



Universidad Austral de Chile

Facultad de Ciencias de la Ingeniería
Escuela de Ingeniería Civil en Informática

ELABORACIÓN DE PIPELINE DE ANÁLISIS BIOINFORMÁTICO PARA IDENTIFICACIÓN DE VARIANTES PATOGENICAS EN EL GEN PKD1 A PARTIR DE SECUENCIACIÓN OXFORD NANOPORE MINION

Proyecto para optar al título de
Ingeniero Civil en Informática

PROFESOR PATROCINANTE:
JORGE MATORANA
DOCTOR EN INFORMÁTICA

PROFESORA CO-PATROCINANTE:
PAOLA KRALL
DOCTORA EN CIENCIAS MENCIÓN BIOLOGÍA
CELULAR Y MOLECULAR

DIEGO BENJAMÍN MILLAR CORONADO

VALDIVIA – CHILE
2024

ÍNDICE

ÍNDICE.....	I
ÍNDICE DE TABLAS.....	III
ÍNDICE DE FIGURAS.....	VI
RESUMEN.....	VII
ABSTRACT.....	VIII
1. INTRODUCCIÓN.....	1
1.1 Contexto y oportunidad.....	1
1.2 Estado del arte.....	2
1.3 Innovación propuesta y su impacto.....	4
1.3.1 Propuesta.....	4
1.3.2 Impacto social-económico.....	4
1.3.3 Impacto científico-tecnológico.....	5
1.4 Objetivos.....	5
1.4.1 Objetivo general.....	5
1.4.2 Objetivos específicos.....	5
2. MARCO TEÓRICO.....	7
2.1 Genética molecular.....	7
2.1.1 ADN.....	7
2.1.2 Dogma central de la biología molecular.....	8
2.1.2.1 Transcripción.....	8
2.1.2.2 Traducción.....	9
2.1.2.3 Replicación.....	10
2.1.3 Herencia.....	11
2.2 Enfermedades renales genéticas.....	12
2.2.1 Tipos de variantes.....	13
2.2.2 PKD1 y Poliquistosis Renal Autosómica Dominante.....	14
2.3 Diagnóstico de enfermedades mediante análisis genético.....	16
2.3.1 Análisis genético.....	16
2.3.2 Interpretación de variantes genéticas.....	17
2.3.3 Secuenciadores.....	20
2.3.3.1 Illumina MiSeq.....	22
2.3.3.2 Oxford Nanopore MinION.....	23
2.3.4 Pipeline de análisis bioinformático.....	24
2.3.4.1 Basecalling.....	25
2.3.4.2 Control de calidad.....	25
2.3.4.3 Trimming y filtrado de lecturas.....	26
2.3.4.4 Mapeo.....	26
2.3.4.5 Variant calling.....	26
2.3.4.6 Clasificación, anotación y filtrado de variantes.....	27

3. REVISIÓN SISTEMÁTICA.....	28
3.1 Objetivo de la revisión.....	28
3.2 Metodología.....	28
3.2.1 Pregunta de revisión.....	28
3.2.2 Fuente y cadena de búsqueda.....	28
3.2.3 Criterios de selección y exclusión.....	30
3.2.4 Extracción de la información.....	30
3.3 Resultados de la revisión.....	30
3.3.1 Artículos aceptados.....	30
3.3.2 Síntesis de los resultados.....	31
4. METODOLOGÍA Y RESULTADOS PARA LA EVALUACIÓN Y SELECCIÓN DE HERRAMIENTAS.....	35
4.1 Casos clínicos.....	35
4.2 Preparación de librerías y secuenciación.....	36
4.3 Búsqueda de herramientas.....	37
4.3.1 Basecalling y demultiplexación.....	37
4.3.2 Control de calidad.....	38
4.3.2.1 Control de calidad post-secuenciación.....	39
4.3.2.2 Control de calidad post-mapeo.....	41
4.3.3 Trimming y filtrado de lecturas.....	44
4.3.4 Mapeo.....	46
4.3.5 Clasificación, anotación y filtrado de variantes.....	52
4.3.6 Sistema de ranking de variantes.....	55
4.3.7 Variant calling.....	58
5. PIPELINE DE ANÁLISIS BIOINFORMÁTICO.....	68
5.1 Optimización de parámetros de trimming y filtrado.....	68
5.2 Pipeline bioinformático para Oxford Nanopore MinION.....	70
5.3 Manual de usuario.....	72
6. CONCLUSIÓN Y TRABAJO FUTURO.....	73
6.1 Conclusión y cumplimiento de objetivos.....	73
6.2 Limitaciones.....	74
6.3 Trabajo futuro.....	75
7. REFERENCIAS.....	76
ANEXO.....	86
Anexo A.....	86
Anexo B.....	112
Anexo C.....	118
Anexo D.....	122
Anexo E.....	124

ÍNDICE DE TABLAS

TABLA	PÁGINA
Tabla 1: Criterios para clasificar una variante como patogénica.....	18
Tabla 2: Criterios para clasificar una variante como benigna.....	19
Tabla 3: Reglas para clasificar variantes según los criterios cumplidos.....	21
Tabla 4: keywords y sinónimos.....	29
Tabla 5: Número de herramientas por utilidad.....	33
Tabla 6: Variantes confirmadas de cada paciente.....	35
Tabla 7: Herramientas para evaluar la calidad de la secuenciación.....	39
Tabla 8: Cumplimiento de requisitos de las distintas herramientas.....	41
Tabla 9: Herramientas para evaluar la calidad del mapeo.....	42
Tabla 10: Motivo de exclusión de herramientas de mapeo.....	47
Tabla 11: Tiempo de ejecución (en minutos) de las herramientas de mapeo para.....	49
Tabla 12: Porcentaje promedio de lecturas mapeadas.....	50
Tabla 13: Porcentaje de INDELs y mismatches en relación con las bases.....	51
Tabla 14: Porcentaje promedio de lecturas recortadas durante el mapeo.....	51
Tabla 15: Herramientas de predicción de patogenicidad.....	53
Tabla 16: Bases de datos de variantes utilizadas.....	55
Tabla 17: Criterios para clasificación de variantes sistema de ranking Yáñez-Krall.....	57
Tabla 18: Herramientas de variant calling seleccionadas.....	59
Tabla 19: Tiempo promedio en minutos de ejecución del variant calling.....	61
Tabla 20: Número de SNPs promedio reportados por cada herramienta.....	62
Tabla 21: Número de INDELs promedio reportados por cada herramienta.....	63
Tabla 22: Variantes confirmadas encontradas y posicionada como principal.....	64
Tabla 23: Número de variantes reportadas.....	66
Tabla 24: Estadísticas de precisión del variant calling.....	66
Tabla 25: VAF y profundidad de las variantes de cada paciente.....	67
Tabla 26: Configuraciones de parámetros para la etapa de trimming y filtrado.....	68
Tabla 27: Porcentaje promedio de bases y lecturas conservadas.....	69
Tabla 28: Porcentaje promedio de lecturas mapeadas.....	69
Tabla 29: Número de variantes reportadas en cada muestra.....	70
Tabla 30: Tiempo de ejecución de cada etapa del pipeline.....	71
Tabla 31: Relación entre archivos de entrada y etapas del pipeline.....	72
Tabla 32: Extracción de información de la Revisión Sistemática - Artículo 1.....	86
Tabla 33: Extracción de información de la Revisión Sistemática - Artículo 2.....	86
Tabla 34: Extracción de información de la Revisión Sistemática - Artículo 3.....	87
Tabla 35: Extracción de información de la Revisión Sistemática - Artículo 4.....	87
Tabla 36: Extracción de información de la Revisión Sistemática - Artículo 5.....	88
Tabla 37: Extracción de información de la Revisión Sistemática - Artículo 6.....	88

Tabla 79: Extracción de información de la Revisión Sistemática - Artículo 48.....	110
Tabla 80: Extracción de información de la Revisión Sistemática - Artículo 49.....	110
Tabla 81: Extracción de información de la Revisión Sistemática - Artículo 50.....	111
Tabla 82: Extracción de información de la Revisión Sistemática - Artículo 51.....	111
Tabla 83: Extracción de información de la Revisión Sistemática - Artículo 52.....	112
Tabla 84: Características de FastQC para control de calidad.....	112
Tabla 85: Características de Fastp para control de calidad.....	113
Tabla 86: Características de MinIONQC para control de calidad.....	113
Tabla 87: Características de NanoPlot para control de calidad.....	114
Tabla 88: Características de NanoQC para control de calidad.....	114
Tabla 89: Características de PycoQC para control de calidad.....	115
Tabla 90: Características de Seqkit para control de calidad.....	115
Tabla 91: Características de Bedtools para analizar la calidad del mapeo.....	116
Tabla 92: Características de DeepTools para analizar la calidad del mapeo.....	116
Tabla 93: Características de Qualimap para analizar la calidad del mapeo.....	116
Tabla 94: Características de Mosdepth para analizar la calidad del mapeo.....	117
Tabla 95: Características de NanoPlot para analizar la calidad del mapeo.....	117
Tabla 96: Características de Picard para analizar la calidad del mapeo.....	117
Tabla 97: Características de Samtools para analizar la calidad del mapeo.....	118
Tabla 98: Características de SeqKit para analizar la calidad del mapeo.....	118
Tabla 99: Características de BBduk para trimming y filtrado.....	118
Tabla 100: Características de Chopper para trimming y filtrado.....	119
Tabla 101: Características de Cutadapt para trimming y filtrado.....	119
Tabla 102: Características de DeepTools para trimming y filtrado.....	119
Tabla 103: Características de Fastp para trimming y filtrado.....	120
Tabla 104: Características de Filtlong para trimming y filtrado.....	120
Tabla 105: Características de Picard para trimming y filtrado.....	121
Tabla 106: Características de Porechop para trimming y filtrado.....	121
Tabla 107: Características de Samtools para trimming y filtrado.....	121
Tabla 108: Características de Seqkit para trimming y filtrado.....	121
Tabla 109: Características de Trimmomatic para trimming y filtrado.....	122
Tabla 110: Motivo de exclusión de herramientas de variant calling.....	122

ÍNDICE DE FIGURAS

FIGURA	PÁGINA
Figura 1: Estructura de la molécula de ADN.....	7
Figura 2: Estructura de referencia de un gen.....	8
Figura 3: Etapas de la expresión genética.....	9
Figura 4: Transcripción de ADN a ARNm.....	9
Figura 5: Tabla de equivalencia de codón a aminoácido.....	10
Figura 6: Proceso de replicación celular.....	11
Figura 7: Ejemplo de herencia, entre dos individuos heterocigotos.....	12
Figura 8: Ejemplo de SNP.....	13
Figura 9: Ejemplo de inserción y delección.....	14
Figura 10: Proteína policistina-1 localizada en la membrana celular.....	15
Figura 11: Diferencias en la supervivencia renal entre individuos con mutaciones...	15
Figura 12: Flujo de trabajo del proceso de análisis genético.....	16
Figura 13: Secuenciador Illumina MiSeq.....	23
Figura 14: Secuenciador Oxford Nanopore MinION.....	23
Figura 15: Secuencia de ADN, con adaptador y enzima fijada.....	24
Figura 16: Señal captada por el paso de la hebra de ADN por el nanoporo.....	24
Figura 17: Pipeline para el análisis bioinformático.....	25
Figura 18: Número de artículos seleccionados versus artículos aceptados.....	31
Figura 19: Número de publicaciones aceptadas por año.....	31
Figura 20: Gráfico del número de publicaciones por año, en el que se utiliza un.....	32
Figura 21: Gráfico del número de publicaciones por año, en el que se utiliza un.....	32
Figura 22: Mapa del gen PKD1 con la posición de los nueve partidores utilizados..	36
Figura 23: Proceso de preparación de librerías para secuenciación de ADN.....	37
Figura 24: Funcionamiento de duplex basecalling.....	38
Figura 25: Indicadores de Calidad de FastQC.....	41
Figura 26: Largo de las lecturas alineados versus largo de las lecturas de las.....	44
Figura 27: Ejemplo funcionamiento ventana deslizante desde 3' a 5'	46
Figura 28: Ejemplo de ambigüedad en el mapeo.....	48
Figura 29: Diagrama de flujo para la calificación de variantes.....	57
Figura 30: Relación entre el variant calling y la anotación de variantes.....	59
Figura 31: Pipeline para el análisis bioinformático.....	71

RESUMEN

La Poliquistosis Renal Autosómica Dominante (**ADPKD**) es la causa genética más frecuente de enfermedad renal en etapa terminal y responsable de una proporción significativa de los casos de enfermedad renal crónica (Mahboob, Rout & Bokhari, 2023). Esta enfermedad impacta negativamente en la calidad de vida de los pacientes y supone una carga económica considerable para los sistemas de salud y los afectados.

El diagnóstico temprano y accesible es fundamental para permitir el acceso a tratamientos oportunos, facilitar la planificación familiar y posibilitar el trasplante renal con donante vivo. La secuenciación genética es una herramienta clave para diagnosticar **ADPKD**. Sin embargo, la accesibilidad a este tipo de diagnóstico está limitada por los costos asociados con la secuenciación genética, así como la posibilidad de que la muestra sea procesada en un laboratorio con la capacidad específica.

En este trabajo se desarrolló un **pipeline bioinformático** para secuenciar el gen **PKD1** utilizando el secuenciador **Oxford Nanopore MinION**, con el objetivo de optimizar el diagnóstico de **ADPKD**. El **pipeline** fue validado con muestras de ocho pacientes que presentaban variantes conocidas, evaluando herramientas específicas en cada etapa del análisis.

El resultado fue un **pipeline** capaz de detectar y clasificar con precisión las variantes en las ocho muestras, completando el análisis en menos de 6 horas. El uso del secuenciador **MinION** permitió cubrir una región más amplia del gen **PKD1**, secuenciando completamente los 46 exones y 41/45 intrones. Además, el enfoque redujo los costos asociados y disminuyó el número de **PCRs** necesarias para la preparación de las muestras.

ABSTRACT

Autosomal Dominant Polycystic Kidney Disease (**ADPKD**) is the most common genetic cause of end-stage renal disease and accounts for a significant proportion of chronic kidney disease cases (Mahboob, Rout & Bokhari, 2023). This condition negatively impacts patients' quality of life and represents a considerable economic burden for healthcare systems and affected individuals.

Early and accessible diagnosis is essential to enable timely treatment, facilitate family planning, and allow for kidney transplantation from a living donor. Genetic sequencing is a key tool for diagnosing **ADPKD**. However, access to this type of diagnosis is limited by the costs associated with genetic sequencing and the need for samples to be processed in a laboratory with the required capabilities.

In this study, a **bioinformatics pipeline** was developed to sequence the **PKD1** gene using the **Oxford Nanopore MinION** sequencer, with the goal of optimizing the diagnosis of **ADPKD**. The **pipeline** was validated using samples from eight patients with known variants, evaluating specific tools at each stage of the analysis.

The result was a **pipeline** capable of accurately detecting and classifying variants in all eight samples, completing the analysis in less than 6 hours. The use of the **MinION** sequencer allowed for broader coverage of the **PKD1** gene, fully sequencing all 46 exons and 41/45 introns. Additionally, this approach reduced associated costs and minimized the number of **PCRs** required for sample preparation.

1. INTRODUCCIÓN

1.1 Contexto y oportunidad

En la última década, la genética ha cobrado cada vez más importancia en el área de la salud, donde su uso en el diagnóstico de enfermedades ha aumentado considerablemente, siendo empleada tanto para diagnosticar enfermedades de origen genético como para identificar virus y bacterias causantes de enfermedades. Este uso se ha extendido ampliamente gracias al incremento y mejora de las técnicas de secuenciación, así como a la reducción de los costos asociados, lo que ha permitido el acceso incluso a laboratorios de menor escala. Además, el aumento de los datos genéticos generados por pacientes y estudios, así como su almacenamiento en bases de datos públicas, ha contribuido a la identificación de variantes (alteraciones en el ADN) responsables de enfermedades.

En esta línea, desde hace aproximadamente seis años, un grupo de docentes de la Universidad Austral de Chile viene ejecutando la iniciativa **GEMINI** (GEnética Molecular Informática para Nefropatías) con el objetivo de mejorar la calidad de vida de los pacientes con sospecha de una enfermedad renal hereditaria y sus familias. En este contexto, a través de dos proyectos con financiamiento regional, **GEMINI** propone brindar acceso a técnicas y tecnologías modernas de medicina personalizada, al tiempo que reduce los costos del presupuesto sanitario nacional e identifica donantes para trasplantes renales (Realizan exitoso trasplante renal, 2024).

GEMINI busca alcanzar estos objetivos mediante el desarrollo, la validación y la implementación de un servicio de diagnóstico de enfermedades renales genéticas (Proyecto GEMINI, 2024). Esto ayuda a las personas a obtener confirmación de una enfermedad en ciertos casos, permitiendo a quienes estén afectados orientar su planificación familiar. Además, facilita el trasplante renal con donante vivo para pacientes en diálisis.

Dado este escenario, **GEMINI** ha centrado su atención en la enfermedad renal hereditaria *Poliquistosis Renal Autosómica Dominante* (**ADPKD**, por sus siglas en inglés), la cual es la causa hereditaria más común de enfermedad renal en el mundo (Mahboob, Rout & Bokhari, 2023). Además, los pacientes diagnosticados con **ADPKD** representan entre un 6-10% de los pacientes en diálisis en Estados Unidos (Mahboob, Rout & Bokhari, 2023) y un 6% de los casos de insuficiencia renal terminal en España (Torres, et al., 2006). Esto implica un costo significativo tanto para el estado como para la calidad de vida del paciente y su familia.

El proceso de diagnóstico llevado a cabo por el equipo de **GEMINI** implica una serie de pasos. Comienza con la extracción de una muestra de ADN de una persona diagnosticada clínicamente con **ADPKD**, el cual se denomina caso índice. Luego, se realiza la amplificación de una región específica del ADN mediante la técnica de **PCR** (*Polymerase Chain Reaction*). Posteriormente, las muestras son analizadas por un secuenciador. Los datos generados por el secuenciador son sometidos a un proceso de análisis que consta de varias etapas, cada una utilizando herramientas especializadas seleccionadas y calibradas en función del secuenciador y la región del

ADN analizada. Finalmente, este proceso de análisis proporciona una lista de variantes genéticas que son candidatas a ser responsables de la enfermedad.

Con este flujo de trabajo, denominado *pipeline*, y con el objetivo de crear un servicio de diagnóstico accesible para quienes lo necesiten, es que en trabajos anteriores se avanzó en diferentes formas de optimizar los tiempos y el esfuerzo de este proceso. En una primera etapa, mediante la automatización del diseño de partidores para la amplificación de las regiones de interés en el ADN, se logró reducir el tiempo de creación de partidores para **PCR**, validando los resultados con el diseño y validación in vitro de partidores para amplificar los exones del gen **PKD2** (Klenner, 2018). Con el éxito del trabajo anterior, se creó un nuevo proyecto llamado **GEMINI-2**, donde se buscó reducir aún más los tiempos en la etapa de **PCR** modificando el *software* anterior para ahora diseñar partidores multiplexables (Castro, 2022), lo que significa que se pueden realizar varias **PCRs** en un solo tubo.

Al seguir avanzando en torno al objetivo, el equipo ha puesto especial atención en los procesos de secuenciación y análisis. En el caso de la secuenciación, actualmente es derivado a un organismo externo, lo que además de significar un costo, aumenta los tiempos de gestión necesarios para coordinar la contratación de un especialista cada vez que sea necesario. Aun así, la ventaja que tiene externalizar el servicio es que se puede contar con un servicio de alta calidad, permitiendo secuenciar las muestras en un secuenciador **Illumina MiSeq**, el cual es un secuenciador denominado de próxima generación (**NGS**), el cual tiene un alto costo (superior a los 120 millones de pesos), y que es ampliamente utilizado para este propósito, ya que genera lecturas de alta calidad. Esta tecnología permite secuenciar lecturas de muestras de varios pacientes al mismo tiempo manteniendo la calidad.

Por otro lado, el secuenciador **Oxford Nanopore MinION** de tercera generación (**TGS**) comercializado desde 2014, se caracteriza por ser un secuenciador portable, de muy bajo costo (alrededor de 1.000 dólares) y que genera lecturas del largo de miles de pares de bases (pbs), pero con una tasa de error mucho mayor a los NGS. No obstante lo anterior, en los últimos años se han presentado muchos avances en su tecnologías, tanto en la química de los materiales utilizados para secuenciar las muestras, como en su *software* de lecturas y corrección de errores que han hecho que la calidad de los datos generados aumente cada año.

Las características del secuenciador **MinION**, junto con su bajo costo, han llevado al equipo a considerar la posibilidad de adquirir este dispositivo y agregarlo a su arsenal de herramientas de diagnóstico. Esto, a su vez, requeriría la creación de un nuevo flujo de trabajo para el análisis bioinformático de las secuencias generadas por este secuenciador. Esta iniciativa abre la oportunidad de establecer un proceso completamente realizable por los miembros del equipo **GEMINI**, desde la toma de muestras hasta el diagnóstico final de la enfermedad.

1.2 Estado del arte

Cuando se habla de diagnóstico de enfermedades genéticas, los secuenciadores por excelencia de los que se tienden a hacer uso, son los secuenciadores **NGS** de la plataforma **Illumina**, dada la gran calidad de los datos que generan, al igual que la

posibilidad de secuenciar múltiples muestras de pacientes al mismo tiempo, reduciendo así los costos y tiempos de secuenciación de las opciones anteriores.

Aun así en los últimos años, los secuenciadores de tercera generación (**TGS**) han ido haciéndose cada vez más presentes en el diagnóstico de enfermedades genéticas, esto debido a que estos nuevos secuenciadores como los de tecnologías de nanoporos, tienen ventajas sobre los secuenciadores **NGS**, como la posibilidad de obtener lecturas de miles de pares de bases (letras del ADN), lo que por ejemplo permite detectar con mayor facilidad tipos de variantes que abarcan varios pares de bases, como lo son las variantes estructurales (**SV**) a través de varios casos donde se ha utilizado esta tecnología (Xiao y Zhou (2020)).

Sumado a lo anterior, al generar lecturas largas, también son menos propensos a secuenciar pseudo genes (regiones similares en el ADN), como es el caso del gen **PKD1** (Xiao & Zhou, 2020), que es el gen más relevante en el diagnóstico de **ADPKD**.

En un reciente estudio, se utilizó un secuenciador con la tecnología de nanoporos para el diagnóstico de 20 individuos con los síndromes de Prader-Willi o Angelman. Paschal et al. (2024) resaltan la importancia de que gracias a las secuenciación de lecturas largas (**LRS**) se pueden analizar diferentes tipos de variantes genéticas a partir de una única fuente de datos, esto trae como beneficio una tasa más alta de diagnóstico y tiempos de respuesta más cortos, además de menores costos.

Por otro lado los secuenciadores de nanoporos también han sido utilizados para la detección de mutaciones en una sola variante (**SNP**), las inserciones y delecciones (**INDEL**). Por ejemplo Nakamura et al. (2024) probaron esta tecnología en el análisis de 33 genomas recopilados de pacientes con sospecha de cáncer hereditario, y comparando los resultados con los obtenidos por tecnologías como **Illumina** o **Sanger** (considerado como *gold standard*) obteniendo resultados bastante similares en la detección de variantes de una letra, incluyendo aquellas clasificadas como patogénicas.

En relación a **ADPKD** y el uso del secuenciador **MinION**, Durkie et al. (2023) secuenciaron los exones del 15 al 33 del gen **PKD1**, con el fin de obtener el haplotipo (Combinaciones de **SNPs** en un mismo cromosoma) de las variantes previamente identificadas, logrando exitosamente confirmar que para uno de los casos que se estaba estudiando, cada parente aportó con una variante.

Finalmente, Helal et al. (2022) destacan la importancia de la identificación precisa y confiable de las variantes del genoma de cada persona, poniendo a prueba seis herramientas de **variant calling**, utilizando los conjuntos de muestras genéticas de referencia **NA12878** y **NA24385**, comúnmente utilizados en la evaluación de herramientas bioinformáticas. Este estudio permite tomar una decisión más objetiva al momento de elegir una herramienta, en este caso evaluadas en genes relacionados con el cáncer de mama. Los resultados de este estudio muestran que tanto la herramienta **Clair3** como **Human-SNP-wf** (que utiliza Clair3) obtuvieron los mejores resultados en cuanto a precisión, recall y f1-score. Además, se puede observar que, en general, las herramientas de detección de variantes tienen un

desempeño más bajo al detectar variantes de inserciones o delecciones (**INDEL**), en comparación con las variaciones de un solo nucleótido (**SNP**).

1.3 Innovación propuesta y su impacto

1.3.1 Propuesta

Se propone el desarrollo de un ***pipeline*** para análisis bioinformático, para secuencias del gen **PKD1**, generadas por el secuenciador **Oxford Nanopore MinION**, para el diagnóstico de **ADPKD**, que representa la enfermedad renal hereditaria más frecuente del mundo y la principal causa genética de ingreso a diálisis.

Con el fin de sobrellevar de mejor manera la baja calidad de las lecturas generadas por el secuenciador (en comparación con las obtenidas con tecnología **NGS**), se llevará a cabo una revisión sistemática de la literatura, a fin de explorar las distintas herramientas disponibles para cada uno de los pasos del análisis, comparándolas en base a ciertas características y eligiendo aquellas que entreguen los mejores resultados para los casos de validación con los que se cuentan.

El ***pipeline*** diseñado podrá ser utilizado por cualquier miembro del equipo **GEMINI** mediante la ejecución de un *script* que, a través de parámetros por línea de comando, permite la ejecución de forma fácil y automática de cada etapa del ***pipeline***, requiriendo así un conocimiento mínimo en bash. Sumado a lo anterior, los miembros del equipo contarán con un manual, el cual contendrá la información necesaria para la correcta ejecución y el correcto entendimiento de cada etapa del ***pipeline***.

Finalmente el ***pipeline*** entregará la lista de variantes encontradas, el tipo de variante (**SNP**, **INDEL**), la posición, y la puntuación según un sistema de *ranking*, para clasificar la prioridad de la variante.

1.3.2 Impacto social-económico

La incorporación del secuenciador **MinION**, sumado a la creación de un ***pipeline*** para el proceso de análisis de **PKD1**, trae consigo múltiples impactos sociales y económicos, que van principalmente impulsados por la reducción de los tiempos y costos que se lograría con esta propuesta. Partiendo desde el proceso de **PCR**, el número de **PCRs** que se realizan por cada paciente para amplificar **PKD1**, se reduciría de 70 a nueve. Esto repercutе en los costos, en el tiempo que le toma al equipo realizar este proceso y además disminuye la probabilidad de cometer errores.

A continuación, en la etapa de secuenciación, al dejar de depender de la contratación de un servicio externo, se reducirán los costos de contratación, además de los tiempos de coordinar la contratación del servicio, como se mencionó anteriormente. También se ha comprobado que el secuenciador **MinION** secuencia las muestras más rápido que el secuenciador **MiSeq** de Illumina, lo que reduciría aún más los tiempos de esta etapa.

Los materiales que ocupa el secuenciador **MinION** para leer las muestras también representan un costo menor que los requeridos para **MiSeq**. A modo de ejemplo, para la secuenciación de un total de 20 muestras con **MiSeq**, el valor aproximado es de 5-6 millones de pesos, versus aproximadamente 1-2 millones de pesos para secuenciar 12 muestras con **MinION**, lo que significa que secuenciar una muestra con MinION es aproximadamente dos veces más barato que con **MiSeq**.

Por último, durante la etapa de análisis también se logra una reducción del tiempo, al contar con un *script* que automatiza todo el proceso, lo que le permite a cualquier miembro del equipo poder llevar a cabo esta etapa.

