41		707					2	3			85	134		30	11	
H			5				23			409					4	14	42	142					35
I			7					13		1	383	5	39	38	11			1	9	99	268		
K		1	4			6	87				1	218		1	42		31	101		3			
L			36					71		3	20		1,237	45		170	6	30	55		110		
M			2		3						62		27		2					116	134		
N			2			69				19	10	30			428				129	13			5
P			45		71				7				171			1,066		62	158	88			
Q	15	139					45			55		32	28			35	480	180		3		3	
R	15	28			91				78	160	5	99	16	1		46	240	583	45	38		106	
S		15		84	46			26	92		12	1	66		158	133		34	1,062	79	17	12	
T			36		211			1		113	22			115	19	41		6	38	1,006			
V			17		142		7	12	13	29		332		58	200						535		
W	13	4				5			2								6	44	2			6	
Y	5	10				20	2		4		55				1				11				352