Finalmente, todas estas reducciones en el costo y los tiempos en diferentes etapas del proceso repercute directamente en el costo y tiempos totales del servicio, lo que se traduce en una mayor accesibilidad para las personas que requieran este análisis.

1.3.3 Impacto científico-tecnológico

Los impactos científico-tecnológicos derivados del desarrollo de este proyecto son impulsados principalmente por la adopción del secuenciador **Oxford Nanopore MinION**, en la realización de los análisis genéticos. Con la utilización de este dispositivo, se abre la posibilidad de descubrir variantes en regiones previamente no analizadas, gracias a sus características como la generación de lecturas largas (*long-read*). Además, posibilita la creación de **pipelines** de análisis que exploran diversos tipos de variantes, abarcando decenas o centenas de pares de bases, facilitando la identificación de variantes estructurales (SV) o variaciones del número de copias (CNV).

1.4 Objetivos

1.4.1 Objetivo general

Mejorar las capacidades de detección y diagnóstico de **ADPKD** a partir de la secuenciación de **PKD1**, mediante un manual que permita realizar el proceso de análisis de la secuencia generadas por el secuenciador MinION.

1.4.2 Objetivos específicos

A partir del objetivo general presentado, se despliega una serie de objetivos específicos, que se centran en aspectos particulares del desarrollo y contribuyen al logro del objetivo general:

1. Manejar el funcionamiento y características principales del secuenciador **Oxford Nanopore MinION** como: funcionamiento, formatos de archivos de salidas, requerimientos de *hardware*, *software*, etc. en un plazo de cuatro semanas.
2. Obtener al menos tres herramientas necesarias para cada una de las distintas etapas del análisis, como con el control de calidad, *trimming*, mapeo de la secuencia, *variant calling*, consulta a bases de datos de variantes y la predicción de patogenicidad, en un plazo de 12 semanas.

3. Establecer un ***pipeline*** de análisis bioinformático, para el diagnóstico de **ADPKD**, mediante el análisis de **PKD1** generadas con el dispositivo **MinION** y con el uso de herramientas halladas en (2) y validado con muestras de pacientes con variantes conocidas y desconocidas, en un plazo de 12 semanas.
4. Permitir que los miembros del equipo puedan aplicar el ***pipeline***, mediante un manual, que facilite la aplicación de este protocolo de análisis post-secuenciación para ser aplicado en futuros análisis de secuenciaciones, en un plazo de cuatro semanas, luego de la validación del funcionamiento del protocolo.

2. MARCO TEÓRICO

2.1 Genética molecular

La genética molecular es el campo de la genética, dedicada a estudiar el **ADN**, así como su estructura, sus genes y funcionamiento. El comienzo de este campo se podría ubicar junto con el descubrimiento del **ADN**, el cual fue aislado por primera vez en 1869 por el biólogo suizo Friedrich Miescher.

2.1.1 ADN

El **ADN**, es una molécula que se encuentra en el núcleo de las células, la cual está conformada por la unión de millones de nucleótidos. A su vez, cada nucleótido está conformado por una azúcar (desoxirribosa), un grupo fosfato y una base nitrogenada [adenina (A), guanina (G), citosina (C) y timina (T)] (Kassem, Girolami & Sanoudou, 2012). El **ADN** presente en el núcleo de las células adopta una estructura de doble hélice (ver Figura 1), la cual se forma a partir de la unión de dos hebras complementarias. En otras palabras, cada base nitrogenada de una hebra “1” se une específicamente con su complementaria en la hebra “2”, siguiendo el criterio de complementariedad de bases (A-T, G-C).

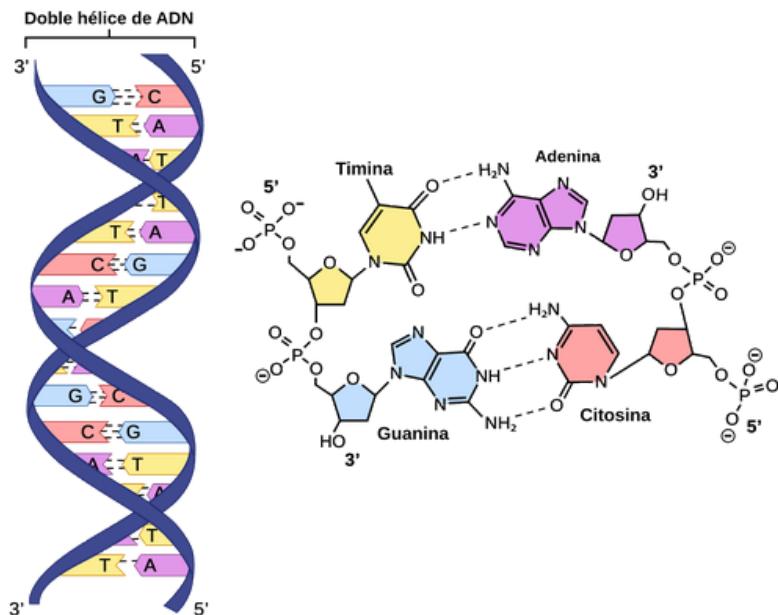


Figura 1: Estructura de la molécula de ADN¹.

El **ADN** humano posee alrededor de $3 \cdot 10^9$ millones de nucleótidos (Kassem et al., 2012), distribuidos en 23 pares de cromosomas. El conjunto total de nucleótidos en un organismo, se conoce como genoma. El genoma humano contiene alrededor de 20.000 genes (Ars, 2021), que son regiones de **ADN** responsables de codificar proteínas. Sin embargo, se estima que solo el 2% del **ADN** total codifica proteínas.

¹ <https://theory.labster.com/es/dna-structure/>

Los genes están compuestos por cuatro tipos de regiones (ver Figura 2): la **región reguladora**, que determina cuándo y en qué cantidad se transcribe un gen específico, la **región promotora**, que es donde se une la maquinaria encargada de la transcripción del gen, además de señalar el inicio de este proceso. Los **exones** son las regiones del gen que codifican proteínas, y el conjunto de todos los exones se conoce como "**exoma**". Por último, los **intrones** son las regiones que se encuentran entre los exones y no tienen función codificante. Estas regiones suelen representar la mayor parte del gen (Kassem et al., 2012).

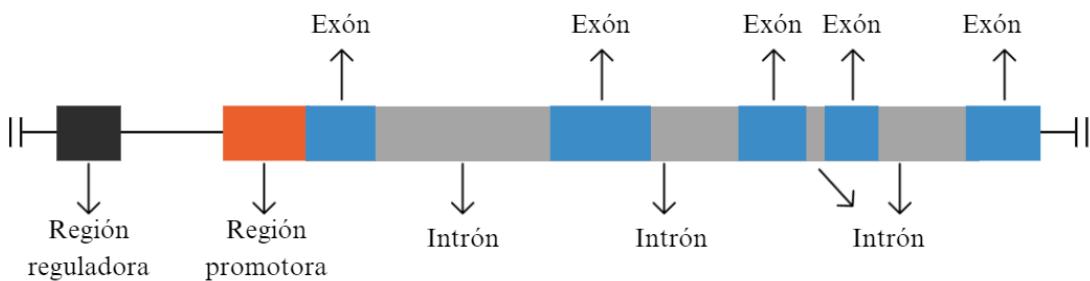


Figura 2: Estructura de referencia de un gen.

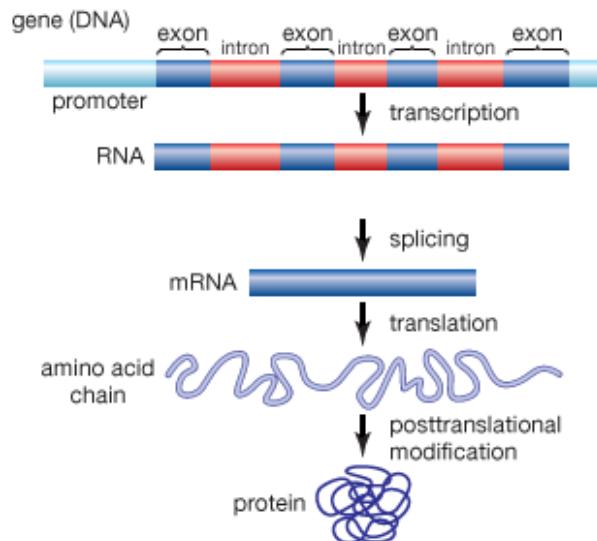
El proceso por el cual los genes codifican proteínas se denomina “**expresión genética**” y constituye lo que se conoce como el **dogma central de la biología molecular**.

2.1.2 Dogma central de la biología molecular

El dogma central de la biología molecular, definido por Francis Crick en 1958 y reformulado en un artículo publicado en 1970, establece que existen tres clases importantes de biopolímeros: el **ADN**, **ARN** y las **proteínas** (Kassem et al., 2012). Este dogma describe el proceso mediante el cual el **ADN** se replica, así como los procesos mediante los cuales se puede realizar la transición de un biopolímero a otro, como la **expresión génica** (Figura 3), que traduce las regiones codificantes de un gen a proteínas, pasando por una serie de etapas hasta formar una cadena de aminoácidos.

2.1.2.1 Transcripción

Para llevar a cabo el proceso de **expresión genética**, primero se debe pasar por la etapa de **transcripción**, la cual inicia cuando una enzima llamada **ARN polimerasa**, reconoce una región promotora del gen (Figura 2). Una vez el **ARN polimerasa** reconoce esta zona, se une a ésta, lo que produce que las hebras de **ADN** se separen a lo largo de una región acotada (Cornejo, 2022). Como se puede observar en la Figura 4, ésto permite que la **ARN polimerasa** genere una nueva cadena complementaria de nucleótidos en sentido 5' a 3', pero donde la base nitrogenada timina se reemplaza por uracilo. Esta nueva cadena se denomina **ARN mensajero (ARNm)**. Una vez se ha terminado de copiar toda la información de todo el gen, la **ARN polimerasa** libera la cadena de **ARNm**, permitiendo que la doble hebra de **ADN**, vuelvan a unirse.



© 2008 Encyclopædia Britannica, Inc.

Figura 3: Etapas de la expresión genética².

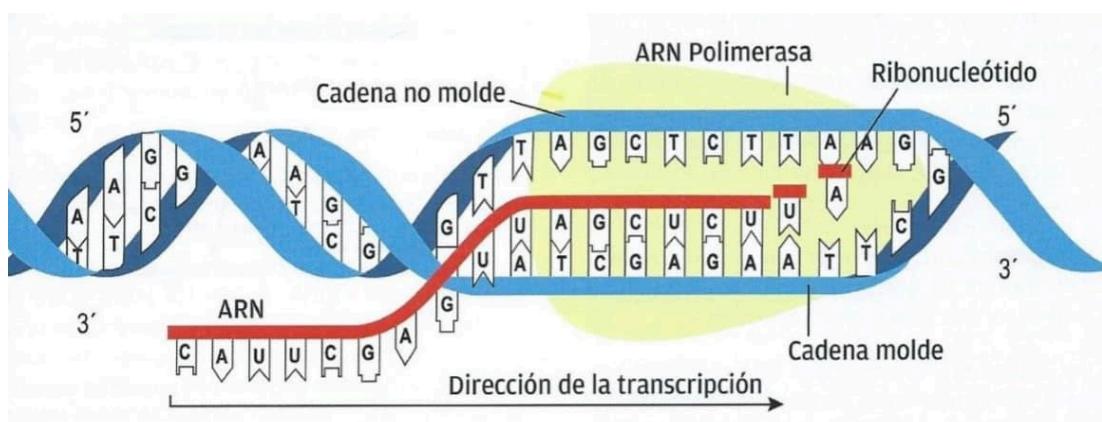


Figura 4: Transcripción de ADN a ARNm³.

Después de que se ha creado esta molécula de **ARNm**, esta pasa por un proceso denominado “**splicing**” donde se remueven los **intrones**, dejando así sólo los **exones** que son las zonas que finalmente codifican la **proteína**. Luego de este paso el **ARNm** sale del núcleo de la célula, hacia el citoplasma, para comenzar el proceso de **traducción**.

2.1.2.2 Traducción

Una vez que el **ARNm** se encuentra en el citoplasma, esta molécula se une a un ribosoma, el cual es el organelo celular encargado de sintetizar una cadena de aminoácidos (**proteína**). El ribosoma comienza a leer la cadena de **ARNm** en conjuntos de tres nucleótidos, denominados **codones**. Cada combinación de codones equivale a uno de los 20 aminoácidos con los que los ribosomas sintetizan las proteínas (ver Figura 5). Cada vez que el ribosoma lee un codón, añade el aminoácido equivalente a la cadena. Sin embargo, también existen codones que

² <https://kids.britannica.com/kids/assembly/view/114928>

³ <https://microbacterium.es/como-funciona-la-transcripcion-del-adn>

marcan el inicio y término de la traducción. Por lo tanto, el proceso termina cuando el ribosoma lee alguno de estos codones de *stop*. En ese momento, el ribosoma deja de agregar aminoácidos a la cadena y la libera como una proteína madura (Kassem et al., 2012).

		Segunda Letra				
		U	C	A	G	
Primera Letra	U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA STOP UAG STOP	UGU Cys UGC Cys UGA STOP UGG Trp	U C A G
	C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
	A	AUU Iso AUC Iso AUA Iso AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
	G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

©BIOINNOVA
innovabiologia.com

Figura 5: Tabla de equivalencia de codón a aminoácido⁴.

2.1.2.3 Replicación

La replicación del **ADN** es el proceso mediante el cual se duplica una molécula de **ADN**. Esto ocurre en conjunto con la división celular, durante la cual la célula genera una copia prácticamente idéntica de sí misma, es decir, una copia para cada uno de los 23 pares de cromosomas. La replicación se inicia cuando un grupo de proteínas especializadas reconoce un punto en el genoma, llamado origen de replicación (Figura 6). Estas proteínas abren el **ADN**, formando dos estructuras en forma de "Y" conocidas como **horquillas de replicación**, lo que da lugar a una **burbuja de replicación**. Este proceso ocurre simultáneamente en todo el cromosoma (Mecanismos Moleculares de la Replicación del ADN | Khan Academy, 2024.). Una vez que se forma la burbuja de replicación, unas enzimas llamadas **ADN polimerasa** se posicionan en cada hebra del **ADN**, en el extremo 3', y comienzan a sintetizar un nuevo par de hebras de **ADN**, finalizando con una copia casi exacta de cada cromosoma.

⁴ <https://www.innovabiologia.com/biodiversidad/diversidad-animal/el-codigo-genetico/>

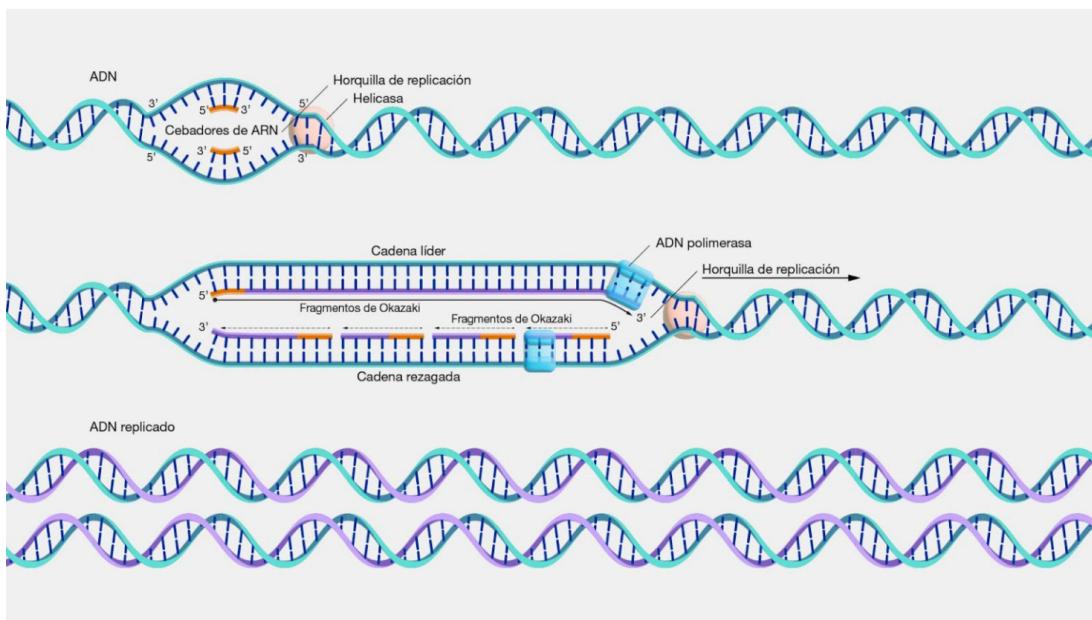


Figura 6: Proceso de replicación celular⁵.

2.1.3 Herencia

Antes del descubrimiento del **ADN**, Gregor Mendel realizó importantes investigaciones en el campo de la genética con sus estudios sobre guisantes, publicados en 1866. Mendel demostró empíricamente la existencia de patrones hereditarios en ciertas características, definiendo la presencia de características **dominantes** y **recesivas**. Además, introdujo el concepto de **alelos** para describir las diferentes variantes de una característica. Cada individuo hereda dos alelos, uno de cada progenitor, para cada característica, por lo tanto, el individuo puede tener dos alelos idénticos (**homocigotos**) o diferentes (**heterocigotos**).

Mediante simples análisis estadísticos y con un conocimiento previo sobre los alelos y sus características, así como sabiendo que cada individuo tiene igual probabilidad de heredar cualquiera de los dos alelos de sus progenitores, es posible calcular la probabilidad de heredar ciertas combinaciones de alelos. Como se muestra en la Figura 7, en el ejemplo existen dos posibles alelos que determinan el color de los guisantes (amarillo o verde). El alelo "amarillo" es dominante y por ende se representa con una letra mayúscula, mientras que el alelo verde es recesivo y se representa con una letra minúscula, lo que significa que se necesita tener dos alelos "verdes" para expresar la característica.

⁵ <https://www.genome.gov/es/genetics-glossary/Replicacion-de-ADN>

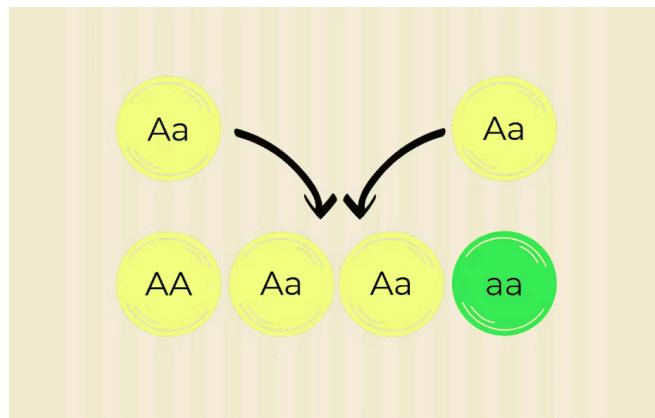


Figura 7: Ejemplo de herencia, entre dos individuos heterocigotos⁶.

2.2 Enfermedades renales genéticas

Cuando se habla de enfermedades genéticas, se hace referencia a enfermedades que tienen su origen en el **ADN**, debido a modificaciones en este. Estas enfermedades pueden ser heredadas (**variante germinal**), como la **ADPKD**, o puede surgir de forma espontánea en alguna célula (**variante somática**), como la mayoría de los cánceres.

En el contexto de las enfermedades renales hereditarias, generalmente se habla también de enfermedades renales raras, definidas como aquellas con una prevalencia menor a un caso cada 2000 personas (según lo definido por la Unión Europea). Aproximadamente el 80% de las enfermedades raras tienen origen genético (Ars, 2021). Estas enfermedades mayoritariamente se diagnostican en la edad pediátrica y son responsables del 35% de muertes en el primer año de vida (Ars, 2021). Además las enfermedades renales raras representan entre el 5% y el 10% de las personas con enfermedad renal crónica, pero **constituyen más del 25% de los pacientes que reciben terapia de reemplazo renal** (Wong et al., 2024).

Las enfermedades renales se pueden clasificar en cinco etapas determinadas por la filtración glomerular estimada (**eGFR**, por sus siglas en inglés), la cual mide qué tan bien funcionan los riñones para filtrar los desechos y el exceso de líquido de la sangre. Cuando un paciente se encuentra en la etapa tres, se considera que tiene una enfermedad crónica moderada, mientras que si se encuentra en la etapa cinco, se considera avanzada. Además, en estas tres últimas etapas es donde se encuentran las tasas más altas de falla renal a los cinco años en pacientes con enfermedades renales genéticas en comparación con aquellos que no tienen un origen genético (Wong et al., 2024).

Sumado a lo anterior, en un estudio realizado por Wong et al (2024) con pacientes del Reino Unido, se observó que los pacientes con enfermedades renales raras tienden a tener una progresión más rápida de la enfermedad. Sin embargo también se observó que la mayoría de los pacientes con alguna enfermedad rara presentan insuficiencia renal alrededor de los 65 años.

⁶ <https://medicoplus.com/ciencia/leyes-mendel>

2.2.1 Tipos de variantes

Las variantes genéticas hacen referencia a las **alteraciones o diferencias que existen entre secuencias de ADN**. En el contexto del diagnóstico de enfermedades, estas diferencias se identifican mediante la comparación con un genoma de referencia (secuencia total de ADN de un organismo que se considera sano) y las secuencias leídas a partir de las muestras del paciente. Existen diferentes tipos de variantes genéticas dependiendo del tipo de alteración que presente. A continuación se presentan algunos de los tipos de variantes más comunes:

SNP (Single Nucleotide Polymorphism): Es el tipo más común de variación genética entre las personas, ocurre aproximadamente en 1 de cada 1000 nucleótidos (*What are single nucleotide polymorphisms (SNPs)? | MedlinePlus Genetics, 2024*) y consiste en un cambio de letra en alguna posición del **ADN** de la persona (ver Figura 8).

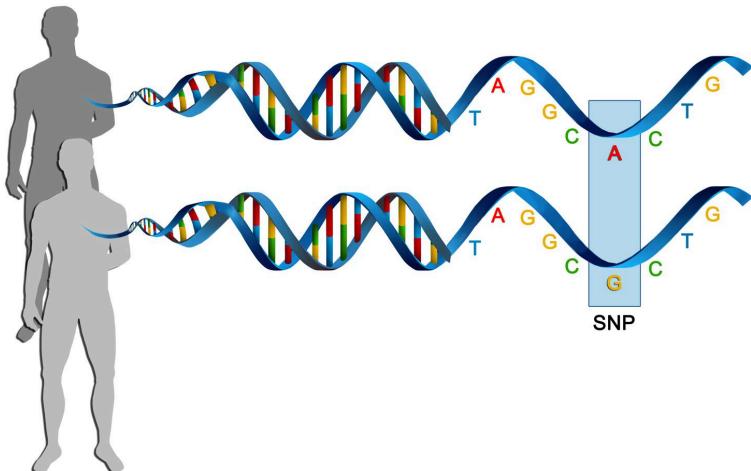


Figura 8: Ejemplo de SNP.⁷

INDEL (Insertion or deletion): Los **INDELS** son un tipo de variante genética que consiste en inserciones o eliminaciones de nucleótidos en el **ADN** (ver Figura 9).

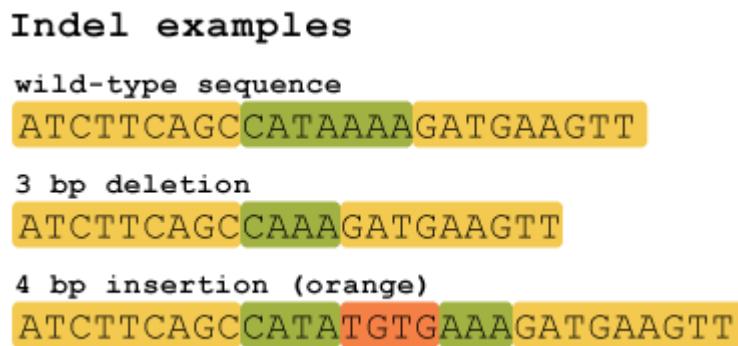


Figura 9: Ejemplo de inserción y delección.⁸

⁷

<https://atlasofscience.org/single-nucleotide-polymorphisms-as-genomic-markers-for-high-throughput-pharmacogenomic-studies/>

⁸ <https://hackbrightacademy.com/blog/indel-finder-how-the-python-version-of-this-program-works/>

Existen más tipos de variantes, como por ejemplo las variantes estructurales (**SV**), que son variaciones de grandes zonas del ADN que pueden ser inserciones, delecciones, polimorfismos u otros. También existen las variaciones del número de copias (**CNV**) que se producen cuando se producen una cantidad superior de copias de una región en el **ADN**. Dada la cantidad de nucleótidos que se ven implicados en estos tipos de mutaciones, es que no son tan comunes como los **SNPs** o **INDELS**.

Por último, la existencia de **una variante en el ADN no asegura que esta cause una enfermedad**. Para poder relacionar una variante con la expresión de una enfermedad, se requiere un proceso de análisis más profundo. Este incluye la búsqueda de información en bases de datos públicas y/o locales que categoricen la variante como causante de la enfermedad. Por otro lado, se pueden evaluar los posibles efectos que podría causar la variante durante la expresión génica. Finalmente, el análisis se puede extender a la familia, analizando a las personas sanas y afectadas, con el fin de verificar que las variantes solo se encuentran en los familiares afectados.

2.2.2 **PKD1** y Poliquistosis Renal Autosómica Dominante

PKD1 es un gen ubicado en el cromosoma 16, contiene 46 exones y tiene un largo de 47,191 nucleótidos. Este gen presenta seis pseudogenes (estructuras en el genoma que se asemejan al gen), lo cual hace que para analizarlo se tenga que realizar previamente **PCRs** adicionales, llamadas **long-range PCRs**, que permiten aislar el gen.

Sobre la función biológica de **PKD1**, se conoce que codifica una proteína llamada **poliquistina-1**, la cual se localiza en la membrana celular de las células renales. Una porción de esta proteína atraviesa la membrana celular y se extiende hacia el exterior de la célula (Figura 10), lo que le permite interactuar con proteínas, carbohidratos y lípidos en el entorno extracelular. Esta interacción facilita que la célula recibe señales del entorno que la rodea y responda de manera apropiada (*PKD1 gene | MedlinePlus Genetics*, 2024).

PKD1 está estrechamente relacionado con la enfermedad hereditaria **ADPKD**, en el que se le pueden atribuir cerca del 85% de los casos (Mahboob et al., 2023). Sumado a lo anterior, se ha identificado que los pacientes diagnosticados con **ADPKD** por variantes en **PKD1** tienen **una progresión más severa de la enfermedad**, frente a los pacientes diagnosticados con variantes en otros genes como **PKD2**, como se puede apreciar en la Figura 11. Se ha reportado que 58 años es la media de edad en que los pacientes con mutaciones en **PKD1** presentan insuficiencia renal, en contraste con la media de 79 años de los pacientes con mutaciones en **PKD2** (Benz & Hartung, 2021).

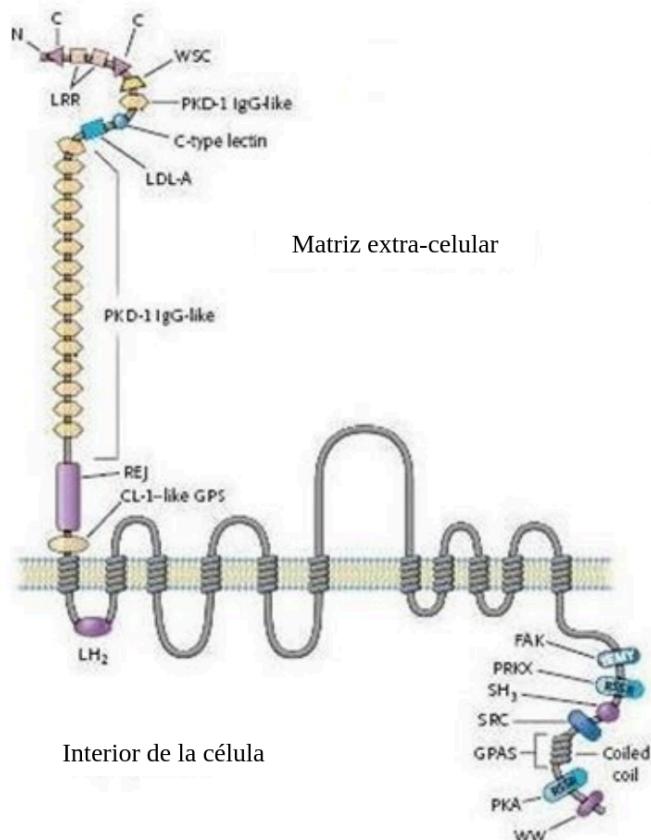


Figura 10: Proteína policistina-1 localizada en la membrana celular⁹.

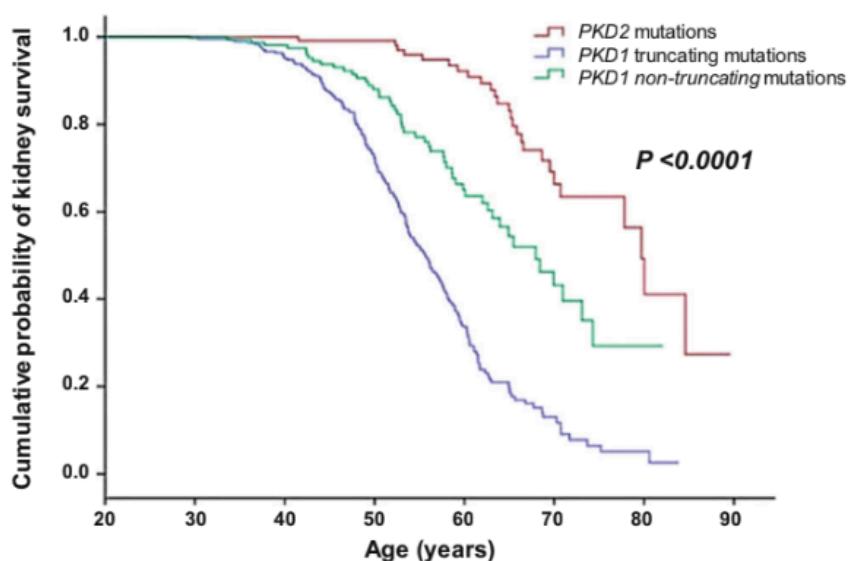


Figura 11: Diferencias en la supervivencia renal entre individuos con mutaciones truncantes de *PKD1*, mutaciones no truncantes de *PKD1* y mutaciones de *PKD2*.¹⁰

⁹

<https://clinical-experimental-nephrology.imedpub.com/articles/genetic-mutations-in-autosomal-dominant-polycystic-kidney-disease-type1.php?aid=24122>

¹⁰ <https://doi.org/10.1007/s00467-020-04869-w>

2.3 Diagnóstico de enfermedades mediante análisis genético

El diagnóstico de enfermedades hereditarias mediante análisis genético requiere una serie de procedimientos y herramientas para obtener la posible confirmación de la enfermedad que se está estudiando. A continuación, se detalla en qué consiste el proceso de análisis, haciendo especial énfasis en la última etapa antes de la entrega de resultados.

2.3.1 Análisis genético

El análisis genético es el proceso por el cual se busca confirmar una enfermedad genética, analizando aquellos genes que causan la enfermedad que se está estudiando. Para esto se sigue un flujo de trabajo (Figura 12), el cual puede variar en ciertos pasos dependiendo de ciertas decisiones que se tomen o el secuenciador que se esté utilizando. De manera global, el flujo de trabajo comienza con la **toma de muestras** del paciente, con el objetivo de extraer el **ADN**, el cual suele obtenerse de una muestra de sangre del paciente.



Figura 12: Flujo de trabajo del proceso de análisis genético.

El segundo paso se denomina **preparación de librerías**, que consiste en preparar los fragmentos de **ADN** para su secuenciación. Para ello, se inicia amplificando la región de interés mediante **PCR**. El número de **PCRs** necesarias para amplificar un gen dependerá del tamaño de los fragmentos de **ADN** que pueda manejar el secuenciador, así como de las características específicas del gen, como su contenido de guanina-citosina o la presencia de pseudogenes. En el caso de querer secuenciar todo el gen o sólo los exones, esto variará. Por lo general, se opta por secuenciar sólo los exones en análisis para el diagnóstico de enfermedades, ya que es en estas regiones donde se encuentran la mayoría de las variantes de importancia clínica.

Después del proceso de **PCR**, si se van a secuenciar conjuntamente muestras de distintas regiones o distintos organismos, se lleva a cabo un proceso similar a la **PCR**. En este proceso se añaden fragmentos de **ADN** llamados **barcoding**, los cuales permiten identificar cada muestra en análisis posteriores. Finalmente, se preparan las muestras conforme a las especificaciones del fabricante del secuenciador. Por lo

general, esto implica la incorporación de una serie de reactivos y la adición de adaptadores, que son fragmentos de **ADN** que se unen a las muestras y facilitan su unión a una celda.

El tercer paso consiste en el proceso de **secuenciación**, durante el cual las muestras se depositan en una *flowcell*. Esta es una celda **receptora de fragmentos de ADN**, y sus características y funcionamiento están directamente relacionados con el tipo de secuenciador utilizado. Posteriormente, el secuenciador comienza a leer los fragmentos de **ADN**, obteniendo la cadena de nucleótidos que los componen y generando finalmente un archivo con la información de las lecturas y su calidad.

Por último, las lecturas generadas por el secuenciador pasan por la etapa de análisis, donde, mediante un nuevo flujo de trabajo, se busca procesar estas lecturas con el objetivo de analizar las distintas variantes encontradas. Esto permite clasificar las variantes en cinco categorías descritas por el *American College of Medical Genetics (ACMG)* y la *Association for Molecular Pathology (AMP)* (Richards et al., 2015) (ver sección a continuación). Las variantes clasificadas en las dos últimas categorías, “**patogénica**” y “**probablemente patogénica**”, se consideran causantes de la enfermedad. Sin embargo, las variantes dentro de la categoría intermedia, es decir, **VUS** (*Variants of Unknown Significance*), requieren un estudio más completo, que generalmente involucra a los familiares que puedan poseer o no la enfermedad, con el fin de confirmar o descartar la patogenicidad de esta variante.

2.3.2 Interpretación de variantes genéticas

En 2015, tras dos años de trabajo, se publicó un artículo que propone estándares y una guía para la interpretación de variantes genéticas. Esta guía fue desarrollada mediante la opinión de expertos, el consenso del grupo de trabajo y la aportación de la comunidad (Richards et al., 2015). Este artículo clasifica las variantes genéticas en cinco categorías, las cuales son:

1. **Patogénica**: causante de enfermedad, con valor diagnóstico para la enfermedad asociada.
2. **Probablemente patogénica**: probablemente causante de enfermedad, a la práctica con valor diagnóstico para la enfermedad asociada.
3. **Variante de significado clínico incierto (VUS)**: con causalidad no definida con la información disponible actualmente.
4. **Probablemente benigna**: probablemente no causante de enfermedad.
5. **Benigna**: no causante de enfermedad.

La guía presenta 16 criterios que sirven como evidencia para justificar la patogenicidad de una variante, los cuales están detallados en la Tabla 1. Estos criterios varían en su impacto, siendo clasificados como "**de apoyo**" (*Supporting*), "**moderado**" (*Moderate*), "**fuerte**" (*Strong*) y "**muy fuerte**" (*Very Strong*). Por otro lado, también presenta 12 criterios detallados en la Tabla 2 para clasificar una variante como benigna, distribuidos en tres categorías: "**de apoyo**" (*Supporting*), "**fuerte**" (*Strong*) e "**independiente**" (*Stand-Alone*).

Tabla 1: Criterios para clasificar una variante como patogénica.

Peso	Código	Criterion
<i>Very strong</i>	PVS1	Variante nula (<i>nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion</i>) en un gen donde la pérdida de funcionalidad es un mecanismo conocido de la enfermedad.
<i>Strong</i>	PS1	La variante produce el mismo cambio en el aminoácido que previamente se había registrado como patogénica.
	PS2	Variante <i>de novo</i> ¹¹ (Confirmado en ambos progenitores) en un paciente con la enfermedad pero sin historial familiar.
	PS3	Estudios <i>in vitro</i> o <i>in vivo</i> que respaldan los efectos dañinos de la variante.
	PS4	La variante tiene una alta prevalencia en individuos afectados con la enfermedad.
<i>Moderate</i>	PM1	La variante se ubica en un punto crítico para la función de la proteína, y las variantes <i>missense</i> ¹² identificadas han resultado ser patogénicas.
	PM2	Ausente en los controles (o a una frecuencia extremadamente baja si es recesiva) en el <i>Exome Sequencing Project</i> , <i>1000 Genomes Project</i> , o <i>Exome Aggregation Consortium</i> .
	PM3	la variante está <i>in trans</i> ¹³ con otra variante categorizada como patogénica (Para patologías recesivas).
	PM4	La variante genera cambios en la longitud de la proteína, debido a inserciones o delecciones <i>in-frame</i> ¹⁴ o variante de tipo <i>stop-loss</i> ¹⁵ .
	PM5	Variante <i>novel missense</i> , donde anteriormente se ha registrado otra <i>missense</i> como patogénica.
	PM6	Variante <i>de novo</i> , pero no ha sido confirmada con los progenitores.

¹¹ **Variante *de novo*:** Variante genética que no está presente en ninguno de los padres, apareciendo por primera vez en un individuo.

¹² **Variante *missense*:** Variante en el ADN, que produce un cambio en el aminoácido resultante.

¹³ **Variantes *in trans*:** Variantes ubicadas en cromosomas homólogos diferentes (una en el cromosoma heredado de la madre y otra en el del padre).

¹⁴ **Mutación *in-frame*:** Inserción, delección o duplicación de nucleótidos en múltiplos de 3, que añade o remueve aminoácidos sin alterar el resto de la secuencia.

¹⁵ **Variante *stop-loss* o *truncating*:** Variante genética que modifica el codón de stop, produciendo una proteína más larga de lo normal.

Peso	Código	Criterio
<i>Supporting</i>	PP1	La variante es compartida por diversos miembros de la familia, también afectados con la enfermedad.
	PP2	Variante missense ubicada en una región conocida por tener pocas variantes <i>missense</i> benignas y las variantes missense son un mecanismo común de la enfermedad.
	PP3	La evidencia computacional sugiere un efecto nocivo sobre el gen.
	PP4	El paciente tiene un fenotipo que coincide con el que causa dicha enfermedad. Además el historial familiar consistente con el modo de herencia del trastorno.
	PP5	La variante ha sido clasificada como patogénica por algún reporte, pero no se presenta evidencia, para hacer una evaluación independiente.

Tabla 2: Criterios para clasificar una variante como benigna.

Peso	Código	Criterio
<i>Stand-alone</i>	BA1	La variante presenta una frecuencia alélica en la población mayor a 5%, según <i>Exome Sequencing Project</i> , 1000 Genomes Project, o Exome Aggregation Consortium
<i>Strong</i>	BS1	La variante tiene una frecuencia más grande que la que se espera para la enfermedad.
	BS2	La variante ha sido observada en adultos sanos, para una enfermedad de detección temprana.
	BS3	Existen estudios que respaldan que no existen efectos dañinos en la función de la proteína o en el <i>splicing</i> .
	BS4	Falta de segregación en los miembros afectados de una familia.
<i>Supporting</i>	BP1	Variante <i>missense</i> en región donde se conoce que las variantes truncadas causan la enfermedad.
	BP2	Para enfermedades dominantes, la detección de una variante in trans con una variante patogénica, puede ser considerada como evidencia de un impacto benigno de la variante.
	BP3	Deleciones o inserciones in-frame en regiones repetitivas sin funciones conocidas.

Peso	Código	Criterio
	BP4	La evidencia computacional sugiere que no hay un impacto sobre el gen o sobre la proteína.
	BP5	Variante encontrada en un caso con una base molecular alternativa para la enfermedad.
	BP6	Una fuente acreditada clasifica la variante como benigna, pero la evidencia no está disponible para hacer una evaluación independiente.
	BP7	Variante <i>synonymous</i> ¹⁶ para la cual los algoritmos de predicción de <i>splicing</i> no predicen ningún impacto en la secuencia consenso de <i>splice</i> ni la creación de un nuevo sitio de <i>splice</i> .

Cada variante encontrada debe ser comparada según los criterios presentados. Sin embargo, existen criterios como PP4, que pueden ser evaluados una sola vez para todas las variantes que presente un paciente. Por otro lado, hay criterios que requieren estudios más extensos, como aquellos que implican la evaluación de familiares, para los cuales a menudo no se dispondrá de esta información. Por último, existen criterios que no son aplicables a la enfermedad, como el PM3, que se aplica únicamente a enfermedades recesivas (Esto excluye a **ADPKD**, que es dominante).

Finalmente, para poder clasificar cada variante en una de las cinco categorías existentes, se debe evaluar los criterios que cumplieron cada variantes en las reglas presentadas en la Tabla 3.

En caso de no identificarse variantes patogénicas pero sí variantes clasificadas como de significado incierto (**VUS**) en individuos diagnosticados con la enfermedad, será necesario extender el análisis a los familiares. Esto permitirá obtener evidencia adicional que contribuya a esclarecer la patogenicidad o benignidad de dichas variantes.

2.3.3 Secuenciadores

Los secuenciadores son las herramientas fundamentales para llevar a cabo los análisis genéticos. La tecnología de secuenciación ha ido evolucionando con los años, pudiendo diferenciarse tres generaciones de secuenciadores. Para empezar, se tienen los secuenciadores de primera generación, donde el más representativo es el secuenciador de **Sanger** (1977), considerado actualmente como el “**gold standard**” de los secuenciadores debido a que produce lecturas de muy alta calidad. Sin embargo, dado el método que utiliza, solo puede leer una secuencia de **ADN** a la vez y aproximadamente de **700 pares de bases** de longitud, lo que lo hace muy costoso y lento para ser utilizado de forma masiva.

¹⁶ **Variante synonymous:** Variante en el ADN que no causa ningún cambio en el aminoácido ni en la proteína.

Tabla 3: Reglas para clasificar variantes según los criterios cumplidos

Clasificación	Regla
Patogénica	1 <i>Very strong</i> (PVS1) AND ≥ 1 <i>Strong</i> (PS1–PS4)
	1 <i>Very strong</i> (PVS1) AND ≥ 2 <i>Moderate</i> (PM1–PM6)
	1 <i>Very strong</i> (PVS1) AND 1 <i>Moderate</i> (PM1–PM6) AND 1 <i>supporting</i> (PP1–PP5)
	1 <i>Very strong</i> (PVS1) AND ≥ 2 <i>Supporting</i> (PP1–PP5)
	≥ 2 <i>Strong</i> (PS1–PS4)
	1 <i>Strong</i> (PS1–PS4) AND ≥ 3 <i>Moderate</i> (PM1–PM6)
	1 <i>Strong</i> (PS1–PS4) AND 2 <i>Moderate</i> (PM1–PM6) AND ≥ 2 <i>Supporting</i> (PP1–PP5)
Probablemente Patogénica	1 <i>Strong</i> (PS1–PS4) AND 1 <i>Moderate</i> (PM1–PM6)
	1 <i>Strong</i> (PS1–PS4) AND 1–2 <i>Moderate</i> (PM1–PM6)
	1 <i>Strong</i> (PS1–PS4) AND ≥ 2 <i>Supporting</i> (PP1–PP5)
	≥ 3 <i>Moderate</i> (PM1–PM6)
	2 <i>Moderate</i> (PM1–PM6) AND ≥ 2 <i>Supporting</i> (PP1–PP5)
Benigna	1 <i>Stand-alone</i> (BA1)
	≥ 2 <i>Strong</i> (BS1–BS4)
Probablemente Benigna	1 <i>Strong</i> (BS1–BS4) AND 1 <i>Supporting</i> (BP1–BP7)
	≥ 2 <i>Supporting</i> (BP1–BP7)
Significado Incierto (VUS)	No se cumplen ninguno de los criterios anteriores.
	La variante presenta criterios contradictorios tanto para ser considerada benigna como patogénica.

Los secuenciadores de segunda generación **NGS**, se dieron a conocer aproximadamente el 2004 y trajeron consigo una reducción importante en los costos por cada base leída, lo que se logra secuenciando miles de secuencias a la vez. En relación a la secuenciación de **Sanger**, existe una pérdida de calidad bastante pequeña.

Otro punto destacado de los secuenciadores **NGS** es su uso generalizado en una amplia gama de análisis, especialmente en el **diagnóstico de enfermedades genéticas**, debido a su calidad y capacidad para secuenciar múltiples muestras simultáneamente. Los secuenciadores más representativos de esta generación son los de **Illumina**, ampliamente utilizados en todos los campos de la genética. Estos secuenciadores están disponibles en varios modelos, adaptados a diferentes tamaños de laboratorios, y pueden generar lecturas de entre **300 y 600 pares de bases**. Sin embargo, debido a sus elevados costos, no están al alcance de todos los laboratorios.

Finalmente, los secuenciadores de tercera generación (**TGS**) surgieron alrededor de 2014, y se dividen en dos tipos de tecnologías distintas, por un lado la **PacBio** con su tecnología de secuenciación de molécula única en tiempo real (**SMRT**) y por otro lado **Oxford Nanopore Technology** con su tecnología de **nanoporos**. Estos nuevos secuenciadores se destacan por su capacidad para generar lecturas mucho más extensas, abarcando **miles de pares de bases**. Además, ofrecen la posibilidad de secuenciar en tiempo real. Sin embargo, presentan una **reducción en la calidad de las lecturas** en comparación con los secuenciadores de nueva generación.

2.3.3.1 Illumina MiSeq

El secuenciador **Illumina MiSeq** (Figura 13), corresponde a un secuenciador de segunda generación, perteneciente a la plataforma **Illumina**. Este es comúnmente utilizado en el diagnóstico de enfermedades genéticas, y en particular por **GEMINI** a través del *core facility* de la Universidad Austral de Chile, **Austral-Omics**. Este secuenciador puede generar lecturas de un máximo de 500 pares de bases, y con una **calidad de lectura promedio del 99,99%**.



Figura 13: Secuenciador Illumina MiSeq¹⁷

¹⁷ <https://illumina.com/systems/sequencing-platforms/miseq.html>

2.3.3.2 Oxford Nanopore MinION

El secuenciador **Oxford Nanopore MinION** (Figura 14), es un secuenciador de tercera generación comercializado desde 2014. Este secuenciador se destaca por ser el único secuenciador portátil gracias a sus dimensiones: tan solo 10.5 cm de largo, 2.3 cm de ancho y 3.3 cm de alto. Además, genera lecturas de miles de pares de bases, con una **calidad de hasta 99,96% de precisión**, dependiendo del kit de preparación de librerías que se utilice. Dada su portabilidad y baja calidad en las lecturas respecto a otros secuenciadores se utiliza principalmente en secuenciaciones de muestras en terreno, para la identificación de microorganismos, donde no se requieren lecturas con un nivel de detalle sobresaliente.



Figura 14: Secuenciador Oxford Nanopore MinION.¹⁸

Para llevar a cabo la secuenciación en este dispositivo, es necesario aplicar el proceso de preparación de librerías, en el que se agregan a las secuencias de **ADN**, los adaptadores (fragmentos de **ADN** que van ligados en los extremos de las secuencias) y una enzima motor que se une a los adaptadores (Figura 15). Existen varios kits de preparación de librerías para el **MinION**, donde la elección dependerá de las necesidades y requisitos del usuario, como tiempo, calidad, largo de las lecturas o número de muestras que se quiere secuenciar a la vez.

Después de la preparación de las librerías, las muestras se depositan en la **flowcell** que cuenta con una membrana resistente a la corriente, que contiene **2048 nanoporos**, posibilitando que cada uno de ellos permita el paso de una hebra de **ADN**. La **enzima motor** es guiada hacia el nanoporo gracias a unos *tethers* incluidos en la membrana y una vez que se acopla a estos, la enzima abre el **ADN** empujando

¹⁸

<https://www.labmedica.es/diagnostico-molecular/articles/294791698/nuevo-analisis-de-deteccion-de-patogenos-empareja-sondas-de-inversion-molecular-secuenciacion-de-ultima-generacion.html>

una hebra a través del nanoporo. Durante este proceso, el **nanoporo** está expuesto a una corriente iónica y cuando la hebra de **ADN** pasa por el **nanoporo**, produce interrupciones en la corriente que varían dependiendo del nucleótido. Aprovechando este principio, unos sensores miden continuamente la corriente, la cual es procesada por herramientas llamadas "**basecallers**", que traducen estas señales en las bases correspondientes (Figura 16).

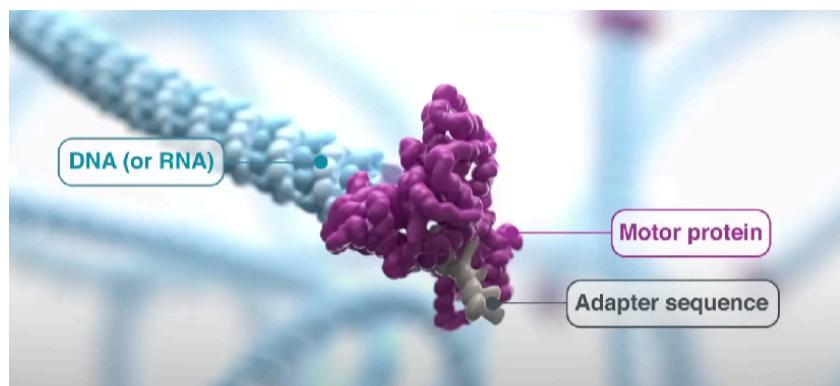


Figura 15: Secuencia de ADN, con adaptador y enzima fijada.¹⁹

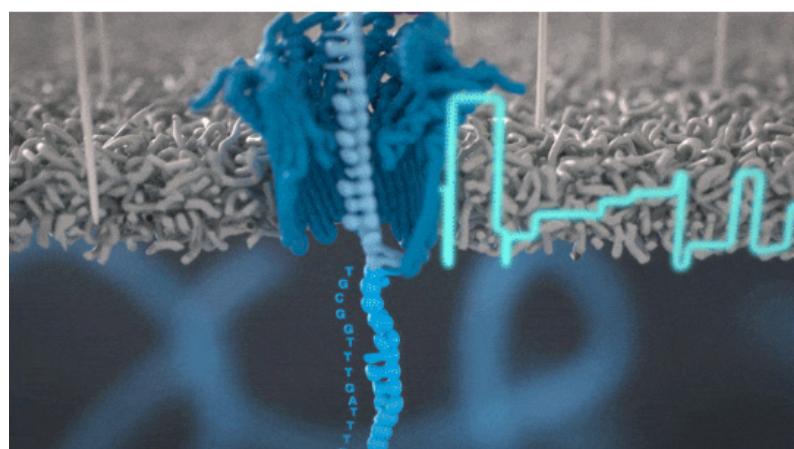


Figura 16: Señal captada por el paso de la hebra de ADN por el nanoporo.²⁰

2.3.4 Pipeline de análisis bioinformático

Una vez se ha completado la secuenciación de las muestras y se dispone de los archivos con las lecturas generadas por el secuenciador, es momento de realizar el análisis de estas lecturas. Las etapas que se llevan a cabo durante el análisis pueden variar según el objetivo y el tipo de secuenciador utilizado. Para el diagnóstico de enfermedades, utilizando un secuenciador de nanoporos, el ***pipeline*** de análisis se puede clasificar en seis etapas (ver Figura 17), como se detalla a continuación.

¹⁹ <https://www.youtube.com/watch?v=qzusVw4Dp8w>

²⁰ <https://nanoporetech.com/platform/technology/basecalling>

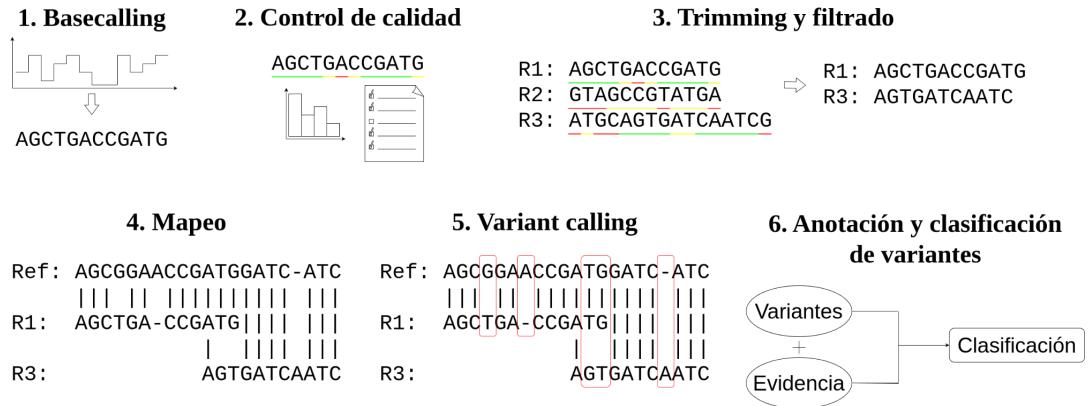


Figura 17: *Pipeline* para el análisis bioinformático.

2.3.4.1 Basecalling

El **basecalling** es el proceso que ocurre una vez se obtienen los datos de las lecturas brutas generadas por el secuenciador. En este proceso, **estos datos son traducidos a una secuencia de bases**, incorporando también la calidad de esta traducción. El tipo de dato bruto que genera el secuenciador depende directamente de la tecnología que utiliza. En el caso de los secuenciadores de nanoporos, estos registran los cambios de corriente causados por las hebras de **ADN** al pasar por los nanoporos.

Sin embargo, este proceso generalmente es invisible para el usuario y se lleva a cabo de forma automática por el secuenciador, a excepción de los secuenciadores de nanoporos, en los que el usuario tiene una mayor libertad para elegir el *software* que utilizará para este proceso, así como ajustar parámetros y configurar archivos de salida.

Esta libertad en la elección del *software* se debe principalmente a que no existe un método único para transformar las señales en bases, por lo que **las distintas herramientas emplean modelos estadísticos y algoritmos basados en redes neuronales recurrentes** para relacionar las señales con una secuencia de bases. Además, los secuenciadores de nanoporos están constantemente mejorando sus materiales, como la **flowcell** y los **kits de preparación de librerías**, lo que genera nuevas herramientas o versiones adaptadas a estos materiales.

2.3.4.2 Control de calidad

Esta etapa **consiste en determinar la calidad de los datos** en momentos específicos del análisis. Es común que se realice una vez que se obtienen las lecturas del secuenciador, las cuales están almacenadas en un archivo en formato **FASTQ²¹**. Se calculan métricas como la calidad de las lecturas para cada una de las bases leídas, así como la longitud de las lecturas más cortas y más largas, la longitud promedio y el número total de lecturas generadas.

²¹ **FASTQ:** Formato de archivo de texto plano que almacena información de secuencias de **ADN** o aminoácidos, además de la calidad de lectura de cada base.

Estas métricas tienen como propósito determinar si la calidad de los datos es suficiente para obtener resultados confiables en el análisis, si es necesario filtrar y recortar ciertas lecturas, o en el peor de los casos, repetir el ensayo.

El control de calidad también se puede llevar a cabo después del proceso de mapeo, con el fin de determinar si las lecturas generadas cubren las regiones de interés y conocer la profundidad de las lecturas (es decir, el número de veces que se leyó una base), así como la calidad del mapeo. Este control de calidad se lleva a cabo analizando la información de las lecturas mapeadas, almacenada en los archivos **SAM**²² o **BAM**²³, generados durante la etapa de mapeo.

2.3.4.3 Trimming y filtrado de lecturas

Esta etapa consiste en recortar aquellas lecturas que puedan corresponder a adaptadores introducidos para realizar el proceso de secuenciación, además se busca podar fragmentos de lecturas que no hayan cumplido con un mínimo de calidad y eliminar aquellas que no cumplan con un largo mínimo.

2.3.4.4 Mapeo

El mapeo corresponde a la etapa en la que las lecturas que han pasado los filtros se alinean contra un genoma de referencia. Esta etapa consta de dos pasos: primero, se crea un índice del genoma de referencia con el objetivo de facilitar la ubicación de las lecturas. Luego, utilizando el índice generado, se alinean las lecturas en el gen de referencia, lo que genera nuevos datos para cada lectura, como la posición de la lectura con respecto al gen de referencia y la calidad del mapeo. Entre otros datos, se genera un reporte mediante una cadena de texto llamada **CIGAR** (*Compact Idiosyncratic Gapped Alignment Report*), que describe las concordancias, discrepancias, inserciones y delecciones con respecto a la secuencia de referencia. Estos datos se almacenan generalmente en un archivo **SAM** (*Sequence Alignment Map*).

2.3.4.5 Variant calling

El *variant calling* (llamado de variantes) tiene por objetivo identificar todas aquellas variantes presentes en las lecturas generadas ya sean **SNPs** o **INDELS**. Esta etapa tiene como desafío distinguir aquellas variantes que realmente están presentes en el paciente, a diferencia de las que pueden ser atribuidas a errores de lecturas conocidas como artefactos. Para esto, las diversas herramientas utilizan diferentes enfoques, desde examinar la calidad de las lecturas en la posición de la variante, hasta analizar las bases vecinas y el genoma de referencia. Esta comparación se basa en métodos estadísticos, así como en el uso de diferentes algoritmos como los Modelos de Markov Ocultos en Pares (*Pair-Hidden Markov Models*) y modelos basados en redes neuronales. Esta etapa genera un archivo **VCF** (*Variant Call Format*) que contiene información referente a la posición, el tipo y la frecuencia alélica de las variantes encontradas, entre otros.

²² **SAM (Sequence Alignment Map):** Formato de archivo para almacenar alineaciones de secuencias de ADN generadas en estudios de secuenciación.

²³ **BAM (Binary Alignment Map):** Versión binaria y comprimida del archivo **SAM**.

2.3.4.6 Clasificación, anotación y filtrado de variantes

Por último esta etapa consiste en clasificar cada una de las variantes encontradas dentro de una de las cinco categorías creadas por *American College of Medical Genetics (ACMG)* y la *Association for Molecular Pathology (AMP)* presentadas en la sección 2.3.2 de este documento (1.Patogénica, 2.Probablemente patogénica, 3.Variante de significado clínico incierto (**VUS**), 4.Probablemente benigna, 5.Benigna).

Para clasificar las variantes, se realizan consultas en bases de datos públicas de variantes y se utiliza *software* de predicción de patogenicidad para recopilar evidencia sobre su posible patogenicidad. Una vez obtenida esta evidencia para cada variante, se anotan estas clasificaciones utilizando herramientas que manipulan archivos **VCF**. Además se puede hacer un filtrado de aquellas que no cumplan con nivel de calidad o profundidad de lectura (Cuántas lecturas contienen la variante).

Para obtener la clasificación final de la variante, se analiza esta evidencia de acuerdo con los criterios establecidos por el bioinformático a cargo o siguiendo el manual de clasificación de variantes basado en las directrices de **ACMG/AMP**, que evalúan diversos criterios en relación con la variante.

Los resultados generados deben ser analizados por el experto a cargo del estudio, ya que en caso de que solo se encuentren variantes clasificadas como **VUS**, se debe evaluar ampliar el análisis al grupo familiar mediante un estudio de segregación para poder obtener más evidencia que permita objetivar la patogenicidad de la variante.

3. REVISIÓN SISTEMÁTICA

3.1 Objetivo de la revisión

El desarrollo de un *pipeline* bioinformático para el análisis de secuencias generadas por el secuenciador *Oxford Nanopore MinION*, con el propósito de diagnosticar de enfermedades genéticas, conlleva una serie de decisiones, siendo una de las más importantes, la selección de herramientas para cada uno de los pasos que posee el proceso de análisis. Por esta razón, se desarrollará una revisión sistemática para identificar las distintas herramientas utilizadas en cada una de las etapas del pipeline, las cuales son: 1) *Basecalling*, 2) Control de calidad, 3) *Trimming* y filtrado de lecturas, 4) Mapeo 5) *Variant calling* 6) Clasificación, anotación y filtrado de variantes.

3.2 Metodología

Para llevar a cabo la revisión sistemática se hizo uso de la plataforma **Parsifal**²⁴, la cual proporciona un conjunto de pasos estructurados que aseguran la reproducibilidad de los resultados. Este proceso comenzó con la formulación de una pregunta de investigación que refleja la necesidad planteada en el objetivo de la revisión, además de servir como guía para definir las palabras clave. Posteriormente, se construyó la cadena de búsqueda basada en las palabras clave con el objetivo de obtener la mayor cantidad de artículos útiles. Finalmente, se establecieron criterios de inclusión y exclusión para filtrar los artículos obtenidos.

3.2.1 Pregunta de revisión

A partir del objetivo de la revisión, se planteó la siguiente pregunta, redactada en inglés para obtener una mayor cantidad de resultados. Además, de esta forma se facilita la búsqueda de las palabras clave que se utilizarán en la cadena de búsqueda.

What tools are used at various stages of a bioinformatics analysis for the diagnosis of genetic diseases, based on sequences ideally generated by the Oxford Nanopore MinION?

3.2.2 Fuente y cadena de búsqueda

Mediante la revisión preliminar del estado del arte, se lograron identificar algunos '*keywords*' comúnmente empleados en el campo de la genética. Estas '*keywords*' se presentan, junto con algunos de sus sinónimos en la Tabla 4.

²⁴ <https://parsif.al/>

Tabla 4: *keywords* y sinónimos

Keyword	Sinónimos
MinION	3GS ONT <i>Oxford Nanopore MinION</i> TGS <i>Third-generation sequencing</i> <i>Nanopore</i>
<i>bioinformatics analysis</i>	<i>genomic analysis</i> <i>sequence analysis</i>
<i>bioinformatics tools</i>	(sin sinónimos)
<i>disease detection</i>	<i>clinical diagnostics</i> <i>disease diagnosis</i> <i>genetic diseases</i>
<i>pipeline</i>	<i>framework</i> <i>protocol</i> <i>workflow</i>

Con los “*keywords*” más importantes y sus sinónimos recopiladas, se creó el siguiente cadena de búsqueda:

```

(
    "MinION" OR "3GS" OR "ONT" OR "Oxford Nanopore MinION" OR
    "TGS" OR "Third-generation sequencing" OR "bioinformatics analysis" OR
    "genomic analysis" OR "sequence analysis" OR "Nanopore" OR "Nanopore
    technology"
)
AND
(
    "pipeline" OR "framework" OR "workflow"
)
AND
(
    "bioinformatics tools" OR "bioinformatics workflows" OR "disease
    detection" OR "clinical diagnostics" OR "disease diagnosis" OR "genetic
    diseases" OR "variant calling" OR "variant detection"
)

```

Luego se utilizó la cadena para la búsqueda de artículos en revistas científicas, para lo que se consultó en la plataforma “**ISI Web of Science**” (WoS), el día viernes 5 de abril del año 2024 sólo considerando el *abstract* de los artículos en la búsqueda.

3.2.3 Criterios de selección y exclusión

Para llevar a cabo la selección imparcial y replicable de los artículos, se fijaron criterios de inclusión y exclusión que debían cumplir los artículos, estos son:

Criterios de Inclusión:

1. Comparación de herramientas para alguna etapa del análisis.
2. *Pipeline* de análisis para el diagnóstico de enfermedades.
3. *Pipeline* que contiene alguna etapa en común con el diagnóstico de enfermedades.
4. Presentación de herramientas para alguna etapa del análisis.

Criterios de Exclusión:

1. No menciona ninguna etapa del proceso de análisis, ni herramienta para el diagnóstico de enfermedades.
2. No menciona herramientas para análisis bioinformático.
3. No se relaciona con lo que se busca en la revisión.
4. Las herramientas que se mencionan no son compatibles con el secuenciador MinION o no se ajustan a las características del secuenciador.

3.2.4 Extracción de la información

Para recopilar la información relevante para la revisión, se llevó a cabo un formulario de extracción de datos, con los siguientes campos:

1. DOI y año.
2. Nombre de la publicación.
3. Tipo de artículo (“Presentación de *pipeline*”, “Comparación de herramientas”, “Presentación de herramienta(s)” o “Optimización de parámetros”).
4. Objetivo del *pipeline* (“No aplica”, “General”, “Otro”, “Diagnóstico de enfermedades”, “Ensamblaje de genoma”, “Detección de bacteria, causante de enfermedades”, “Detección de variantes” o “Detección de virus”).
5. Tipo de secuenciador que se usó.
6. Herramienta y usos.
7. Hallazgos extras.

3.3 Resultados de la revisión

3.3.1 Artículos aceptados

Mediante el *string* de búsqueda implementado, **se obtuvieron 104 artículos**, de los cuales se aceptaron 52 (Figura 18) bajo los criterios de inclusión y exclusión. Además en la Figura 19 se puede apreciar la distribución de los años de los artículos, donde se observa un aumento en el número de artículos por año, y una disminución en los dos últimos años, lo que se puede explicar debido a la fecha en la que fue realizada la búsqueda y el retardo en la indexación de publicaciones en la que la plataforma WoS.

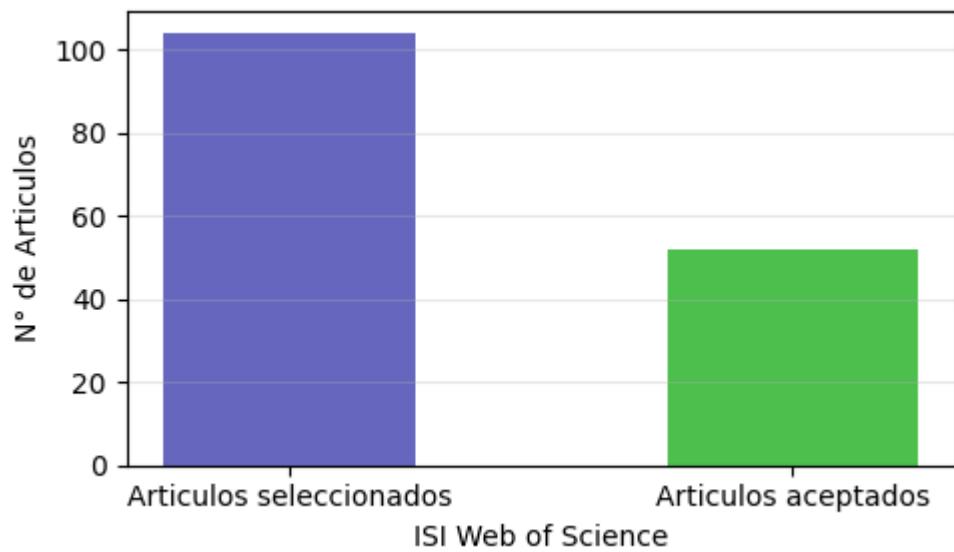


Figura 18: Número de artículos seleccionados versus artículos aceptados.

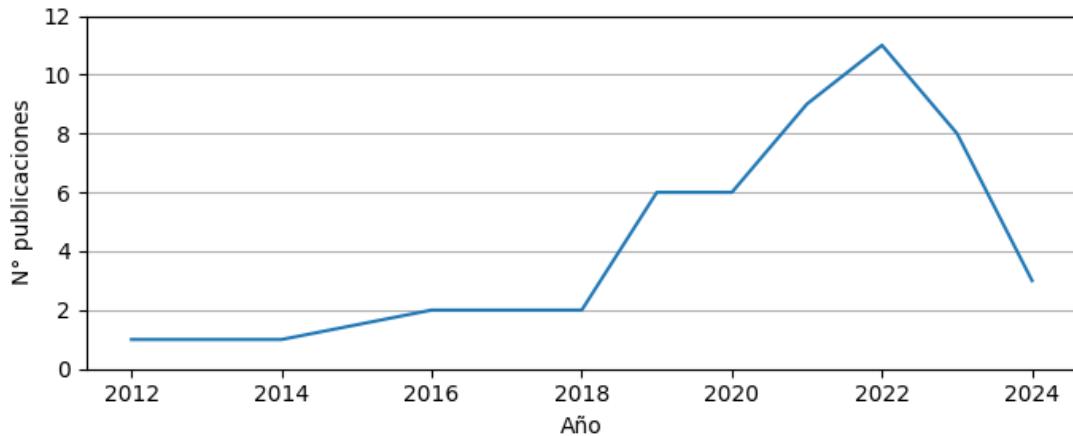


Figura 19: Número de publicaciones aceptadas por año.

La información extraída de los artículos, consistente en la información definida en el punto 3.2.4, se puede consultar en el Anexo A.

3.3.2 Síntesis de los resultados

Como se mencionó anteriormente, el número de publicaciones que declaran el empleo de algún secuenciador ha ido en aumento con los años (Figura 19), esta tendencia también se puede observar en el uso de secuenciadores de nanoporos como el **MinION** (Figura 20). Además si se filtran aquellos artículos que hacen uso de esta tecnología para buscar variantes en el **ADN**, también se puede observar un aumento en el número de artículos por año (Figura 21). Esto puede atribuirse al mejoramiento de la calidad de las lecturas generadas por el secuenciador, lo que ha posibilitado su uso para la detección de variantes causantes de enfermedades genéticas, como mencionan Orsini et al. (2018) y Салахов et al. (2022).

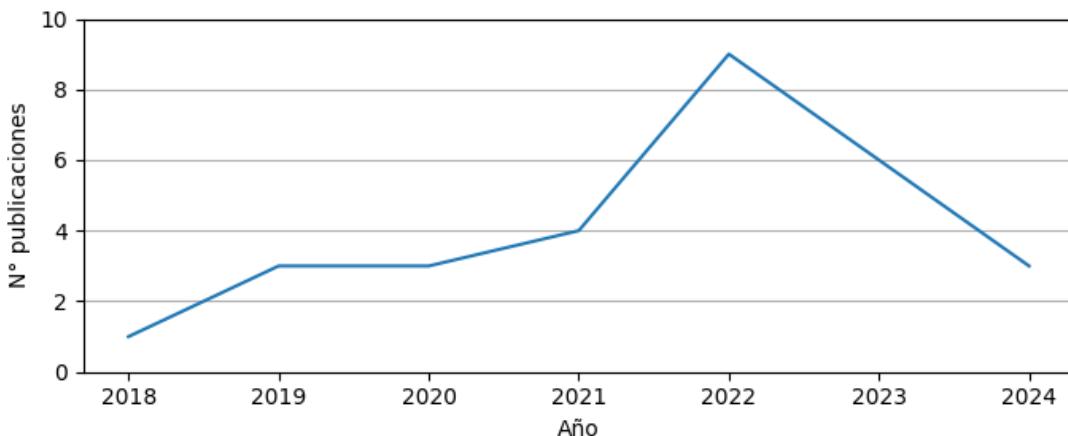


Figura 20: Gráfico del número de publicaciones por año, en el que se utiliza un secuenciador de nanoporos.

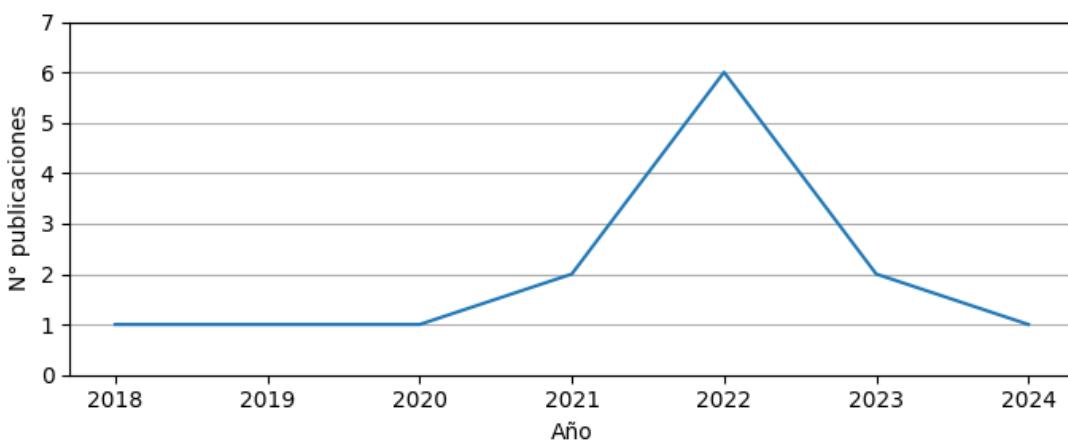


Figura 21: Gráfico del número de publicaciones por año, en el que se utiliza un secuenciador de nanoporos para la detección de variantes.

En cuanto al número de herramientas halladas para cada una de las etapas principales que contempla un ***pipeline* de análisis bioinformático** para el diagnóstico de una enfermedad genética, estas se pueden ver en la Tabla 5.

A pesar de que se encontró una gran variedad de herramientas para cada etapa, en todas ellas existen herramientas que se repiten con mayor frecuencia. En donde se ve más marcado este punto es la etapa de mapeo, en donde se puede observar dos herramientas que se repiten constantemente. Su uso responde al secuenciador que se utilice: en los artículos que se utilizaban secuenciadores **NGS**, en su mayoría se ocupaba la herramienta **BWA** (Burrows-Wheeler Aligner), mientras que los artículos que reportan el uso del secuenciador **MinION**, se observa que **Minimap2** es la herramienta más popular para la etapa de mapeo. Esto puede tener que ver con la especificidad en las herramientas, dado que como Carbo et al. (2023) mencionan, **Minimap2** está diseñado para el análisis de secuencias de plataformas con una tasa de error de lectura relativamente alta, como lo es el secuenciador **MinION**.

Tabla 5: Número de herramientas por utilidad.

Utilidad	Número de herramientas	Herramienta más utilizada
<i>Basecalling</i>	3	Guppy
Control de calidad	14	FastQC
<i>Trimming</i> y filtrado	10	Trimmomatic / Samtools
Mapeo	27	BWA / Minimap2
<i>Variant calling</i>	41	GATK
Clasificación, anotación y filtrado de variantes	36	ANNOVAR / SnpEff
Base de datos de variantes	17	ClinVar
<i>SV/CNV calling</i>	22	Sniffles
Herramientas extras	23	Samtools

Esto mismo ocurre en las herramientas de *variant calling*, donde en secuenciadores **NGS**, se puede observar un mayor uso de las herramientas **GATK** y **BCFtools**. Mientras que en los secuenciadores de nanoporos se puede observar más diversidad en las herramientas, pero también una tendencia por herramientas como **Medaka**, **Longshot**, **DeepVariant** y **Clair3**.

Es relevante destacar que **Medaka**, **DeepVariant** y **Clair3** emplean algoritmos basados en redes neuronales, como señalan Ramachandran et al. (2021) y Helal et al. (2022). Esta elección sugiere una inclinación hacia herramientas con enfoques predictivos en el contexto de los secuenciadores de nanoporos. Esta tendencia podría explicarse por la menor calidad de las lecturas obtenidas en estos secuenciadores, lo que conlleva a la búsqueda de herramientas más sofisticadas capaces de distinguir entre los errores de lectura y las variantes reales.

La variabilidad en los distintos enfoques utilizados por las distintas herramientas también ocasiona que los resultados de este proceso puedan ser variables, incluso el rendimiento de la misma herramienta podría variar dependiendo del gen o región que se está analizando. Sumado a lo anterior, las herramientas reciben parámetros que pueden alterar aún más sus resultados finales, lo que causa que la elección de la herramienta a utilizar no pueda ser tomada sólo en base descripciones de funcionamiento, resultados específicos o generales de publicaciones y por lo tanto se necesite basar en evidencia empírica para el caso particular que se está estudiando.

Como se mencionó anteriormente, los parámetros con los que se utilizan las herramientas también repercuten en la calidad de los resultados. Esto presenta un nuevo problema el cual es la elección de los mejores parámetros para cada herramienta. En esta línea Svensson et al. (2019), mencionan que si bien es ventajoso personalizar las herramientas para una situación específica, no siempre es obvio qué efecto tendrá el cambio de parámetros en el resultado. Esto sumado a que en un *pipeline* secuencial, como lo es este tipo de análisis, varias de las etapas contienen herramientas las cuales poseen parámetros que pueden ser optimizados, lo que aumenta drásticamente el número de combinaciones posibles de parámetros, por lo que, la mejor forma de optimizarlos es usar técnicas como ***grid search*** o ***Doepipeline***, mencionada en su artículo.

También se encontró que existen herramientas optimizadas para maximizar los resultados de datos provenientes de ciertos secuenciadores. Esto principalmente ocurre en las etapas de mapeo y *variant calling*, donde las herramientas se adaptan para trabajar con las características de las lecturas que generan los secuenciadores. Algunas de estas características pueden ser el largo de las lecturas generadas o la tasa de errores que se pueden generar. Por lo tanto, una herramienta optimizada para **NGS** y por lo tanto, que presente muy buenos resultados para el análisis de un gen determinado, desde lecturas de un secuenciador de nueva generación, podría entregar resultados muy diferentes para el mismo caso, pero con lecturas de un secuenciador como el **MinION**. Otro caso semejante es el de herramientas que reciben como *input* archivos en formato de una herramienta determinada, como es el caso de la herramienta para llamado de variantes **SOAPsnp** que recibe, como se menciona en Dolled-Filhart et al. (2013), archivos **SOAP** provenientes de la herramienta de mapeo **SOAP3**, lo que imposibilita la opción de utilizar **SOAPsnp** con otra herramienta de mapeo.

Finalmente, se identificaron herramientas que aunque no están directamente relacionadas con una etapa específica del proceso, desempeñan un papel crucial para garantizar el éxito del *pipeline* de análisis. Algunas de ellas se enfocan en la conversión de formatos, como **VCFtools**, **Poretools toolkit** y **pod5 package**. Otras herramientas, como **Nanopolish** o **Cerebro**, pueden corregir errores de lectura o de variantes. Asimismo, se encontraron herramientas como **Cutadapt**, que resultan valiosas para llevar a cabo la demultiplexación de las lecturas generadas por el secuenciador. Este último proceso es especialmente relevante cuando se ha realizado la secuenciación de múltiples regiones del **ADN** o muestras provenientes de diferentes pacientes de manera simultánea.

4. METODOLOGÍA Y RESULTADOS PARA LA EVALUACIÓN Y SELECCIÓN DE HERRAMIENTAS

En este capítulo se realiza un repaso de los casos clínicos disponibles para el desarrollo y validación del funcionamiento del *pipeline*. Además, se proporcionarán detalles sobre los materiales utilizados en la secuenciación de las muestras y las características de dicho proceso. Posteriormente, se exponen las herramientas identificadas en cada etapa, seguidas por la descripción de los métodos empleados para su evaluación y selección en cada fase del *pipeline* de análisis.

El objetivo principal de este capítulo es identificar el conjunto de herramientas que logre los mejores resultados en cada una de las etapas correspondientes, permitiendo detectar la variante previamente identificada en cada uno de los pacientes.

4.1 Casos clínicos

Con el objetivo de evaluar el funcionamiento de las herramientas y validar el *pipeline* final, se cuenta con un grupo de ocho pacientes diagnosticados con **ADPKD**. Este grupo incluye pacientes con distintos tipos de variantes (Tabla 6), tales como **SNPs**, **inserciones** y **deleciones**. Además, dos de estos pacientes presentan **variantes en regiones intrónicas** del gen. Estas muestras permiten abarcar la mayoría de los escenarios posibles que podrían surgir al buscar variantes en nuevos pacientes.

Tabla 6: Variantes confirmadas de cada paciente.

Paciente	Ubicación	Tipo	Ref	Alt	Efecto
S1	Exón 2	SNP	T	C	<i>Missense</i>
S2	Exón 25	SNP	A	C	<i>Missense</i>
S3	Exón 17	SNP	C	T	<i>Missense</i>
S4	Exón 29	Deleción	GGCTG	-	<i>Truncating</i>
S5	Intrón 37	SNP	C	A	<i>Truncating</i>
S6	Intrón 22	SNP	A	G	<i>Truncating</i>
S7	Exón 40	Inserción	-	G	<i>Truncating</i>
S8	Exón 38	SNP	G	A	<i>Missense</i>

Si bien las muestras fueron escogidas para ser representativas de distintas situaciones, **esta información no fue provista al tesista al momento de estudiarlas**, de tal forma de validar la efectividad del proceso mediante una estrategia a ciegas.

4.2 Preparación de librerías y secuenciación

Para llevar a cabo el análisis del gen *PKD1* en las muestras de los distintos pacientes, se realizó una secuenciación dirigida (*targeted sequencing*), la cual permite obtener una mayor profundidad de cobertura en una región específica de interés. Con este fin, se buscó amplificar los 46 exones del gen, lográndolo mediante nueve PCRs de lecturas largas (*long-read PCR*) (ver Figura 22), propuestos por Tan et al. (2012). Como resultado de lo anterior, también se obtuvo una gran parte de secciones intrónicas, lo cual permite estudiar no sólo los exones del gen *PKD1*, sino también las regiones intrónicas completas o contiguas a cada uno de los 46 exones.

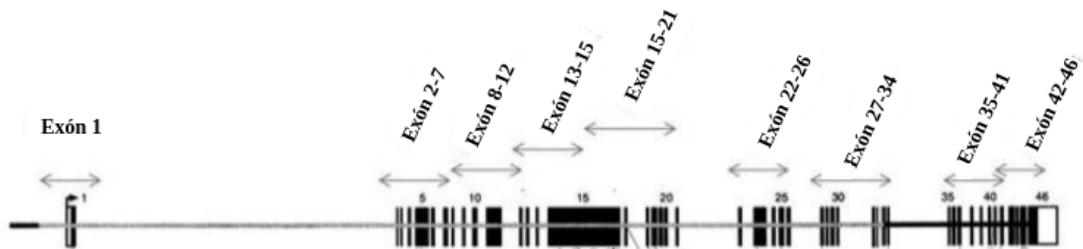


Figura 22: Mapa del gen *PKD1* con la posición de los nueve partidores utilizados para amplificar los exones del gen mediante *long-reads PCR*²⁵.

Una vez amplificadas las regiones de interés, se utilizó el kit de librerías *Native Barcoding Kit 24 V14* para preparar las muestras. Este kit está diseñado para obtener una calidad superior a Q20, lo que se traduce en más del 99% de precisión en las lecturas. Además, permite secuenciar hasta 24 muestras de diferentes pacientes en una misma *flowcell*. Por último, este kit permite leer ambas hebras del ADN, insertando adaptadores en ambos extremos de cada hebra (como se muestra en la Figura 23), que posibilita mejorar aún más la calidad de las lecturas mediante el método de *duplex basecalling* que se explica más adelante.

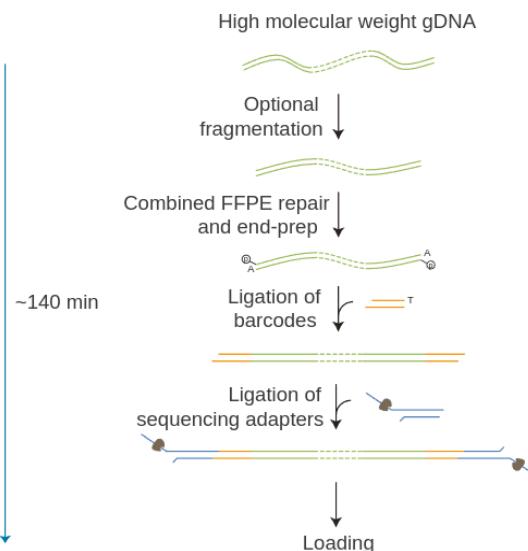


Figura 23: Proceso de preparación de librerías para secuenciación de ADN²⁶.

²⁵ <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3391417/>

²⁶ <https://store.nanoporetech.com/native-barcoding-kit-24-v14.html>

Finalmente, las ocho muestras fueron secuenciadas utilizando una única *flowcell*, correspondiente a la **versión R10.4.1**, qué era la más reciente al momento de realizar la secuenciación. Esta *flowcell* se diferencia de la versión anterior en que presenta dos sensores por poro, lo que mejora la precisión de las lecturas.

4.3 Búsqueda de herramientas

A continuación, se describen los requisitos y métodos empleados para la búsqueda y selección de herramientas en cada etapa, con el propósito de elegir aquellas que, en conjunto, permitan la correcta identificación de las variantes previamente detectadas en los casos clínicos con los que se cuenta. Dada la variabilidad en los objetivos, funcionamientos y métodos de las herramientas utilizadas en cada etapa, se optó por establecer criterios de selección exclusivos para cada una de ellas.

No obstante lo anterior, existen criterios generales que se aplican a todas las etapas del proceso. En primer lugar, la prioridad principal radica en seleccionar aquellas herramientas que garanticen los mejores resultados. Si dos herramientas cumplen la misma función y producen resultados idénticos, se seleccionará aquella con los menores tiempos de ejecución. Finalmente, en el caso de que una tarea requiera el uso de una combinación de herramientas para alcanzar un resultado, se optará por el conjunto de herramientas más pequeño.

Por último, la ejecución de cada una de las herramientas se llevó a cabo en el supercomputador Patagón²⁷ de la Universidad Austral de Chile (**FONDEQUIP EQM180042**). Este enfoque permitió acceder a un entorno de ejecución más controlado, facilitando la medición precisa de los tiempos y el funcionamiento de las distintas herramientas.

4.3.1 Basecalling y demultiplexación

Como mencionamos anteriormente, la etapa de **basecalling** implica transformar las señales eléctricas generadas por las hebras de ADN al pasar por los nanoporos en secuencias de letras para su posterior análisis.

Durante la revisión sistemática, se **hallaron tres herramientas de basecalling: Albacore, Guppy y Dorado**. Sin embargo, es importante señalar que cada una de estas herramientas fue desarrollada por *Oxford Nanopore Technologies*, siendo cada una un reemplazo de la anterior. Con esto, se consiguió mejorar progresivamente la precisión y eficiencia del proceso de *basecalling*. Actualmente, **Dorado es la mejor opción para realizar el basecalling**, ya que fue desarrollada y entrenada para ser utilizada con la *flowcell R10.4* y el **kit de preparación de librería v14**, el cual se utilizó para esta secuenciación.

²⁷ <https://patagon.uach.cl/>

Además, **Dorado** introduce un nuevo enfoque llamado *duplex basecalling*, que utiliza información tanto de la señal como de la traducción de ambas hebras del ADN para combinarlas en una sola lectura, mejorando así la precisión (Figura 24). Basandonos en estos puntos, se decidió utilizar **Dorado** como herramienta para la etapa de **basecalling**, aprovechando su capacidad de *duplex basecalling*.

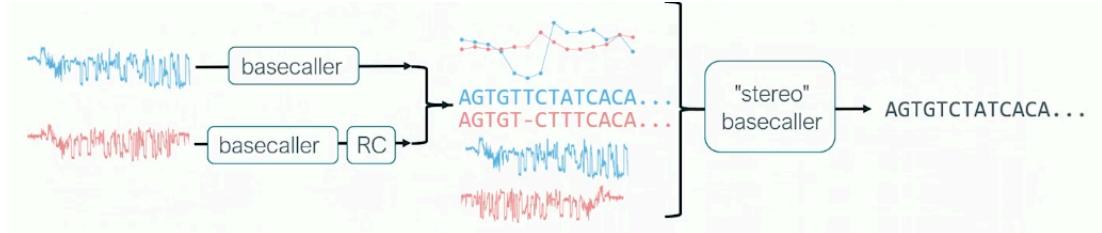


Figura 24: Funcionamiento de *duplex basecalling*²⁸.

Posterior al **basecalling**, se procedió a realizar el proceso de **demultiplexación**, que consiste en clasificar las lecturas según los pacientes a los que pertenecen. Este proceso se llevó a cabo utilizando la herramienta **Dorado**, la cual separa las muestras en diferentes archivos según los barcodes que encuentra en las lecturas. **Dorado ofrece dos enfoques** para este proceso: buscar los *barcodes* en cualquiera de los dos extremos de las lecturas, o asegurarse de encontrar ambos barcodes en cada lectura antes de clasificarla. El segundo enfoque reduce la tasa de falsos positivos al clasificar las lecturas, pero también disminuye la cantidad de lecturas clasificadas para cada *barcode*. Por lo tanto, **se decidió utilizar el primer enfoque**, priorizando así un mayor número de lecturas para cada paciente.

Finalmente, se generó un archivo para cada paciente y otro archivo para las lecturas que no pudieron ser asignadas a ninguno de los ocho pacientes debido a que no se les identificó *barcode*.

4.3.2 Control de calidad

Como se mencionó en el punto 2.3.4.2, esta etapa tiene por objetivo evaluar la calidad de los datos, con el fin de poder tomar las decisiones pertinentes referente a continuar o no con el análisis, o el nivel de *trimming* y filtrado que se va a aplicar a las muestras. Por lo tanto, es importante que las herramientas seleccionadas para esta etapa dispongan de una serie de métricas enfocadas en distintos atributos de las muestras, **permitiendo al usuario evaluar rápidamente la calidad de ellas**. Se consideraron las siguientes métricas para aceptar las lecturas como válidas:

- R1. Distribución de calidad al interior de las lecturas.
- R2. Distribución del largo de las lecturas.
- R3. Porcentaje de GC (guanina o citosina).
- R4. Porcentaje de lecturas duplicadas.
- R5. Porcentaje de adaptadores presentes.

Además, se espera aplicar métricas relacionadas con el mapeo, que serán útiles para determinar si se cubrieron las regiones de interés, en este caso, los exones del gen **PKD1** y la mayoría de los intrones. Destacando que solo cuatro intrones no se han

²⁸ <https://www.youtube.com/watch?v=8DVMG7FEBys>

secuenciado completamente, como se muestra en la Figura 22. También se analizará la profundidad de lectura promedio obtenida, lo cual permitirá tener una mayor confiabilidad de las bases leídas por el secuenciador. Por lo tanto es importante que la herramienta entregue información respecto de:

- RM1. Cobertura del mapeo.
- RM2. Profundidad promedio.

4.3.2.1 Control de calidad post-secuenciación

En relación con el primer objetivo, que consiste en evaluar la calidad de los datos post-secuenciación, se procedió a recopilar información sobre cada una de las herramientas identificadas en la revisión sistemática. Se encontró que ocho de ellas generan datos estadísticos útiles para evaluar la calidad de las secuencias. La información de los formatos de entrada, formatos de salida y datos estadísticos generados puede ser revisada en el Anexo B (desde la Tabla 84 hasta la Tabla 90). Sin embargo, cabe destacar que la herramienta **NanoOK** no fue evaluada, ya que no admite archivos en formato **FASTQ** (con más de una lectura), **POD5²⁹** o **sequence_summary.txt³⁰**. En la Tabla 7 se puede ver una breve descripción de cada una de las herramientas evaluadas en esta etapa.

Tabla 7: Herramientas para evaluar la calidad de la secuenciación.

Herramienta	Descripción
FastQC	Herramienta que genera reportes para el control de calidad.
Fastp	Herramienta que genera un reporte para el control de calidad y a su vez realiza el <i>trimming</i> y filtrado de lecturas.
MinIONQC	Librería de R para el control de calidad de secuencias provenientes de secuenciadores MinION y PromethION .
NanoPlot	Herramienta para generar reportes que permiten visualizar la calidad de lectura largas y el mapeo de estas.
NanoQC	Herramienta de control de calidad de lecturas largas.
PycoQC	Herramienta que genera reportes de control de calidad a partir de resúmenes de secuenciación.
Seqkit	Conjunto de herramientas para la manipulación de archivos FASTQ .

En general se puede observar que las herramientas reciben como entrada archivos en formato **FASTQ** o un archivo **sequence_summary.txt**, siendo **FASTQ** utilizado por la mayoría del *software*. Las herramientas **MinIONQC** y **PycoQC** que hacen uso del

²⁹ **POD5**: Formato utilizado por secuenciadores **ONT**, para almacenar datos de señales eléctricas crudas generadas durante el proceso de secuenciación.

³⁰ **sequence_summary.txt**: Archivo generado por el *software* MinKNOW. Este archivo contiene información detallada sobre las lecturas obtenidas, incluyendo datos sobre la calidad, la temporalidad y el canal de la *flowcell* que produjo cada lectura.

archivo ***sequence_summary.txt***, aprovechan la información extra de este archivo, para graficar la evolución temporal de los datos producidos por el secuenciador, además de los resultados generados en cada canal de la *flowcell*. Sin embargo, aunque esta información podría ser útil para lograr un uso más eficiente del secuenciador, esto tiene el inconveniente que se desvía del propósito principal del *pipeline* y podría causar que el usuario pierda el enfoque en los datos verdaderamente importantes para el análisis de las secuencias. Además, esta información ya es proyectada en tiempo de ejecución por el *software MinKNOW* el cual es responsable del funcionamiento del secuenciador **MinION**.

Una de las limitaciones al utilizar herramientas que aceptan el archivo ***sequence_summary.txt*** como entrada es que este archivo se genera antes de realizar el *basecalling*. Debido a que la etapa de *basecalling* utiliza redes neuronales para convertir las señales eléctricas en secuencias de ADN, la calidad de las lecturas puede variar dependiendo del modelo de red utilizado. Esto implica que la información de calidad proporcionada en el ***sequence_summary.txt*** podría no reflejar con precisión los resultados finales después de realizar el *basecalling*. Además, si se secuencian varias muestras de manera conjunta, estas no estarán demultiplexadas en el archivo inicial, lo que impide un análisis separado de la calidad de cada muestra individual antes de completar la etapa de demultiplexación. Por lo tanto, las herramientas **MinIONQC** y **PycoQC** no son adecuadas para este *pipeline*.

Por otro lado, también se puede observar que todos los programas generan reportes que incluyen información sobre la calidad general y/o gráficos con las distintas distribuciones de calidad o largo de las lecturas. Esto a excepción del *software Seqkit*, el cual entrega un reporte en formato **TSV**, solo con valores en texto plano. El resto de los reportes se presentan en archivos **HTML**. Entre estos *software* se puede resaltar los reportes elaborados por la herramienta **Fastp**, el cual, mediante el uso de un solo comando permite además realizar la etapa de **trimming** y **filtrado** de las lecturas, incluyendo en el reporte la comparación de los resultados obtenidos antes y después de recortar y filtrar las secuencias de peor calidad.

Al evaluar los requisitos que estas herramientas deben cumplir, los cuales fueron propuestos como indispensables al inicio del punto 4.3.1. Se tiene como resultado lo expuesto en la Tabla 8, donde se observa que la única herramienta en cumplir con todos los requisitos propuestos, es **FastQC**.

Finalmente, a partir de los resultados presentados en la Tabla 8, se puede concluir que la herramienta **FastQC** es la mejor opción para analizar la calidad de las lecturas, ya que cumple con todos los requisitos esenciales y reportes de información. Aunque **FastQC** presenta una interfaz algo desactualizada en comparación con otras alternativas, cuenta con características que facilitan el análisis de los resultados, mostrando indicadores que permiten evaluar rápidamente diversas características de los parámetros de las lecturas (Figura 25).

Tabla 8: Cumplimiento de requisitos de las distintas herramientas.

Herramienta	R1	R2	R3	R4	R5
FastQC	X	X	X	X	X
Fastp	X		X	X	X
NanoPlot	X	X			
NanoQC		X			
Seqkit		X	X		

Requisitos:

- R1. Distribución de calidad al interior de las lecturas.
- R2. Distribución del largo de las lecturas.
- R3. Porcentaje de GC (guanina o citosina).
- R4. Porcentaje de lecturas duplicadas.
- R5. Porcentaje de adaptadores presentes.

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per tile sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)

Figura 25: Indicadores de Calidad de FastQC.

4.3.2.2 Control de calidad post-mapeo

Como se mencionó al principio de esta sección, en esta etapa se busca evaluar la calidad del mapeo utilizando dos métricas principales: la **cobertura** (porcentaje de la región objetivo que se ha leído) y la **profundidad** (número de veces que se ha leído cada base de la región objetivo). Para lograrlo, se recolectaron ocho herramientas extraídas de la revisión sistemática, las cuales generan métricas que analizan distintos atributos de las lecturas mapeadas.

En la Tabla 9 se puede ver una breve descripción de cada una de las herramientas evaluadas en esta etapa. De igual manera, en el Anexo B (desde la Tabla 91 a la Tabla 98), se presenta una descripción más detallada de las características de estas herramientas para el análisis de la calidad del mapeo, junto con un enlace a su repositorio o sitio oficial.

Tabla 9: Herramientas para evaluar la calidad del mapeo.

Herramienta	Descripción
Bedtools	Herramienta para realizar operaciones sobre archivos BED ³¹ .
Deeptools	Herramienta para realizar análisis de cobertura y profundidad de datos de secuenciación de alto rendimiento.
Mosdepth	Herramienta para calcular la profundidad para WGS ³² , exome o target sequence .
NanoPlot	Herramienta para graficar secuencias de lectura largas.
Picard	Herramienta para manipular archivos SAM , BAM , CRAM y VCF .
Qualimap	Herramienta que analiza los datos provenientes del mapeo, para tener una visión general de estos y facilitar la toma de decisiones en análisis posteriores.
Samtools	Herramienta para manipular archivos SAM , BAM y CRAM .
SeqKit	Herramienta para manipular archivos FASTA , FASTQ y BAM .

Una vez recopiladas las características de las diferentes herramientas para el análisis de la calidad de las lecturas, se procedió a ejecutar cada una. Esto permitió observar de manera más clara las ventajas y desventajas de cada herramienta, así como evaluar con mayor detalle las distintas características que ofrecen.

Entre las características más relevantes observadas se encuentra la capacidad de reportar el número y/o porcentaje de lecturas mapeadas (**Bedtools**, **Qualimap**, **Samtools**, **Seqkit** y **Picard**). Además, algunas herramientas permiten obtener un conteo más preciso de las lecturas mapeadas en una o varias regiones específicas (**Bedtools**, **Qualimap** y **Samtools**). Esta última funcionalidad es especialmente importante al trabajar con amplicones, como en este caso, ya que permite verificar si todas las regiones long-range han sido adecuadamente cubiertas.

En cuanto a la profundidad de lectura, **Bedtools**, **Mosdepth** y **Samtools** generan un informe detallado del número de veces que se leyó cada base en una región específica. Por otro lado, **Deeptools** proporciona únicamente los cuartiles relacionados con la distribución de la profundidad de lectura alcanzada en las diferentes bases.

Estas dos características se consideran fundamentales para determinar si se puede continuar con la ejecución del *pipeline*. Por lo tanto, la herramienta o herramientas

³¹ **BED (Browser Extensible Data)**: Formato de archivo, para almacenar y representar regiones genómicas de manera sencilla y flexible.

³² **WGS (Whole Genome Sequence)**: Técnica de secuenciación de ADN que permite obtener la secuencia completa del material genético de un organismo.

seleccionadas para esta etapa deben ser alguna de las mencionadas previamente (**Bedtools**, **Deeptools**, **Mosdepth**, **Qualimap**, **Picard**, **Samtools**, **Seqkit**).

Tanto **Seqkit** como **Picard** fueron descartadas ya que, a diferencia de **Bedtools**, **Qualimap** o **Samtools**, no permiten visualizar el número de lecturas mapeadas por región, una característica clave en este caso para confirmar que se han mapeado correctamente las nueve regiones *long-range*. De manera similar, **Deeptools** también fue descartada por ofrecer información más limitada en comparación con **Bedtools**, **Mosdepth** y **Samtools**.

Es importante señalar que las herramientas **Deeptools** y **Mosdepth** requieren archivos **BAM** indexados como entrada, lo que implica el uso adicional de herramientas como **Samtools** o **Picard** para realizar la indexación. Por lo tanto, dado que herramientas como **Bedtools** y **Samtools** pueden proporcionar ambas métricas sin necesidad de procesamiento adicional, no parece conveniente optar por **Deeptools**, **Mosdepth** ni, de manera similar, por **Qualimap**.

Finalmente, las herramientas seleccionadas para esta etapa son **Bedtools** y **Samtools**. **Bedtools** proporciona información sobre la **cobertura y la profundidad promedio** de las lecturas mapeadas en un formato de salida fácilmente procesable de manera automatizada. Por su parte, **Samtools** ofrece **información tanto sobre las lecturas no mapeadas como sobre la profundidad de lectura por base**, y, al igual que **Bedtools**, permite una automatización sencilla.

Por otro lado, **hay herramientas que presentan características interesantes para el análisis de las lecturas mapeadas**, aunque no sean necesariamente útiles para este *pipeline* en particular. Sin embargo, vale la pena mencionarlas, ya que podrían ser útiles en etapas posteriores del desarrollo de este pipeline y podrían tener un valor añadido para ciertos objetivos específicos.

Una de estas características es el número de inserciones, delecciones y mismatches encontrados (**Qualimap** y **Samtools**). Estas métricas pueden ser útiles al comparar los mapeos generados por distintas herramientas, ya que, por ejemplo, un mayor número de inserciones y delecciones podría indicar que la herramienta es más flexible al intentar mapear las lecturas.

Otra característica interesante es el número de lecturas recortadas (**Qualimap**) por las herramientas de alineamiento. Recortar demasiadas lecturas podría hacer que se pierdan las regiones con mayores diferencias respecto al genoma de referencia, conservando únicamente aquellas con más similitudes y, por ende, reduciendo la detección de posibles variantes.

Finalmente, **NanoPlot** ofrece una característica similar al mostrar la distribución del tamaño de las lecturas frente al tamaño de las lecturas mapeadas (con posibles recortes) (Figura 26). Esto permite analizar el nivel de recorte aplicado por las herramientas de mapeo a las lecturas.

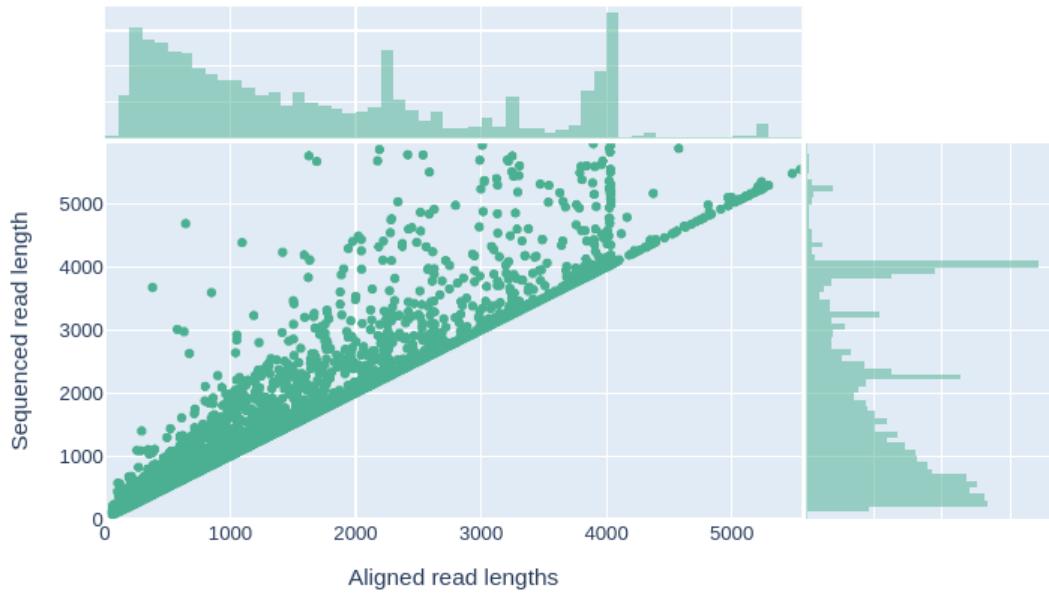


Figura 26: Largo de las lecturas alineados versus largo de las lecturas de las secuencias.

Estas características, aunque son información interesante para analizar, no resultan muy útiles para este tipo de *pipeline*, ya que el objetivo principal es verificar la cobertura de las regiones de interés y la profundidad alcanzada para poder continuar con la etapa de *variant calling*. Sin embargo, es valioso tener en cuenta estas herramientas para la construcción del *pipeline*, ya que permiten comparar las herramientas de mapeo, analizar los datos generados por el secuenciador en uso, y evaluar visualmente los resultados del mapeo en ciertas regiones. Con esta información, se pueden realizar mejoras en la cobertura y profundidad en futuros análisis.

4.3.3 Trimming y filtrado de lecturas

Considerando lo expuesto en el punto 2.3.4.3, es crucial que la herramienta seleccionada para esta etapa permita descartar secuencias que no sean relevantes para el análisis o que carezcan de la calidad necesaria para confiar en su información. Además, debe permitir podar aquellos fragmentos de las lecturas que no cumplen con la calidad deseada, conservando así las mejores partes. Por lo tanto, los requisitos necesarios para la selección de esta herramienta son:

1. **Filtrado de lecturas por calidad mínima:** Se eliminan lecturas que no alcanzan un umbral mínimo de calidad promedio en sus bases.
2. **Filtrado de lecturas por largo mínimo y máximo:** Se descartan lecturas que sean más cortas o más largas que un tamaño especificado.
3. **Poda (*Trimming*) de lecturas en base a calidad mínima:** Se eliminan bases desde los extremos de las lecturas que no cumplen con un umbral de calidad.

Tras revisar las herramientas identificadas en la revisión sistemática, se encontraron un total de diez herramientas pertenecientes a dicha etapa, además de una adicional, que también permiten filtrar y/o podar lecturas basándose en características como la calidad y la longitud. Estas herramientas son **BBDuk**, **Chopper**, **Cutadapt**, **DeepTools**, **Fastp**, **Filtlong**, **Picard**, **Porechop**, **Samtools**, **Seqkit** y **Trimmomatic**.

Al igual que en el punto 4.3.2, la información sobre las herramientas, incluidos los archivos de entrada, archivos de salida y sus características, está descrita desde la Tabla 99 a la Tabla 109 en el Anexo C.

Estas herramientas pueden ser clasificadas en dos grupos. En el primero se encuentran las que aceptan archivos **FASTA** o **FASTQ** y permiten podar o filtrar lecturas de estos archivos. En el segundo están las herramientas que reciben como entrada archivos con lecturas mapeadas, en formato **SAM** o **BAM**, que solo permiten filtrar lecturas, ya sea en base a sus características o si corresponden a lecturas duplicadas. En este último grupo, solo se encuentran tres herramientas, de las cuales **DeepTools**, **Picard** y **Samtools** presentan las mismas características. Sin embargo, **Samtools** es una herramienta especialmente diseñada para la manipulación de archivos **SAM**, **BAM** y **CRAM**, por lo que esta debería ser la herramienta elegida si se llegara a necesitar filtrar lecturas ya mapeadas.

Volviendo al primer grupo, este está compuesto por las ocho herramientas restantes, que comparten varias características entre sí. De todas ellas, destaca **Porechop**, la cual está principalmente enfocada en la demultiplexación, el filtrado y poda de adaptadores. No obstante, como se mencionó en el punto 4.3.1, la herramienta **Dorado** presenta la opción de demultiplexar las muestras y además elimina los adaptadores y *barcodes* presentes en estas, por lo que el uso de **Porechop**, para este *pipeline*, no se ve justificado.

Siguiendo con las herramientas restantes, al evaluar los requisitos expresados al principio de la sección, se observa que todas cumplen con el requisito de filtrar las lecturas que no cumplan con un largo mínimo o máximo. Sin embargo, solo dos herramientas (**Cutadapt** y **Trimmomatic**) no tienen la opción de filtrar lecturas en base a una calidad promedio mínima. Por lo tanto, debido a la importancia de este requerimiento, dichas herramientas quedan descartadas.

Por otro lado, en cuanto a la poda de lecturas, hay características compartidas por varias herramientas, como podar una cierta cantidad de bases al inicio o final de una lectura. Esto puede ser importante para datos provenientes de secuenciadores que tienen un patrón de calidad, como los de Illumina, en los que la calidad desciende al final de las lecturas. No obstante, esto no parece aplicarse en el caso del **MinION**.

En cambio, existen características que, aunque la base es la misma en varias herramientas, presentan ciertos cambios a la hora de llevar a cabo estas tareas, como la evaluación de las lecturas mediante ventana deslizante, que es realizada por las herramientas **Fastp**, **Filtlong** y **Trimmomatic**. Esta característica consiste en recorrer las lecturas una posición a la vez y evaluar la calidad promedio de un cierto número de bases a partir de esa posición. En caso de no cumplir con una calidad mínima, la lectura es podada en el caso de **Fastp** y **Trimmomatic**, o filtrada en el caso de **Filtlong** (Figura 27). Otra diferencia es que la herramienta **Fastp** permite ejecutar varias ventanas deslizantes, recortando las lecturas en ambas direcciones, mientras que **Trimmomatic** solo permite ejecutar la ventana en dirección de 3' a 5'. Esto supone que **Fastp** puede ser más flexible y efectivo a la hora de preservar las partes de mayor calidad de las lecturas.

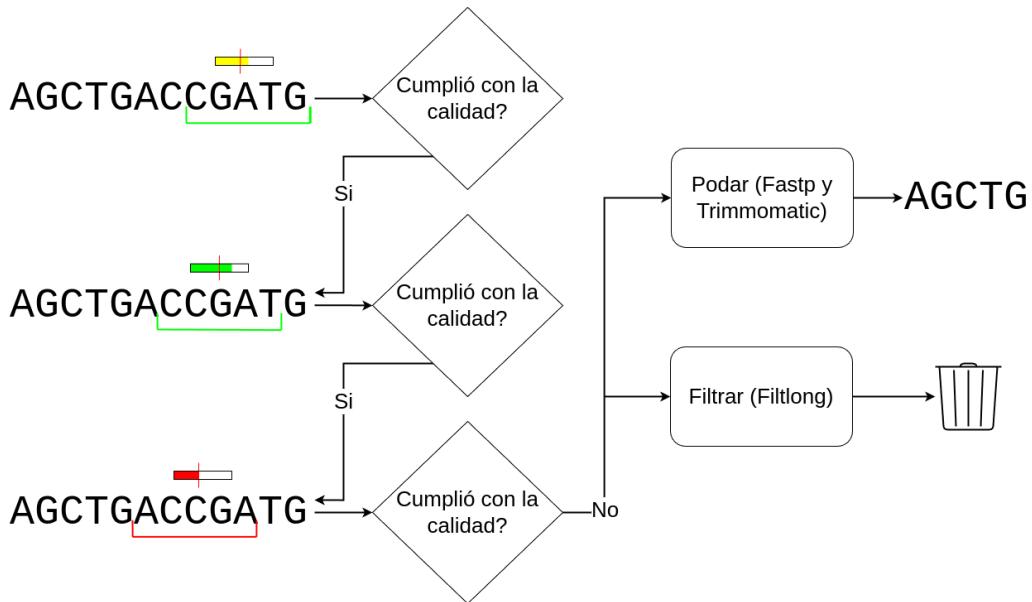


Figura 27: Ejemplo funcionamiento ventana deslizante desde 3' a 5'.

Otra característica práctica es la presentada por **Cutadapt**, que poda los extremos de las lecturas si no cumplen con una calidad mínima. Esto se logra mediante un algoritmo que, a diferencia de otros como Trimmomatic, no detiene la poda cuando encuentra una base que cumple con la calidad mínima si anteriormente había muchas bases con mala calidad. Esto implementa una poda más agresiva desde los extremos. Esto produce lecturas más cortas, pero de mejor calidad, mediante la eliminación de una mayor cantidad de bases de menor calidad a costa de algunas con buena calidad. Esta característica la hace una herramienta a tener en cuenta para la selección final.

Finalmente, la herramienta **Fastp** presenta una opción para filtrar las lecturas con un cierto porcentaje de bases no cualificadas, entendidas como aquellas que no superan un umbral de calidad. Esta opción permitiría conservar lecturas en las que la mayoría de las bases son de buena calidad, aún cuando entre ellas se encuentren bases de calidad muy baja que hacen que el promedio de calidad disminuya debajo de un umbral y sea podada o filtrada por otros algoritmos.

En base a todo lo mencionado anteriormente y a las características de las herramientas recopiladas en el Anexo C, aunque varias de ellas cumplen con los requisitos básicos establecidos al inicio de esta sección, y algunas, como **Cutadapt**, ofrecen características interesantes para la poda de regiones de baja calidad, se considera que **Fastp** destaca por la amplia gama de funcionalidades que ofrece. Esta herramienta permite filtrar y podar lecturas de manera eficiente, facilitando la preservación de la mayoría de las lecturas y asegurando que se conserven los fragmentos de mayor calidad.

4.3.4 Mapeo

Durante la revisión sistemática se hallaron un total de 27 herramientas utilizadas para el mapeo de lecturas con un genoma de referencia. Sin embargo, de estas 28 herramientas, cinco no cumplen con esta tarea. En particular, en el caso de **Samtools**, esta herramienta se utiliza para obtener diferentes estadísticas o manipular los

archivos obtenidos de la etapa de mapeo. Por otro lado, **Duplomap** permite realinear lecturas ya mapeadas, mejorando la precisión de las lecturas largas alineadas en regiones duplicadas. Finalmente, las herramientas **Last** y **MashMap** se utilizan para identificar regiones similares entre secuencias largas. Por lo tanto estas herramientas no serán consideradas en esta etapa.

Por otro lado, se descartaron otras siete herramientas por diferentes motivos, como no contar con sitio web al cual acceder para descargar la herramienta, o que no han recibido soporte en más de 10 años. Estas herramientas incluyen **SHRIMP**, **MAQ**, **Stampy**, **ELAND**, **SARUMAN**, **SOAP2** y **SOAP3**. Finalmente, se descartaron ocho herramientas adicionales, por diversos motivos detallados en la Tabla 10

Tabla 10: Motivo de exclusión de herramientas de mapeo.

Herramientas	Motivos de exclusión
Bowtie	Versión anterior de Bowtie2 .
MAFFT	Herramienta para alineamiento múltiple.
mrFAST	Requiere que todas las lecturas sean del mismo largo.
MUMmer	No genera archivo de salida en formato SAM o BAM .
NovoAlign	Herramienta comercial.
Pbmm2	Adaptación de la herramienta Minimap2 para lecturas provenientes del secuenciador PacBio.
STAR	Herramienta para mapear secuencias de ARN.
SEAL	Utiliza otras herramientas ya incorporadas para mapear.

Las herramientas que finalmente fueron seleccionadas para evaluación son: **Bowtie2**, **BWA**, **GraphMap**, **Minimap2**, **NGMLR**, **LRA**, **Vulcan** y **Winnowmap**. Un total de 6 herramientas están diseñadas para mapear lecturas largas, exceptuando por **BWA**. **BWA** es una de las herramientas más populares de mapeo, ocupada principalmente para mapear secuencias de algunos cientos de pares de bases, como lo son las lecturas generadas por los secuenciadores de **Illumina**. Por otro lado, la herramienta **Bowtie2** no está desarrollada principalmente para mapear lecturas largas, pero puede llegar a soportar lecturas de unos miles de pares de bases.

Para realizar la comparación entre las herramientas, se mapearon las lecturas obtenidas de las muestras de ocho pacientes contra el genoma de referencia **GRCh38/hg38**, que incluye la información genética de los 23 cromosomas del genoma humano. No obstante, antes de proceder con el mapeo, se filtraron las lecturas con una longitud superior a 6000 pb, ya que, dado su tamaño, probablemente correspondían a moléculas de **ADN** residual de regiones no relacionadas con las regiones objetivos que se observan en la Figura 22.

Por otro lado, debido a los errores inherentes en las lecturas y las diferencias naturales entre el genoma de los pacientes y el genoma de referencia, no es posible afirmar que un mapeo sea más correcto que otro. Siempre se busca la explicación más simple en el proceso de mapeo, sin embargo, esto **no excluye la posibilidad de obtener diferentes mapeos** que, pese a su diversidad, pueden compartir un nivel de complejidad similar (como se muestra en la Figura 28). Por lo tanto, para realizar una comparación y evaluación adecuada de las herramientas, se optó por seleccionar dos herramientas que destacaron en diferentes características. Los mapeos resultantes se emplearon en las etapas posteriores, con el fin de evaluar el impacto de dichas características.

Read: GCTGCACGCGATAATACCAGTGGAC

Reference: GCTCCATGCGGTAAATACCAGTAGGC
 ||| ||| ||| ||||||||| | |

Map 1: GCTGCACGCGATAATACCAGTGGAC

Reference: GCTCCATGCGGTAAATACCAGTAGG-C
 ||| ||| ||| ||||||||| ||| |

Map 2: GCTGCACGCGATAATACCAGT-GGAC

Figura 28: Ejemplo de ambigüedad en el mapeo³³.

La característica más importante que se consideró para la selección de herramientas fue el tiempo de ejecución. Tal como señalan Lunter y Goodson (2010), las **herramientas de mapeo tienden a ser rápidas o a ser sensibles, pero rara vez coexisten ambas** características. Esta observación es reafirmada por Fu, Mahmoud, Muraliraman, Sedlazeck y Treangen (2021), quienes mencionan que: “Para maximizar el potencial de la secuenciación de lecturas largas en este contexto, **han surgido nuevos métodos de mapeo que se han enfocado principalmente en la velocidad o en la precisión**”.

Una menor sensibilidad o precisión a la hora de mapear las lecturas conduce a sesgos de mapeo no deseados, particularmente para lecturas de regiones con mayor divergencia y para lecturas que contienen indels (Lunter & Goodson, 2010). Lunter y Goodson (2010) también señalan que: “En cualquier experimento, una fracción de las lecturas presentará tasas de error elevadas, y poder incluir de manera confiable los datos de estas lecturas mejora la potencia de los análisis posteriores y reduce el costo total de la secuenciación”. Esto adquiere especial relevancia considerando que los secuenciadores de nanoporos, conocidos por sus altas tasas de error, son posiblemente los secuenciadores con mayores desafíos en este aspecto.

De esta forma, se optó por seleccionar una herramienta con tiempos de ejecución más largos y otra con tiempos de ejecución más cortos. El objetivo es poder en

³³ Ejemplo adaptado de :<https://github.com/BenLangmead/bowtie2>

etapas posteriores, determinar si invertir más tiempo en el mapeo (lo que implicaría una mayor sensibilidad), permite obtener mejores resultados del *variant calling*.

Tras ejecutar las herramientas en cada una de las muestras, se registraron los tiempos de ejecución, los cuales se presentan en la Tabla 11. Se observa una variación significativa entre los tiempos de ejecución de las distintas herramientas. **Minimap2** destaca por ser la más rápida, mientras que **GraphMap** muestra los tiempos más prolongados, **aproximadamente 160 veces mayor que Minimap2**.

Tabla 11: Tiempo de ejecución (en minutos) de las herramientas de mapeo para cada una de las muestras.

Herramienta	S1	S2	S3	S4	S5	S6	S7	S8
Bowtie2	709	926	1022	1509	919	1400	722	900
BWA	58	75	106	116	90	114	74	99
GraphMap	1145	1522	1932	1741	1288	1305	1793	1564
LRA	13	16	24	24	18	24	14	15
Minimap2	7	8	12	11	9	11	9	9
NGMLR	41	48	62	72	54	70	59	57
Vulcan	22	20	26	28	21	27	23	22
Winnowmap	7	9	13	12	10	12	11	11

Es importante destacar que, para **el cálculo de los tiempos, no se incluyó el tiempo de indexación**, ya que este proceso se realiza una única vez por cada genoma de referencia utilizado. Además, como se mencionó anteriormente, la selección del tiempo como métrica de comparación se hizo principalmente para evaluar la sensibilidad de las herramientas; por lo tanto, el tiempo de indexación no aporta información relevante para este análisis.

Después de ejecutar y comparar los tiempos de ejecución de cada herramienta, se calcularon diversas métricas que permitieron caracterizar de manera más precisa los resultados obtenidos. Estas métricas fueron generadas utilizando herramientas previamente mencionadas, como **Samtools**, **Bedtools** y **Qualimap**.

Las métricas utilizadas para analizar el número de lecturas mapeadas incluyeron el **porcentaje promedio de lecturas mapeadas** por las herramientas de forma general, así como el **porcentaje de lecturas mapeadas específicamente en las regiones objetivo**. Finalmente, se consideró el **porcentaje promedio de las lecturas mapeadas conservadas luego de filtrar por las regiones objetivo**.

Los resultados obtenidos se pueden apreciar en la Tabla 12. En ella no se observa un patrón claro entre el porcentaje de lecturas mapeadas en todo el genoma y el tiempo de ejecución de las herramientas. Sin embargo, al analizar el porcentaje promedio de

lecturas mapeadas en las regiones objetivo, se evidencia una relación entre el tiempo de ejecución y el porcentaje de lecturas mapeadas en las regiones objetivo: tres de las cuatro herramientas con tiempos de ejecución más largos mapearon un menor promedio de lecturas en comparación con las herramientas más rápidas. A su vez, tres de las cuatro herramientas más rápidas lograron un porcentaje promedio mayor de lecturas mapeadas en las regiones objetivo.

Tabla 12: Porcentaje promedio de lecturas mapeadas.

Herramienta	Porcentaje promedio de lecturas mapeadas	Porcentaje promedio de lecturas mapeadas (Situadas en la regiones objetivo)	Porcentaje promedio de las lecturas mapeadas conservadas después del filtrado por regiones objetivo
Bowtie2	65,5%	37,7%	57%
BWA	92,5%	46,5%	50,8%
GraphMap	88,8%	36,9%	42,1%
LRA	72,7%	38,7%	53%
Minimap2	85,7%	44,4%	52,4%
NGMLR	78%	37,4%	48,9%
Vulcan	96,9%	49,7%	51,3%
Winnowmap	82,7%	42,2%	51,7%

Al considerar el porcentaje de lecturas mapeadas conservadas después de filtrar por las regiones objetivo (columna tres), se observa que **Bowtie2** retiene un mayor porcentaje de estas lecturas en comparación con las demás herramientas, mientras que **GraphMap** retiene el menor porcentaje. Esto podría indicar una mayor sensibilidad y precisión por parte de **Bowtie2** en el mapeo de lecturas.

Otra métrica obtenida a partir de las muestras mapeadas fue el porcentaje de inserciones o delecciones (**INDELS**) y el porcentaje de discrepancias entre el genoma de referencia y las lecturas alineadas (**mismatch**) (Tabla 13). En estas métricas, destaca notablemente el resultado obtenido por **GraphMap**, que presenta los valores más altos. Este resultado contrasta con lo que se esperaría de una herramienta con alta sensibilidad, que debería mapear las lecturas de manera que se ajusten mejor al genoma de referencia y, por lo tanto, contengan menos errores. Este comportamiento se observa, por ejemplo, en **Bowtie2**, que muestra un menor porcentaje de **INDELS**.

Por último, se evaluó el porcentaje de lecturas recortadas por las herramientas de mapeo (Tabla 14). Los resultados más destacados son los obtenidos por la herramienta **LRA**, que generó recortes en casi todas las lecturas alineadas, y por otro

lado, **Bowtie2**, que no recortó ninguna de las lecturas alineadas. Además, el resultado obtenido por **Bowtie2** se diferencia enormemente del de las demás herramientas, siendo **BWA** la que le sigue con un 56% de las lecturas recortadas. El resultado de **Bowtie2** podría interpretarse como un efecto de su alta sensibilidad al buscar el mejor ajuste para las lecturas.

Tabla 13: Porcentaje de INDELs y *mismatches* en relación con las bases mapeadas.

Herramienta	Porcentaje de INDELs	Porcentaje de <i>mismatches</i>
Bowtie2	1,06%	5,8%
BWA	1,36%	4,96%
GraphMap	3,37%	7,6%
LRA	2,3%	4,27%
Minimap2	1,83%	5,5%
NGMLR	2,23%	4,92%
Vulcan	1,88%	5,11%
Winnowmap	1,8%	5,36%

Tabla 14: Porcentaje promedio de lecturas recortadas durante el mapeo.

Herramienta	Porcentaje promedio de lecturas recortados
Bowtie2	0%
BWA	56%
GraphMap	61,7%
LRA	99,7%
Minimap2	79,5%
NGMLR	94,2%
Vulcan	80,1%
Winnowmap	79,2%

Con base en lo planteado al inicio de la sección y en las métricas analizadas, en las siguientes etapas se utilizaran las muestras mapeadas con la herramienta **Minimap2**, con el objetivo de comprobar si el mapeo realizado por una herramienta de bajo tiempo de ejecución ofrece la calidad necesaria para la detección de variantes genéticas.

Por otro lado, aunque inicialmente se había considerado seleccionar también la herramienta con el tiempo de ejecución más largo, el análisis de los resultados obtenidos por **Bowtie2** y **GraphMap** llevó a la conclusión de que **Bowtie2** ofrece resultados más consistentes con los esperados de una herramienta más sensible y precisa. Por lo tanto, **Bowtie2** será la segunda herramienta utilizada en las etapas posteriores, a pesar de que **GraphMap** haya presentado tiempos de ejecución más elevados.

4.3.5 Clasificación, anotación y filtrado de variantes

Antes de comparar y evaluar las herramientas de *variant calling*, es necesario definir las herramientas para la clasificación y anotación de variantes, así como el sistema de *ranking* de variantes (sección siguiente), lo que permitirá ejecutar el *pipeline* de principio a fin.

Conforme se indicó en el punto 2.3.3.6, esta fase tiene como objetivo clasificar una variante dentro de las cinco categorías establecidas por **ACMG/AMP**. Sin embargo, **para este pipeline, se decidió dividir esta etapa en dos fases**. En la primera fase, se busca recolectar evidencia que pueda contribuir a la clasificación de la variante. Posteriormente, en la segunda fase, detallada en el punto 4.3.7, se procede a puntuar las variantes basándose en la evidencia recopilada.

Los *software* empleados en esta primera fase pueden clasificarse en tres tipos: “**herramientas de anotación y filtrado de variantes**”, “**herramientas de predicción de patogenicidad**” y “**bases de datos**”. A continuación, se detalla el propósito de cada uno de los tipos mencionados, así como los requisitos para la selección de las herramientas.

Anotación y filtrado de variantes: Los programas de esta categoría permiten la anotación en el archivo **VCF** de la información relacionada con la clasificación de la variante, así como el filtrado de aquellas que no cumplan con los parámetros de calidad o profundidad establecidos.

Herramientas de predicción de patogenicidad: En esta categoría se incluyen todas aquellas herramientas que buscan predecir, mediante diversos enfoques, los efectos que tendrán las variantes cuando el gen sea traducido a proteína. Estos *software* pueden proporcionar una clasificación siguiendo los estándares propuestos por la **ACMG/AMP** o pueden tener una clasificación propia. Sin embargo, como se menciona en la guía de la **ACMG/AMP**, descrita anteriormente, no se debe sobreestimar la evidencia computacional, ya que muchos de estas herramientas podrían estar utilizando los mismos algoritmos o las herramientas podrían no haber sido validadas frente a variantes patogénicas bien establecidas (Richards et al, 2015).

Bases de datos: Existen diferentes tipos de bases de datos de variantes. Por ejemplo, algunas almacenan la frecuencia de las variantes en la población general, mientras que otras contienen la clasificación de variantes previamente identificadas y verificadas. También existen bases de datos de la misma organización que pueden contener variantes de pacientes anteriores o incluso de familiares del paciente. Esta información, que ha sido recopilada y verificada previamente, tiene un gran valor.

Por lo tanto, **es esencial contar con un conjunto de bases de datos que ofrezca una amplia cantidad y diversidad de datos**, facilitando la toma de decisiones con respecto a una variante.

Para esta etapa, **se decidió mantener las herramientas y bases de datos utilizadas en el sistema de ranking** de variantes descrito en el punto 4.3.7. Esta elección se basa en la consideración de que el conjunto seleccionado abarca de manera adecuada los criterios necesarios para clasificar las variantes genéticas identificadas.

A continuación se presentan las herramientas utilizadas para cada uno de los tres propósitos mencionados previamente, comenzando con las herramientas de anotación y filtrado de variantes. Para esta tarea, se empleará **ANNOVAR**, una herramienta que permite anotar las funciones y evidencias asociadas a las variantes. Esto se logra mediante la integración de un amplio catálogo de bases de datos y herramientas de predicción de patogenicidad.

De este modo, los *software* de predicción de patogenicidad que se utilizarán están integrados en la base de datos **dbNSFP**, la cual puede ser descargada desde **ANNOVAR**. Esta base de datos **incluye un total de 21 herramientas para la predicción funcional y la anotación de variantes** no sinónimas y sitios de corte y empalme en el genoma humano. Las 21 herramientas se pueden consultar en la Tabla 15.

Tabla 15: Herramientas de predicción de patogenicidad.

Herramienta	Descripción
CADD	Califica la patogenicidad de las variantes de un solo nucleótido, sustituciones de múltiples nucleótidos e INDELs en el genoma humano.
DANN	Califica la patogenicidad de SNPs , MNPs e INDELs , a través de redes neuronales.
Eigen	Herramienta que integra anotaciones genómicas funcionales para variantes codificantes y no codificantes.
FATHMM	Predice las consecuencias funcionales de las variantes codificantes, variantes de un solo nucleótido <i>no synonymous</i> y las variantes no codificantes.
fathmm-MKL	Predice las consecuencias funcionales de SNPs codificantes y no codificantes.
FitCons	Evalúa las consecuencias funcionales de variantes genómicas
GenoCanyon	Predice el efecto funcional de variantes genómicas.

Herramienta	Descripción
GERP++	Identifica elementos restringidos ³⁴ , a través de un enfoque de "genómica comparativa" (Davydov et al., 2010).
LRT	Evalúa la evidencia de que una variante genética afecta la función de una proteína mediante métodos estadísticos.
MetaLR	Integra nueve puntuaciones de patogenicidad de variantes independientes y la información de frecuencia alélica para predecir la patogenicidad de variantes <i>missense</i> .
MetaSVM	Predice el impacto funcional de las sustituciones de aminoácidos en las proteínas.
MutationAssessor	Predice el impacto funcional de las sustituciones de aminoácidos en las proteínas.
MutationTaster	Predice el efecto funcional de variantes genéticas.
M-CAP	Clasifica la patogenicidad clínica.
SIFT	Predice el impacto funcional de las sustituciones de aminoácidos en las proteínas.
SiPhy	Analiza alineamientos de secuencias múltiples y destaca bases o pequeñas regiones que están bajo selección al observar la reducción en las tasas de sustitución y detectar patrones de sustitución inesperados (Siphy, 2010).
PhyloP	Calcula p-values de conservación o aceleración basados en un alineamiento.
PolyPhen2 HDIV	Diagnóstica enfermedades mendelianas.
PolyPhen2 HVAR	Evalúa alelos raros en loci potencialmente involucrados en fenotipos complejos.
PROVEAN	Predice si un SNP o un INDEL tiene un impacto en la función biológica de una proteína.
VEST	Clasifica y prioriza variantes missense raras con probable implicación en enfermedades humanas basado en aprendizaje supervisado (Carter et al., 2013).

Por otro lado, se utilizaron un total de ocho bases de datos (Tabla 16), de las cuales nueve están disponibles para descarga a través de **ANNOVAR**, con la excepción de la base de datos de la **Clínica Mayo**, a la cual se puede acceder únicamente a través

³⁴ **Elementos restringidos:** Regiones del genoma que han sido altamente conservadas a lo largo de la evolución debido a la selección natural.

de un navegador. Este conjunto de bases de datos permite abarcar todas las características necesarias de una variante, desde su ubicación en el genoma y su frecuencia en la población mundial, hasta la existencia o ausencia de evidencia sobre su patogenicidad.

Tabla 16: Bases de datos de variantes utilizadas.

Base de datos	Descripción
Abraom	Base de datos de frecuencia de variantes genéticas en poblaciones.
Esp	Base de datos de frecuencia de variantes genéticas en poblaciones.
dbSNP	Base de datos que contiene información sobre SNP , microsatélites e inserciones y delecciones a pequeña escala, junto con información de publicaciones, frecuencia en la población, consecuencia molecular y mapeo genómico y RefSeq , tanto para variaciones comunes como para mutaciones clínicas
Clínica Mayo	Base de datos de variantes genéticas asociadas a ADPKD .
ClinVar	Base de datos de variantes germinales y somáticas asociadas a distintos fenotipos.
GnomAD	Base de datos de frecuencia de variantes genéticas en poblaciones.
RefGene	Específica genes codificantes y no codificantes de proteínas humanas conocidas.
1000G	Base de datos de frecuencia de variantes genéticas en poblaciones.

Finalmente, para realizar una síntesis de los resultados basada en toda la evidencia proporcionada por las herramientas y bases de datos integradas en **ANNOVAR**, se utilizará **InterVar**. Esta herramienta, que se puede incorporar a **ANNOVAR**, permite clasificar las variantes de tipo missense en los cinco grupos propuestos por la **ACMG/AMP** mediante la ejecución automática de la guía descrita anteriormente.

Con este conjunto de herramientas, se genera un archivo que compila toda la evidencia y las predicciones para cada variante, información que es esencial para la siguiente etapa, la cual tiene como objetivo identificar una variante como candidata responsable de la enfermedad.

4.3.6 Sistema de ranking de variantes

El sistema de clasificación de variantes es el método empleado por el equipo de **GEMINI** para **evaluar y priorizar las variantes según su potencial patogénico**.

Como se mencionó en la sección anterior, esta etapa es parte del proceso de **clasificación, anotación y filtrado de variantes**. Este sistema fue diseñado originalmente para clasificar variantes de muestras generadas por el secuenciador **Illumina MiSeq** y fue desarrollado por el bioinformático Cristian Yañez en colaboración con Paola Krall.

Las variantes que se busca clasificar en esta sección son aquellas identificadas previamente por los programas de *variant calling* (siguiente sección) y que cuentan con anotaciones generadas a partir de *software* de predicción de patogenicidad y/o bases de datos de variantes (punto 4.3.5).

El método de evaluación de variantes se fundamenta en diversos factores, como la clasificación de las variantes en distintas bases de datos, las predicciones de patogenicidad proporcionadas por programas especializados, el tipo de variante y su ubicación en el gen. Además, se considera la clasificación final de la variante tras aplicar la guía propuesta por la **ACMG/AMP** mediante el *software* **InterVar**.

Este sistema de *ranking* asigna puntaje a las variantes con un algoritmo personalizado. El cual califica las variantes con un puntaje que va de 0 a 100 puntos, donde un puntaje de 100 es considerado como una variante patogénica, mientras que un puntaje igual a cero se interpreta como una variante benigna.

Antes de proceder a la clasificación de las variantes, se realiza un filtrado inicial. Se excluyen aquellas **variantes con una frecuencia alélica en la población igual o superior al 5%**. Además, se eliminan las **variantes que tienen una cobertura de lectura menor o igual a 30**, así como aquellas con una **frecuencia alélica de la variante (VAF) menor o igual a 0.09 o mayor a 0.8**. Estos criterios se aplican para filtrar posibles errores de secuenciación.

En el caso de enfermedades de perfil dominante, se espera que la **VAF** sea cercana al 50%. Sin embargo, este valor puede variar significativamente debido a factores en la preparación de las muestras y el proceso de secuenciación. Por ello, suelen aplicarse márgenes más amplios, ajustados según la experiencia del especialista y la tecnología de secuenciación utilizada. En este *pipeline*, **los valores de filtrado como VAF y cobertura de lectura, se ajustarán con base en los resultados obtenidos** tras el análisis de las muestras.

Finalmente, el sistema de *ranking* propone los siguientes criterios de evaluación recopilados en la Tabla 17. Tras evaluar cada variante en cada uno de los criterios, se le asigna puntaje a la variante, siguiendo el siguiente diagrama de flujo presentado en la Figura 29.

Siguiendo estas instrucciones para cada una de las variantes encontradas, se obtiene una lista de las variantes, la cual puede ser ordenada según la probabilidad de ser responsables de la enfermedad, en base a la evidencia presentada. Sin embargo, solo se puede tener un alto grado de certeza de que son variantes patogénicas aquellas que han sido calificadas con 80 puntos o más (Códigos C1 a C4). Si no se obtiene ninguna variante que supere ese puntaje, entonces se deberá evaluar la posibilidad de realizar un nuevo análisis que incluya a los familiares afectados y sanos, con el fin de

conseguir más evidencia que pueda respaldar la patogenicidad de alguna de las variantes encontradas.

Tabla 17: Criterios para clasificación de variantes sistema de *ranking* Yáñez-Krall.

Código	Criterio
C1	RefGene clasifica la variante como “ <i>Stopgain</i> ”.
C2	Clinvar , Intervar o Clínica Mayo clasifican la variante como “ Patogénica ”.
C3	Mutation Taster la clasifica como “A” (“ <i>disease_causing_automatic</i> ”).
C4	RefGene la clasifica como “ <i>splicing</i> ”, “ <i>frameshift deletion</i> ”, “ <i>frameshift insertion</i> ”.
C5	Mutation Taster la clasificada como “D” (“ <i>disease_causing</i> ”).
C6	LRT la clasifica como “D” (<i>Deleterious</i>).
C7	Sift la clasifica como “D” (<i>Deleterious</i>).
C8	Variante clasificada como “ <i>frameshift deletion</i> ” o “ <i>frameshift insertion</i> ”.

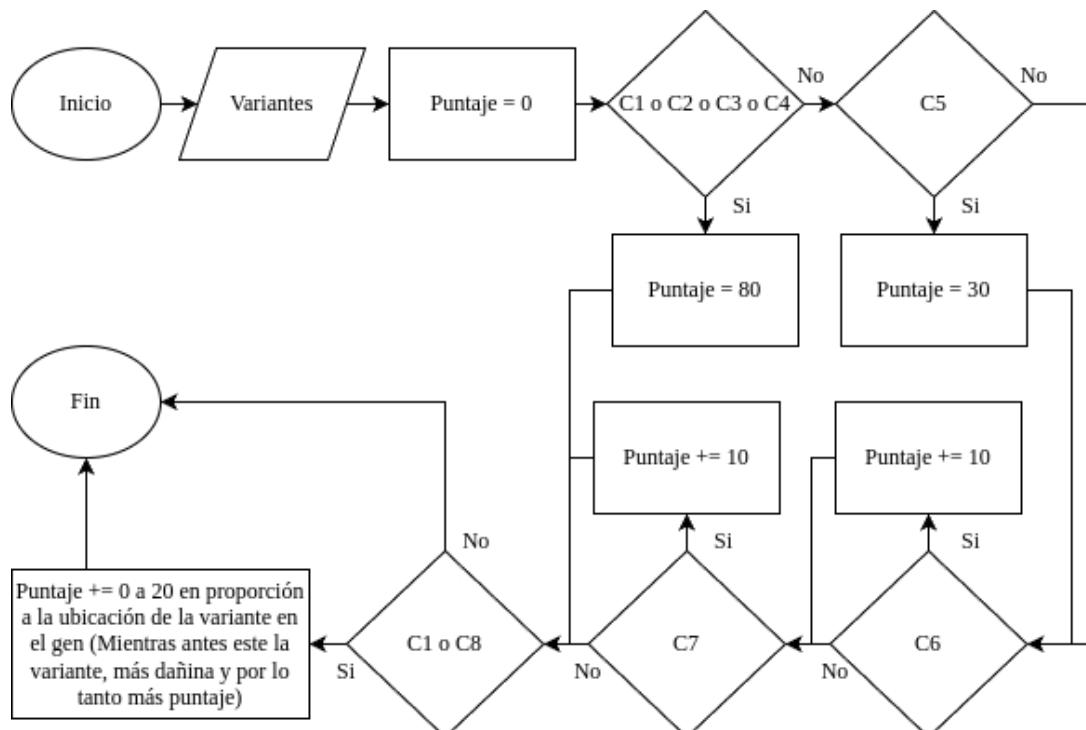


Figura 29: Diagrama de flujo para la calificación de variantes.

Para facilitar la evaluación de las herramientas y, posteriormente, la valoración de las variantes por parte de los miembros del equipo **GEMINI**, se automatizó la ejecución del sistema de *ranking* de variantes mediante un *script* en **Python**. Este *script* toma

como entrada uno de los archivos finales generados por **ANNOVAR**, que contiene toda la información de las variantes identificadas por las herramientas mencionadas en el apartado 4.3.6. Adicionalmente, se desarrolló un segundo *script* que extrae las variantes almacenadas en la base de datos de la **Clínica Mayo**, automatizando así la búsqueda y comparación de las variantes encontradas con las registradas en dicha base de datos.

Finalmente, el script de calificación de variantes incorpora la información de la **Clínica Mayo** para las variantes y las evalúa de acuerdo con el diagrama mostrado en la Figura 29. El proceso concluye entregando una lista de variantes con su respectivo puntaje y la información utilizada para la calificación. Estas variantes se presentan en un archivo en formato **TSV**, que puede visualizarse fácilmente en **Excel**. Las variantes están ordenadas según varios criterios: si cumplen con la frecuencia alélica mínima requerida, si cumplen con la frecuencia alélica de variante establecida, si satisfacen la profundidad de lectura requerida, su nivel de patogenicidad y su posición en el genoma.

4.3.7 Variant calling

La etapa de identificación de variantes, conocida como ***variant calling***, es una de las fases más críticas y variables en términos de resultados, ya que las herramientas emplean distintos enfoques y heurísticas para diferenciar entre errores de secuenciación y variantes genéticas reales. Además, es fundamental que las variantes causales se identifiquen claramente como las de mayor relevancia, una vez aplicado el sistema de *ranking* descrito en la sección anterior.

Para comparar y evaluar las herramientas de *variant calling*, fue necesario definir previamente las herramientas de las etapas previas y posteriores del *pipeline*. Esto permite evaluar no solo si las herramientas de ***variant calling*** identifican correctamente las variantes (anotándolas y clasificándolas con **ANNOVAR**), sino también si el **sistema de ranking de variantes** destaca como causales las variantes previamente detectadas en los casos clínicos presentados en el punto 4.1 (ver Figura 30).

En esta etapa, se utilizarán las muestras mapeadas con las dos herramientas seleccionadas previamente (**Bowtie2** y **Minimap2**). Esto permitirá evaluar cuál de estos mapeos proporciona los mejores resultados para optimizar el rendimiento de las herramientas de *variant calling*. Por lo tanto, los resultados obtenidos en esta fase también determinarán en la elección final de la herramienta de mapeo.

Durante la revisión sistemática descrita en el punto 3, se hallaron un total de 41 herramientas relacionadas con dicha etapa, de las cuales se descartaron 28 por diversos motivos, entre ellos un soporte descontinuado, ser versiones anteriores de herramientas más modernas o la imposibilidad de detectar algún tipo específico de variante (ver Anexo D, Tabla 110 para mayores detalles). La Tabla 18 presenta las 12 herramientas seleccionadas junto a una breve descripción de ellas.

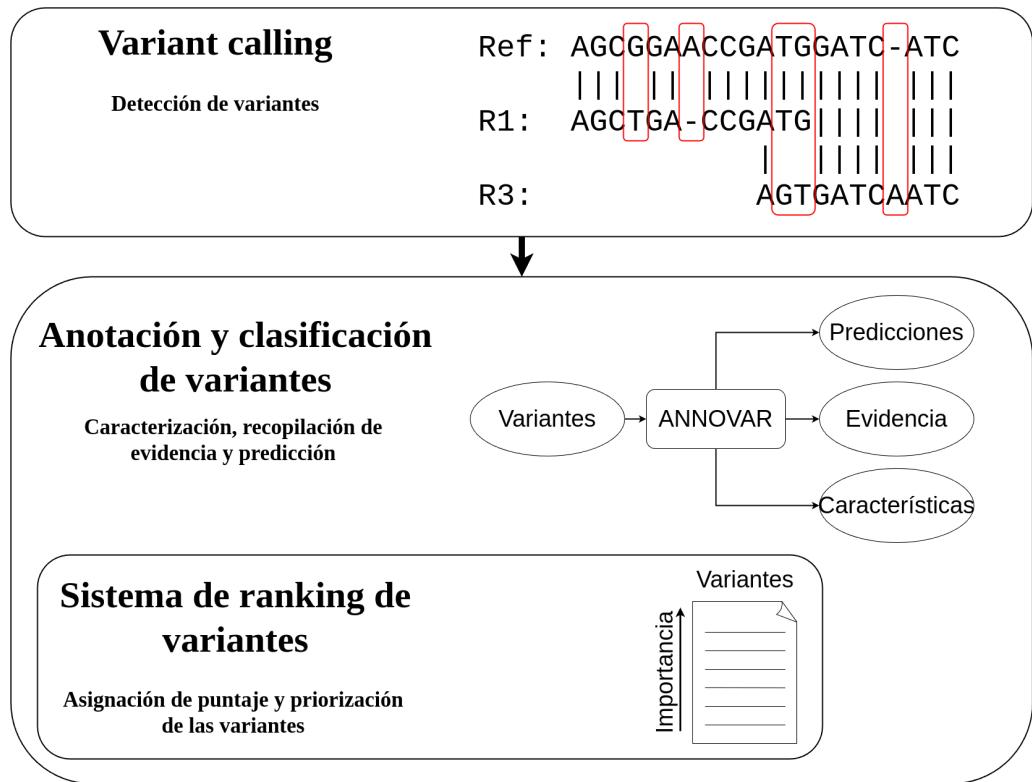


Figura 30: Relación entre el *variant calling* y la anotación de variantes.

Una vez seleccionadas las herramientas a comparar, se procedió a ejecutar cada una de ellas con sus configuraciones por defecto, salvo en los casos en que se requerían parámetros específicos para datos de secuenciación por nanoporo. Este fue el caso de **DeepVariant**, que **utiliza modelos pre-entrenados según la flowcell empleada**, y de **Clair3**, que además **depende del modelo de basecalling utilizado**. Todas las herramientas se ejecutaron sobre **CPU**, a excepción de **NanoSNP**, que requiere **GPU**, lo que debería verse reflejado en una mayor velocidad de procesamiento.

Tabla 18: Herramientas de variant calling seleccionadas.

Herramienta	Descripción
Bcftools	Herramienta para la detección de variantes y manipulación de archivos en el formato VCF .
Clair3	Detector de variantes pequeñas en línea germinal para lecturas largas. Clair3 combina dos categorías: pileup calling que maneja la mayoría de las variantes candidatas y full-alignment, que aborda los candidatos complicados para maximizar la precisión
Deepvariant	Herramienta de <i>variant calling</i> basada en redes neuronales que recibe lecturas alineadas y genera tensores de imagen de acumulación. Clasifica cada tensor mediante una red neuronal convolucional y, finalmente, reporta los resultados en un archivo estándar en formato VCF o gVCF (Poplin et al., 2018).

Herramienta	Descripción
Freebayes	Herramienta de <i>variant calling</i> con enfoque bayesiano. Funciona basándose en las secuencias literales de las lecturas alineadas a un objetivo particular, no en su alineación precisa (Garrison & Marth, 2012).
GATK	Detector de variantes mediante el ensamblaje local de haplotipos de novo en una región activa.
Lofreq	Herramienta de variant calling rápida y sensible para inferir SNVs ³⁵ e indels a partir de datos de secuenciación de nueva generación.
Longshot	<i>Variant caller</i> para genomas diploides utilizando lecturas largas propensas a errores. Solo para SNPs .
NanoCaller	<i>Variant caller</i> que implementa una red neuronal convolucional profunda para la detección de SNPs/indels a partir de datos de secuenciación de lecturas largas.
NanoSNP	Método de detección de SNPs basado en aprendizaje profundo para identificar SNPs a partir de lecturas de secuenciación Nanopore de baja cobertura.
PEPPER-Margin-Deepvariant	Pipeline para el llamado de variante, a través de tres herramientas: PEPPER : detector de variantes a través de redes neuronales recurrentes. Margin : Genera un archivo de alineación etiquetado por haplotipos utilizando un modelo oculto de Markov. DeepVariant : Detector de variantes a través de una red neuronal.
Snippy	Herramienta de <i>variant calling</i> para lecturas de secuenciación de nueva generación (NGS). Utiliza Freebayes para el llamado de variantes.
VarScan2	<i>Variant caller</i> para datos de secuenciación de nueva generación.

Tras la ejecución de cada herramienta, se obtuvieron los tiempos promedio de procesamiento para las ocho muestras, los cuales se presentan en la Tabla 19. En dicha tabla se observa la variabilidad de los tiempos según la herramienta utilizada, con resultados que van desde menos de un minuto hasta casi 14 horas, como fue el caso de **Freebayes** en las muestras procesadas con **Minimap2**.

³⁵ **SNV (Single Nucleotide Variation)**: Es una variación de un solo nucleótido en el genoma, sin requisito de frecuencia mínima en la población. A diferencia de los **SNPs**, que deben estar presentes en al menos el 1% y son siempre variantes heredables (germinales), las **SNVs** pueden incluir tanto variantes heredadas como mutaciones somáticas.

Tabla 19: Tiempo promedio en minutos de ejecución del *variant calling*.

Herramienta	Tiempo promedio de ejecución con muestras de Bowtie2	Tiempo promedio de ejecución con muestras de Minimap2
Bcftools	1,04	0,48
Clair3	1,98	3,0
Deepvariant	0,80	0,68
Freebayes	402,07	837,35
GATK	20,30	117,01
Lofreq	93,84	172,23
Longshot	0,60	2,71
NanoCaller	32,49	34,2
NanoSNP	3,98	3,94
PEPPER-Margin-Deepvariant	6,23	16,94
Snippy	48,34	59,81
VarScan2	5,17	5,03

En esta tabla no se observa una diferencia significativa en los tiempos de ejecución de las herramientas entre los dos grupos de muestras. Sin embargo, en la mayoría de los casos, las herramientas de *variant calling* parecen ejecutarse más rápido en las muestras mapeadas con **Bowtie2**.

Para analizar el número de variantes encontradas, primero se descartaron aquellas que no superaron los filtros establecidos por las propias herramientas. Este proceso se realizó utilizando **Bcftools** y se aplicó tanto a las muestras analizadas con **DeepVariant** como a las procesadas con **PEPPER-Margin-DeepVariant**.

Al analizar el número promedio de SNPs reportados por las herramientas de llamado de variantes (Tabla 20), se **observa que, en general, más de la mitad de los reportes presentan menos de 30 variantes**, lo cual podría considerarse un número esperado, dado que se estima una variante de tipo SNP por cada mil bases, y la región secuenciada abarca aproximadamente 29,600 bases. Sin embargo, también se presentan casos extremos, como **Bcftools** y **NanoCaller** con las muestras mapeadas con **Minimap2**, que reportan en promedio 22.515 y 18.605 variantes, respectivamente. Esto podría indicar que los parámetros de filtrado de variantes son demasiado permisivos. Además, **Freebayes** también destaca por reportar una gran cantidad de variantes en ambos grupos de muestras.

Tabla 20: Número de SNPs promedio reportados por cada herramienta.

Herramienta	Bowtie2	Minimap2
Bcftools	104	22515
Clair3	54	25
Deepvariant	22	22
Freebayes	3972	1546
GATK	19	21
Lofreq	21	84
Longshot	105	85
NanoCaller	351	18605
NanoSNP	18	9
PEPPER-Margin -Deepvariant	19	13
Snippy	13	5
VarScan2	136	16057

Al analizar el número de variantes reportadas según la herramienta de mapeo, se observa que, a pesar de la suposición de que un menor número de *mismatches* resultaría en un menor número de variantes reportadas, como se mencionó en el punto 4.3.4, los resultados de la Tabla 20 no reflejan dicha relación.

La Tabla 21 muestra el promedio de **INDELs** reportados por las distintas herramientas. Se observa que este número es generalmente bajo en la mayoría de los casos, lo cual es consistente con la frecuencia conocida de los **INDELs**, que es aproximadamente 10 veces menor que la de los **SNPs** (un **INDEL** cada 10.000 bases), lo que sugiere un promedio esperado de alrededor de tres **INDELs**. Sin embargo, las herramientas **Freebayes** y **Lofreq** reportan valores que se alejan considerablemente de este promedio esperado, lo que podría indicar que casi la totalidad de los **INDELs** reportados provienen de errores de secuenciación o de mapeo.

Al igual que en el caso anterior, no se refleja lo planteado en el punto 4.3.4, ya que el número de **INDELs** reportados es similar entre ambas herramientas de mapeo.

Finalmente, se evaluó el número de variantes confirmadas detectadas por cada herramienta (Tabla 22). Para ello, se aplicó el sistema de *ranking* de variantes, descrito en el punto 4.3.7, a todas las variantes reportadas que superaron los filtros establecidos por las propias herramientas de llamado de variantes.

Tabla 21: Número de INDELs promedio reportados por cada herramienta.

Herramienta	Bowtie2	Minimap2
Bcftools	0	1
Clair3	2	3
Deepvariant	1	2
Freebayes	199	211
GATK	21	21
Lofreq	17	7679
Longshot	0	0
NanoCaller	0	1
NanoSNP	0	0
PEPPER-Margin -Deepvariant	2	3
Snippy	5	0
VarScan2	19	5

En la Tabla 22 se puede observar que, con ambas herramientas de mapeo, alguna herramienta de *variant calling* logra identificar las ocho variantes. Inclusive, **existe al menos una combinación para cada herramienta de mapeo, que consigue clasificar las ocho variantes como variantes candidatas (VarScan2 con Bowtie2, y Clair3 y DeepVariant con Minimap2)**.

Entre los resultados, se destaca que las tres herramientas que lograron identificar y clasificar correctamente las ocho variantes solo lo hicieron en un grupo de muestras, ya sea las mapeadas con **Bowtie2** o con **Minimap2**. Por otro lado, también resaltan los resultados obtenidos por herramientas de *variant calling* específicas para NGS, como **GATK**, **Lofreq** y **VarScan2**. En la mayoría de los casos, estas herramientas detectaron correctamente las variantes, pero no lograron clasificarlas adecuadamente como candidatas, excepto **VarScan2**, que en las muestras mapeadas con **Bowtie2** identificó y puntuó correctamente las ocho variantes.

Evidentemente, **la elección de la herramienta de llamado de variantes se limitará a aquellas que lograron identificar y clasificar correctamente las variantes**, en este caso, **Clair3**, **DeepVariant** y **VarScan2**. Por lo tanto, el análisis comenzó evaluando los resultados desde la perspectiva de la herramienta de mapeo. Para las muestras mapeadas con **Bowtie2**, solo una herramienta (**VarScan2**) logró encontrar y clasificar correctamente las variantes, mientras que para las muestras mapeadas con **Minimap2**, dos herramientas (**Clair3** y **DeepVariant**) lo hicieron.

Tabla 22: Variantes confirmadas encontradas y posicionada como principal.

Herramienta	Variantes encontradas (Bowtie2)	Variantes posicionadas como principal (Bowtie2)	Variantes encontradas (Minimap2)	Variantes posicionadas como principal (Minimap2)
Bcf-tools	6 / 8	6 / 8	2 / 8	2 / 8
Clair3	7 / 8	6 / 8	8 / 8	8 / 8
Deepvariant	7 / 8	6 / 8	8 / 8	8 / 8
Freebayes	8 / 8	0 / 8	8 / 8	0 / 8
GATK	8 / 8	2 / 8	7 / 8	0 / 8
Lofreq	8 / 8	2 / 8	8 / 8	0 / 8
Longshot	5 / 8	3 / 8	5 / 8	5 / 8
NanoCaller	2 / 8	2 / 8	2 / 8	2 / 8
NanoSNP	2 / 8	2 / 8	0 / 8	0 / 8
PEPPER-Margin -Deepvariant	5 / 8	4 / 8	4 / 8	4 / 8
Snippy	0 / 8	0 / 8	0 / 8	0 / 8
VarScan2	8 / 8	8 / 8	2 / 8	0 / 8

Además, al analizar el tiempo promedio que tomó mapear cada muestra (Tabla 11), la evidencia sugiere que un mayor tiempo de ejecución no se traduce necesariamente en mejores resultados de *variant calling*. Por lo tanto, tanto **Bowtie2** como **VarScan2** no parecen ser las mejores opciones para el *pipeline* bioinformático.

Para decidir entre **Clair3** y **DeepVariant**, se tomaron en cuenta varios de los resultados obtenidos en el *variant calling*, como el número de **SNPs**, **INDELs** (ver Tabla 20 y Tabla 21) e incluso el tiempo de ejecución (ver Tabla 19). Sin embargo, surgieron dos problemas al realizar estas comparaciones. Primero, según los resultados de las tablas de **SNPs** e **INDELs**, no parecía haber una relación clara entre el número de variantes encontradas y la posibilidad de clasificarlas como variantes candidatas. Esto se puede observar en el caso de **VarScan2** con las muestras de **Bowtie2**, que logró identificar todas las variantes y clasificarlas correctamente, a pesar de ser la herramienta que encontró el mayor número de variantes.

Segundo, aunque **Clair3** fue, en promedio, 4.4 veces más lenta que **DeepVariant**, los tiempos de ejecución son considerablemente bajos y no representa un gran sacrificio

el elegir **Clair3**. Por lo tanto, se concluyó que la rapidez de ejecución no era un buen indicador para elegir entre una u otra herramienta.

De este modo, y dado que ambas herramientas habían obtenido el mejor resultado posible, se intentó tomar la decisión sobre cuál elegir basándose en la información disponible en la bibliografía científica. En este contexto, Barbitoff et al. (2022) realizaron un benchmark de nueve herramientas de *variant calling*, entre ellas **Clair3** y **DeepVariant**, utilizando un extenso conjunto de datasets, incluidos aquellos proporcionados por **GIAB** (*Genome In A Bottle*), considerados como *gold standard*. Es así como, en este estudio, **DeepVariant** obtuvo los mejores resultados en ambos tipos de datasets empleados (**WGS** y **WES**).

Es importante señalar que **el estudio de Barbitoff et al. se basó en muestras secuenciadas con tecnología Illumina**, lo cual puede influir en la comparabilidad de los resultados con las muestras de nanopore utilizadas en nuestra investigación. Además, las versiones de las herramientas evaluadas en el estudio son distintas de las que se consideraron en ese trabajo, lo que **introduce otra capa de variabilidad**. Estas limitaciones sugieren que, a pesar de la evidencia presentada en la literatura, **la aplicabilidad de los resultados a nuestro contexto específico podría no ser directa**.

Con el propósito de decidir entre ambas herramientas, se decidió realizar un *benchmark* de estas herramientas utilizando uno de los datasets de **GIAB**, en particular **GM12878**, proporcionado por *Oxford Nanopore Technologies*³⁶. Este dataset fue secuenciado con un kit de preparación de librerías V14 y *flowcells* R10.4.1, similares a los utilizados en este trabajo. Debido al tamaño de las muestras, se descargaron con las lecturas ya mapeadas. No obstante, al igual que el *pipeline* empleado en este estudio, se utilizó **Dorado** con el modelo "sup" y **Minimap2** para el mapeo, lo cual asegura la comparabilidad de los resultados.

Tras ejecutar el *variant calling*, se procedió a comparar las variantes detectadas por las herramientas con un *dataset* de variantes confirmadas, limitado a las regiones codificantes del genoma humano. Este archivo **VCF** contenía **21.012 variantes**, de las cuales **20.592 eran SNPs y 420 eran INDELs**. Al igual que en comparaciones previas, se descartaron las variantes que no superaron los filtros aplicados por las herramientas, y se compararon los archivos **VCF** generados con el conjunto de variantes reales (ver Tabla 23).

En la tabla se observa que **DeepVariant** reporta una mayor cantidad de variantes reales, siendo al mismo tiempo la herramienta que identifica la menor cantidad total de variantes (tanto **INDELs** como **SNPs**). Para evaluar el desempeño, se utilizaron métricas como **precisión**³⁷, **recall**³⁸ y **F1-score**³⁹, calculadas con **Hap.py**, una

³⁶ <https://labs.epi2me.io/giab-2023.05/>

³⁷ **Precisión:** Representa el porcentaje de valores que se han clasificado como positivos son realmente positivos.

³⁸ **Recall:** Representa el porcentaje de valores positivos correctamente clasificados como positivos.

³⁹ **F1-score:** Resume la precisión y el recall en un solo valor, mostrando qué tan bien el modelo equilibra ambos aspectos en la clasificación de los positivos.

herramienta comúnmente empleada para este tipo de evaluaciones. Los resultados obtenidos se presentan en la Tabla 24.

Tabla 23: Número de variantes reportadas

Herramienta	Tipo de variante	Variantes totales reportadas	Variantes reales reportadas
DeepVariant	INDEL	641	385
	SNP	24802	20479
Clair3	INDEL	5901	336
	SNP	48280	19659

Tabla 24: Estadísticas de precisión del *variant calling*.

Herramienta	Tipo de variante	Precisión	Recall	F1-Score
DeepVariant	INDEL	0,917	0,61	0,732
	SNP	0,995	0,826	0,902
Clair3	INDEL	0,8	0,174	0,286
	SNP	0,955	0,597	0,735

Los datos muestran claramente la superioridad de **DeepVariant** en cuanto a los resultados. Aunque **Clair3** logra una alta precisión para ambos tipos de variantes, reporta numerosas variantes inexistentes, lo que aplicado al entorno clínico, incrementa la probabilidad de clasificar erróneamente una variante como causal tras la clasificación y el sistema de *ranking* de variantes.

Después de comparar en profundidad las herramientas **DeepVariant** y **Clair3**, se concluye que **DeepVariant** es la opción más confiable, demostrando un mejor desempeño en distintos datasets, independientemente del origen de los datos. Por último, se selecciona **Minimap2** como la herramienta de mapeo elegida para el *pipeline*.

Con base en los resultados de frecuencia alélica de la variante (**VAF**) y la profundidad reportados por la herramienta de *variant calling* seleccionada (Tabla 25), se definieron los parámetros de filtrado para el sistema de *ranking*. En este proceso, se **filtraran las variantes con un VAF menor a 0.3 o mayor a 0.7**. Por otro lado, **no se aplicará un criterio adicional de profundidad**, dado que se considera que la herramienta **DeepVariant** ya realiza un filtrado exhaustivo, reteniendo únicamente las lecturas más representativas.

Tabla 25: VAF y profundidad de las variantes de cada paciente.

Paciente	VAF	Profundidad de lectura reportada
S1	0.56	205
S2	0.42	264
S3	0,5	6
S4	0.48	386
S5	0.61	213
S6	0.47	362
S7	0.51	323
S8	0.52	107

5. PIPELINE DE ANÁLISIS BIOINFORMÁTICO

En este capítulo se presenta el *pipeline* final, desarrollado a partir de la búsqueda y evaluación de herramientas descritas en el capítulo 4. Además, se presenta el manual detallado junto con el *script* diseñado para permitir la replicación de este *pipeline* por cualquier miembro del equipo GEMINI.

5.1 Optimización de parámetros de trimming y filtrado

Una vez seleccionadas las herramientas para cada etapa del *pipeline*, se procedió a optimizar ciertos parámetros con el objetivo de mejorar la precisión y efectividad en futuras aplicaciones. En particular, se consideró la etapa de **trimming y filtrado** de variantes, que, como se ha mencionado previamente, es fundamental para eliminar lecturas que no cumplen con los criterios de calidad definidos. Esta etapa contribuye a asegurar que solo las lecturas de alta calidad sean retenidas, lo cual es **esencial para garantizar la confiabilidad de los resultados obtenidos en el pipeline**.

Se evaluaron cuatro configuraciones de parámetros que incrementan gradualmente los requisitos de calidad mínimos para las lecturas. Estos parámetros incluyen filtros de longitud mínima y máxima, así como filtros que analizan la calidad de las lecturas, como el porcentaje máximo permitido de bases con baja calidad. Además, se aplicaron recortes utilizando ventanas deslizantes (explicadas en el punto 4.3.3), que analizan la calidad promedio de un número específico de bases desde ambos extremos de la lectura, eliminando las secciones que no cumplen con el umbral de calidad establecido. Los valores específicos de cada filtro aplicado se detallan en la Tabla 26.

Tabla 26: Configuraciones de parámetros para la etapa de *trimming* y filtrado.

Criterios	C1	C2	C3	C4
Largo mínimo de las lecturas.			30	
Largo máximo de las lecturas.			6000	
Porcentaje máximo de bases no cualificadas de una lectura.			40%	
Calidad mínima de una base para ser considerada cualificada (en escala Phred).	7	10	15	20
Largo de la ventana deslizante (en bases)			10	
Calidad promedio mínima de la ventana (en escala Phred).	7	10	15	20

Tras aplicar cada uno de estos filtros a las 8 muestras, se comparó el promedio del porcentaje de lecturas y bases filtradas (ver Tabla 27). Como era de esperarse, a

medida que aumentaba la exigencia de los parámetros de filtrado, disminuye el número de lecturas obtenidas. Sin embargo, **las diferencias entre las distintas configuraciones no fueron extremas**: entre la configuración C1 y C4 hubo solo un 3% de diferencia en la cantidad de bases y aproximadamente un 1.5% en la cantidad de lecturas conservadas. Esto puede deberse a la buena calidad de las muestras, lo cual podría variar en caso de trabajar con muestras de menor calidad.

Tabla 27: Porcentaje promedio de bases y lecturas conservadas.

Configuración	Porcentaje promedio de bases conservadas	Porcentaje promedio de lecturas conservadas
C1	94,63%	98,84%
C2	94,21%	98,77%
C3	93,22%	98,36%
C4	91,77%	97,45%

A continuación, se mapearon las muestras contra el genoma de referencia **GRCh38/h38** y se comparó el porcentaje promedio de lecturas mapeadas en general y aquellas que corresponden a la región objetivo (ver Tabla 28). Los resultados muestran que, aunque las diferencias entre las configuraciones son pequeñas, una diferencia del 0.1% en el porcentaje de lecturas puede representar aproximadamente 500 lecturas, lo cual podría ser relevante a la hora de maximizar la profundidad de lectura. En este sentido, la configuración C4 logra mapear un mayor porcentaje de lecturas en general, mientras que la configuración C1 logra un mayor porcentaje de lecturas en las regiones objetivo.

Tabla 28: Porcentaje promedio de lecturas mapeadas.

Configuración	Porcentaje promedio de lecturas mapeadas	Porcentaje promedio de lecturas mapeadas en la regiones objetivo
C1	85,84%	44,524%
C2	85,838%	44,505%
C3	85,873%	44,482%
C4	85,962%	44,471%

Finalmente, se verificó que las cuatro configuraciones identificaran y clasificarán correctamente las variantes, y luego se comparó el número de variantes reportadas (ver Tabla 29). Contrario a lo esperado, **un filtrado y recorte más estricto de las lecturas no resultó en una menor cantidad de variantes**. De hecho, con la configuración más exigente (C4), el número de variantes reportadas aumentó en la mayoría de los casos.

Tabla 29: Número de variantes reportadas en cada muestra.

Configuración	S1	S2	S3	S4	S5	S6	S7	S8
C1	13	19	21	22	17	26	53	18
C2	14	18	22	23	20	25	51	18
C3	13	19	22	23	20	24	53	19
C4	13	21	22	23	19	26	50	20

Finalmente, en base a las evaluaciones realizadas, se concluyó que la configuración **C1** es la más adecuada para ser incorporada al *pipeline*. Esto se debe a que, como se observó en las comparaciones, aplicar valores más estrictos en el filtrado y recorte de lecturas no parece tener un impacto significativo en los resultados, ya que no se evidencia un aumento en el porcentaje de lecturas mapeadas en las regiones objetivo. Además, criterios más estrictos en el filtrado y recorte de lecturas tampoco parecen reducir el número de variantes reportadas. De hecho, como se muestra en la Tabla 29, estos criterios parecen aumentar el número de variantes, lo que podría traducirse en una mayor probabilidad de que el sistema de *ranking* de variantes identifique incorrectamente una variante como candidata para ser responsable de la enfermedad.

5.2 Pipeline bioinformático para Oxford Nanopore MinION

Luego de realizar una búsqueda exhaustiva y una comparación detallada de las herramientas disponibles para cada etapa del *pipeline*, desde el *basecalling* hasta la clasificación y anotación de las variantes detectadas, **se seleccionaron un total de 30 herramientas**. De estas, 23 corresponden a la etapa de clasificación y anotación de variantes. Además, **se emplearon ocho bases de datos de diferentes tipos**, con el propósito de identificar diversas características de las variantes.

El flujo de ejecución de este *pipeline* se ilustra en la Figura 31, donde se muestra la interacción entre cada herramienta. Se especifican los archivos de entrada y salida para cada etapa, proporcionando una visión clara de cómo los datos fluyen a lo largo del proceso.

Este *pipeline* logró cumplir con la visión propuesta, detectando y reportando correctamente las variantes en las ocho muestras analizadas. La ejecución completa del *pipeline* tardó aproximadamente **5 horas y 34 minutos**. Esto incluye el tiempo de ejecución de cada etapa para cada una de las ocho muestras. Entre ellas, el *basecalling* fue la etapa que más tiempo consumió, con aproximadamente 3 horas y 47 minutos. Los tiempos que demoró cada etapa se pueden consultar en la Tabla 30; cabe destacar que, a partir de la etapa de control de calidad hasta el sistema de *ranking*, el tiempo refleja la ejecución de cada una de las ocho muestras.

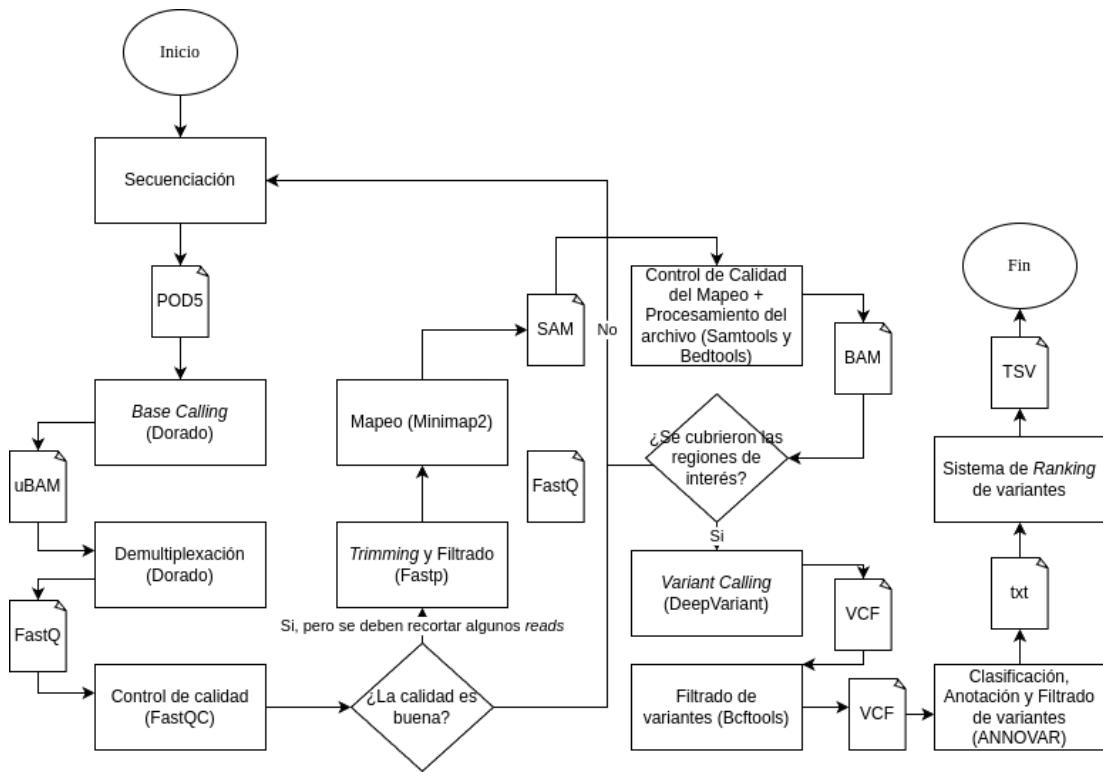


Figura 31: Pipeline para el análisis bioinformático.

Tabla 30: Tiempo de ejecución de cada etapa del *pipeline*.

Etapa	Tiempo (Minutos)
<i>Basecalling</i>	227,4
Demultiplexación	4,8
Control de calidad	3,7
<i>Trimming</i> y filtrado	1,9
Mapeo	70,4
Control de calidad del mapeo	13,8
<i>Variant calling</i>	4,7
Anotaciones	7,2
Sistema de <i>ranking</i> de variantes	0,7

5.3 Manual de usuario

Una vez seleccionadas las herramientas, se elaboró un manual de usuario que fue validado por Paola Krall, quien es una de las principales usuarias de este *pipeline*. Este manual está disponible en el Anexo E, que detalla cada uno de los pasos necesarios para la ejecución del *pipeline*. Este manual permite a cualquier miembro del equipo comprender el funcionamiento del *pipeline* y aplicarlo a nuevas muestras. Además, el manual incluye una sección dedicada a tanto a los requisitos de *hardware* y *software*.

Junto con la elaboración del manual, se desarrolló un *script* en **Bash** que automatiza la ejecución de cada una de las etapas del *pipeline*, utilizando un archivo de configuración que contiene las variables necesarias, como la ubicación de las muestras provenientes del secuenciador. Además, el *script* permite la ejecución a partir de cinco etapas distintas, dependiendo del tipo de archivo de entrada, como se detalla en la Tabla 31. Este *script* también incorpora herramientas previamente desarrolladas, mencionadas en el apartado 4.3.7, para ejecutar el sistema de *ranking* y obtener la calificación de las variantes.

Tabla 31: Relación entre archivos de entrada y etapas del *pipeline*.

Tipo de archivo	Etapa de partida
POD5	<i>Basecalling</i>
FASTQ	Control de calidad
SAM	Control de calidad del mapeo
BAM	<i>Variant calling</i>
VCF	Anotación y clasificación de variantes

6. CONCLUSIÓN Y TRABAJO FUTURO

A continuación, se presenta una breve conclusión tras la realización de este trabajo, junto con propuestas para trabajos futuros que pueden desarrollarse a partir de los resultados obtenidos.

6.1 Conclusión y cumplimiento de objetivos

Al inicio de este trabajo, se estableció como objetivo mejorar las capacidades de detección y diagnóstico de la **ADPKD** a partir de la secuenciación del gen **PKD1**, mediante un manual que documente la realización del proceso de análisis de las secuencias generadas por el secuenciador **MinION**. Para lograrlo, se plantearon cuatro objetivos específicos, los cuales contribuyen al cumplimiento del objetivo general. A continuación, se revisarán cada uno de los objetivos específicos con el fin de examinar su cumplimiento.

El primer objetivo específico era “Manejar el funcionamiento y características principales del secuenciador *Oxford Nanopore MinION* como: funcionamiento, formatos de archivos de salidas, requerimientos de *hardware, software, etc*”. Este objetivo permitiría adentrarse en el campo de la secuenciación por nanoporos, facilitando la comprensión de sus diferencias con respecto a los secuenciadores **NGS**. El cumplimiento de este objetivo se refleja en el capítulo 2, donde se revisa el funcionamiento de esta tecnología.

El segundo objetivo específico consistió en “Obtener al menos tres herramientas necesarias para cada una de las distintas etapas del análisis, como con el control de calidad, *trimming*, mapeo de la secuencia, *variant calling*, consulta a bases de datos de variantes y la predicción de patogenicidad”. Este objetivo se logró mediante una revisión sistemática de la literatura, presentada en el capítulo 3, donde se hallaron 14 herramientas para el control de calidad, 11 para el *trimming* y filtrado de variantes, 27 herramientas de mapeo, 41 herramientas de *variant calling*, 17 bases de datos y 36 herramientas de predicción de patogenicidad y clasificación de variantes.

El tercer objetivo específico era “Establecer un *pipeline* de análisis bioinformático, para el diagnóstico de **ADPKD**, mediante el análisis de **PKD1** generadas con el dispositivo **MinION** y con el uso de herramientas halladas en el objetivo específico dos y validado con muestras de pacientes con variantes conocidas y desconocidas.”. Este objetivo se ve plasmado en el capítulo 5, donde, luego de realizar una evaluación de cada una de las herramientas para cada etapa, concluyendo con una lista de herramientas que forman el *pipeline* de análisis final, las cuales alcanzaron un 100% sensibilidad y especificidad en las ocho muestras utilizadas para validar su funcionamiento.

El cuarto y último objetivo específico era “Permitir que los miembros del equipo puedan aplicar el *pipeline*, mediante un manual, que facilite la aplicación de este protocolo de análisis post-secuenciación para ser aplicado en futuros análisis de secuenciaciones”. El cumplimiento de este objetivo se evidencia con la redacción del Anexo E, en donde se describe paso a paso la ejecución completa del *pipeline*, junto con la incorporación de *scripts* que automatizan tanto el proceso completo como

etapas específicas, como la aplicación del sistema de *ranking* de variantes y la consulta en la base de datos de la Clínica Mayo.

Como conclusión, este trabajo cumplió con el objetivo general, permitiendo la exploración del secuenciador *Oxford Nanopore MinION* en el entorno clínico para el diagnóstico de enfermedades hereditarias. Además, se proporcionó al equipo de **GEMINI** un *pipeline* capaz de procesar los datos generados por el secuenciador y entregar una lista de variantes clasificadas y reportadas según su relevancia patogénica. Como resultado, se logró el éxito en la identificación y clasificación de las variantes conocidas de las ocho muestras de los pacientes de prueba.

Finalmente, durante el desarrollo de este trabajo, se evidenciaron las ventajas del uso del secuenciador **MinION**, como la reducción del número de **PCRs** de 70 a solo nueve. Esto repercutió directamente en la disminución de costos, tiempos y en la reducción de posibles errores en la preparación de muestras. Por otro lado, se incrementaron significativamente las regiones cubiertas del gen **PKD1**, logrando secuenciar todos los exones y 41 de los 45 intrones, mientras que los cuatro intrones restantes se secuenciaron parcialmente, con una profundidad de lectura superior a **1000x..**

6.2 Limitaciones

A pesar de los resultados satisfactorios obtenidos, **es importante reconocer ciertas limitaciones en este trabajo**. A continuación, se describen los aspectos que podrían influir en el desempeño y la aplicabilidad del *pipeline* desarrollado.

Aunque algunas herramientas de *variant calling* lograron detectar las variantes en las ocho muestras analizadas, esto no garantiza un rendimiento óptimo en todos los casos. **Los resultados obtenidos dependen, en gran medida, de la calidad de la secuenciación**, lo que puede comprometer la precisión en escenarios con datos de menor calidad.

La elección de una herramienta de *variant calling* debe ser revisada periódicamente. Muchas de las herramientas más recientes emplean enfoques basados en *deep learning*, con modelos entrenados específicamente para un secuenciador o características particulares. Además, las continuas actualizaciones en los materiales y software del secuenciador **MinION** exigen que las herramientas de *variant calling* se mantengan actualizadas. Por lo tanto, es recomendable evaluar de forma empírica (como se hizo en este trabajo) o a través de la literatura científica cuál es la mejor opción cada cierto tiempo.

Este principio también aplica a la etapa de *basecalling*. Aunque en la actualidad la opción más confiable suele ser la proporcionada por el fabricante del secuenciador, **es fundamental mantener actualizados tanto la herramienta** como los modelos utilizados en esta etapa para garantizar la mayor precisión posible.

Finalmente, este trabajo tuvo como objetivo principal la búsqueda de variantes catalogadas como **SNPs** e **INDELS**, que representan la mayoría de los casos de

ADPKD. Sin embargo, es importante señalar que este enfoque dejó fuera otros tipos de alteraciones genéticas, como las **variantes estructurales** o las **variaciones en el número de copias**.

6.3 Trabajo futuro

Existen varias áreas que pueden ser exploradas y optimizadas en investigaciones futuras. A continuación se describen algunas propuestas para el desarrollo y mejora del *pipeline* y sus aplicaciones.

El *pipeline* desarrollado en este trabajo se centró exclusivamente en el gen ***PKD1***, ya que representa aproximadamente el 85% de los casos de **ADPKD**. Sin embargo, como propuesta para trabajos futuros, se podría investigar la aplicación de este *pipeline* en muestras de pacientes cuyas variantes patogénicas se encuentren en el gen ***PKD2***, que representa alrededor del 15% de los casos, e incluso en ***IFT140***, que representa menos del 1% de los casos.

Por otro lado, tal como se mencionó anteriormente en la sección de limitaciones, este trabajo sólo consideró variantes de tipo **SNPs** e **INDELS**, que constituyen la mayoría de los casos. Por lo tanto, se sugiere también como trabajo futuro la ampliación del *pipeline* mediante la incorporación de herramientas capaces de detectar y clasificar **variantes estructurales** y **variaciones en el número de copias**. Esto contribuiría a desarrollar un *pipeline* más robusto para la detección y diagnóstico de **ADPKD**.

Finalmente, como resultado de la revisión sistemática, se identificó una amplia variedad de herramientas, muchas de las cuales no fueron utilizadas ni evaluadas en este *pipeline*. No obstante, es importante señalar que algunas de estas herramientas podrían ser de gran utilidad, e incluso esenciales, cuando se trabaja con datos que presentan una alta tasa de error, como los generados por el secuenciador **MinION**. Entre ellas se destacan **DuploMap**, que realinea secuencias mapeadas para mejorar la precisión del análisis, y **GLIMPSE**, que predice variantes genéticas no observadas directamente en los datos de secuenciación. La incorporación de estas herramientas en futuros estudios podría enriquecer el análisis y aumentar la detección de variantes relevantes.

7. REFERENCIAS

- Andrews, T. D., Jeelall, Y., Talaulikar, D., Goodnow, C. C., & Field, M. A. (2016). DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations. *PeerJ*, 4, e2074. <https://doi.org/10.7717/peerj.2074>
- Ars E. (2021). Enfoque Genético de las Enfermedades Renales Hereditarias. En: Lorenzo V., López Gómez JM (Eds). Nefrología al día. ISSN: 2659-2606. Disponible en: <https://www.nefrologiaaldia.org/359>
- Barbitoff, Y. A., Abasov, R., Tvorogova, V. E., Glotov, A. S., & Predeus, A. V. (2022). Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. *BMC Genomics*, 23(1). <https://doi.org/10.1186/s12864-022-08365-3>
- Benz, E. G., & Hartung, E. A. (2021). Predictors of progression in autosomal dominant and autosomal recessive polycystic kidney disease. *Pediatric Nephrology*, 36(9), 2639-2658. <https://doi.org/10.1007/s00467-020-04869-w>
- Bimber, B. N., Yan, M. Y., Peterson, S. M., & Ferguson, B. (2019). mGAP: the macaque genotype and phenotype resource, a framework for accessing and interpreting macaque variant data, and identifying new models of human disease. *BMC Genomics*, 20(1). <https://doi.org/10.1186/s12864-019-5559-7>
- Bogaerts, B., Van Den Bossche, A., Verhaegen, B., Delbrassinne, L., Mattheus, W., Nouws, S., Godfroid, M., Hoffman, S., Fraiture, M., De Keersmaecker, S. C. J., & Vanneste, K. (2024). Closing the gap: Oxford Nanopore Technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. *Journal Of Clinical Microbiology*. <https://doi.org/10.1128/jcm.01576-23>
- Buza, T., Tonui, T., Stomeo, F., Tiambo, C. K., Katani, R., Schilling, M. A., Lyimo, B., Gwakisa, P., Cattadori, I. M., Buza, J., & Kapur, V. (2019). iMAP: an integrated bioinformatics and visualization pipeline for microbiome data analysis. *BMC Bioinformatics*, 20(1). <https://doi.org/10.1186/s12859-019-2965-4>
- Cao, Y., Ha, S. Y., So, C., For, T. M., Tang, C., Zhang, H., Liang, R., Yang, J., Chung, B. H., Chan, G. C., Lau, Y., García-Barceló, M., Shiu-Kwan, E., Sucharitchan, P., Hirankarn, N., & Yang, W. (2022). NGS4THAL, a One-Stop Molecular Diagnosis and Carrier Screening Tool for Thalassemia and Other Hemoglobinopathies by Next-Generation Sequencing. *The Journal Of Molecular Diagnostics/The Journal Of Molecular Diagnostics*, 24(10), 1089-1099. <https://doi.org/10.1016/j.jmoldx.2022.06.006>
- Carbo, E. C., Mourik, K., Boers, S. A., Munnink, B. B. O., Nieuwenhuijse, D. F., Jonges, M., Welkers, M. R., Matamoros, S., Van Harinxma Thoe Slooten, J.,

- Kraakman, M. E. M., Karelotti, E., Van Der Meer, D., Veldkamp, K. E., Kroes, A. C., Sidorov, I. A., & De Vries, J. J. (2023). A comparison of five Illumina, Ion Torrent, and nanopore sequencing technology-based approaches for whole genome sequencing of SARS-CoV-2. *European Journal Of Clinical Microbiology & Infectious Diseases*, 42(6), 701-713. <https://doi.org/10.1007/s10096-023-04590-0>
- Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., & Karchin, R. (2013). Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genomics*, 14(S3). <https://doi.org/10.1186/1471-2164-14-s3-s3>
- Салахов, Р. Р., Голубенко, М. В., Валиахметов, Н. Р., Pavlyukova, Е. Н., Зарубин, А. А., Бабушкина, Н. П., Кучер, А. Н., Слепцов, А. А., & Назаренко, М. С. (2022). Application of Long-Read Nanopore Sequencing to the Search for Mutations in Hypertrophic Cardiomyopathy. *International Journal Of Molecular Sciences*, 23(24), 15845. <https://doi.org/10.3390/ijms232415845>
- Castro, I. (2022). Diseño de paneles de partidores multiplexados para diagnóstico genético. Trabajo para optar al título de Ingeniero Civil en Informática, UACH.
- Cherif, E., Thiam, F., Salma, M., Rivera-Ingraham, G. A., Justy, F., Deremarque, T., Breugnot, D., Doudou, J., Gozlan, R. E., & Combe, M. (2022). ONTdeCIPHER: an amplicon-based nanopore sequencing pipeline for tracking pathogen variants. *Bioinformatics*, 38(7), 2033-2035. <https://doi.org/10.1093/bioinformatics/btac043>
- Cheung, M. K., & Kwan, H. S. (2012). Fighting Outbreaks with Bacterial Genomics: Case Review and Workflow Proposal. *Public Health Genomics*, 15(6), 341-351. <https://doi.org/10.1159/000342770>
- Cornejo, S. M. (2022, 5 mayo). ¿Cómo funciona la transcripción genética del ADN? *Microbacterium*. <https://microbacterium.es/como-funciona-la-transcripcion-del-adn>
- Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Computational Biology*, 6(12), e1001025. <https://doi.org/10.1371/journal.pcbi.1001025>
- De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Stražišar, M., Sleegers, K., & Van Broeckhoven, C. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Research*, 29(7), 1178-1187. <https://doi.org/10.1101/gr.244939.118>
- De Souza, V. B. C., Jordan, B., Tseng, E., Nelson, E. A., Hirschi, K. K., Sheynkman, G., & Robinson, M. D. (2023). Transformation of alignment files improves performance of variant callers for long-read RNA sequencing data. *Genome Biology*, 24(1). <https://doi.org/10.1186/s13059-023-02923-y>

Dolled-Filhart, M., Lee, M., Ou-Yang, C., Haraksingh, R., & Lin, J. (2013). Computational and Bioinformatics Frameworks for Next-Generation Whole Exome and Genome Sequencing. *The Scientific World Journal*, 2013, 1-10. <https://doi.org/10.1155/2013/730210>

Durkie, M., Watson, C. M., Winship, P. R., Hogg, A. C., Nyanhete, R., Cooley, S., Valluru, M. K., Shaw-Smith, C., Bingham, C., Gilchrist, M., Kenny, J., Consortium, G. E. R., & Ong, A. (2023). The Common PKD1 p.(Ile3167Phe) Variant Is Hypomorphic and Associated with Very Early Onset, Biallelic Polycystic Kidney Disease. *Human Mutation*, 2023, 1-8. <https://doi.org/10.1155/2023/5597005>

Försti, A., Kumar, A., Paramasivam, N., Schlesner, M., Catalano, C., Dymerska, D., Lubiński, J., Eils, R., & Hemminki, K. (2016). Pedigree based DNA sequencing pipeline for germline genomes of cancer families. *Heredity Cancer In Clinical Practice*, 14(1). <https://doi.org/10.1186/s13053-016-0058-1>

Fu, Y., Mahmoud, M., Muraliraman, V. V., Sedlazeck, F. J., & Treangen, T. J. (2021). Vulcan: Improved long-read mapping and structural variant calling via dual-mode alignment. *Gigascience*, 10(9). <https://doi.org/10.1093/gigascience/giab063>

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.1207.3907>

Goenka, S. D., Gorzynski, J. E., Shafin, K., Fisk, D. G., Pesout, T., Jensen, T., Monlong, J., Chang, P., Baid, G., Bernstein, J. A., Christle, J. W., Dalton, K., Garalde, D. R., Grove, M. E., Guillory, J., Kolesnikov, A., Nattestad, M., Ruzhnikov, M., Samadi, M., . . . Ashley, E. A. (2022). Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing. *Nature Biotechnology*, 40(7), 1035-1041. <https://doi.org/10.1038/s41587-022-01221-5>

Greig, D. R., Jenkins, C., Gharbia, S. E., & Dallman, T. J. (2019). Comparison of single-nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin-producing *Escherichia coli*. *Gigascience*, 8(8). <https://doi.org/10.1093/gigascience/giz104>

Grumaz, C., Hoffmann, A., Vainshtein, Y., Kopp, M., Grumaz, S., Stevens, P., Decker, S., Weigand, M. A., Hofer, S., Brenner, T., & Sohn, K. (2020). Rapid Next-Generation Sequencing-Based Diagnostics of Bacteremia in Septic Patients. *The Journal Of Molecular Diagnostics/The Journal Of Molecular Diagnostics*, 22(3), 405-418. <https://doi.org/10.1016/j.jmoldx.2019.12.006>

- Helal, A. A., Saad, B. T., Saad, M. T., Mosaad, G. S., & Aboshanab, K. M. (2022). Evaluation of the Available Variant Calling Tools for Oxford Nanopore Sequencing in Breast Cancer. *Genes*, 13(9), 1583. <https://doi.org/10.3390/genes13091583>
- Holmqvist, I., Bäckerholm, A., Tian, Y., Xie, G., Thorell, K., & Tang, K. (2021). FLAME: long-read bioinformatics tool for comprehensive spliceome characterization. *RNA*, 27(10), 1127-1139. <https://doi.org/10.1261/rna.078800.121>
- Hosch, S., Wagner, P., Giger, J. N., Dubach, N., Saavedra, E., Perno, C. F., Gody, J., Pagonendji, M., Ngoagouni, C., Ndoua, C., Nsanzabana, C., Vickos, U., Daubenberger, C., & Schindler, T. (2024). PHARE: a bioinformatics pipeline for compositional profiling of multiclonal Plasmodium falciparum infections from long-read Nanopore sequencing data. *The Journal Of Antimicrobial Chemotherapy/Journal Of Antimicrobial Chemotherapy*. <https://doi.org/10.1093/jac/dkae060>
- Kassem, H. S., Girolami, F., & Sanoudou, D. (2012). Molecular genetics made simple. *Global Cardiology Science & Practice*, 2012(1), 6. <https://doi.org/10.5339/gcsp.2012.6>
- Kirov, I., Kolganova, E., Dudnikov, M., Yurkevich, O. Y., Amosova, A. V., & Muravenko, O. V. (2022). A Pipeline NanoTRF as a New Tool for De Novo Satellite DNA Identification in the Raw Nanopore Sequencing Reads of Plant Genomes. *Plants*, 11(16), 2103. <https://doi.org/10.3390/plants11162103>
- Klenner, J. (2018). Definición de primers para particionamiento de ADN. Trabajo para optar al título de Ingeniero Civil en Informática, UACh.
- Lama, M., Pulusu, C. P., Khamari, B., Peketi, A. S. K., Kumar, P., Nagaraja, V., & Bulagonda, E. P. (2021). Genomic and phylogenetic analysis of a multidrug-resistant Burkholderia contaminans strain isolated from a patient with ocular infection. *Journal Of Global Antimicrobial Resistance*, 25, 323-325. <https://doi.org/10.1016/j.jgar.2021.04.004>
- Leekitcharoenphon, P., Sørensen, G., Löfström, C., Battisti, A., Szabó, I., Wasyl, D., Slowey, R., Zhao, S., Brisabois, A., Kornschober, C., Kärssin, A., Jánosi, S., Cerny, T., Svendsen, C. A., Pedersen, K., Aarestrup, F. M., & Hendriksen, R. S. (2019). Cross-Border Transmission of *Salmonella Choleraesuis* var. Kunzendorf in European Pigs and Wild Boar: Infection, Genetics, and Evolution. *Frontiers In Microbiology*, 10. <https://doi.org/10.3389/fmicb.2019.00179>
- Lepuschitz, S., Weinmaier, T., Mrazek, K., Beisken, S., Weinberger, J., & Posch, A. E. (2020). Analytical Performance Validation of Next-Generation Sequencing Based Clinical Microbiology Assays Using a K-mer Analysis Workflow. *Frontiers In Microbiology*, 11. <https://doi.org/10.3389/fmicb.2020.01883>

- Leung, A. W., Leung, H. Y., Wong, C., Zheng, Z., Lui, W., Luk, H., Lo, I. F. M., Luo, R., & Lam, T. (2022). ECNano: A cost-effective workflow for target enrichment sequencing and accurate variant calling on 4800 clinically significant genes using a single MinION flowcell. *BMC Medical Genomics*, 15(1). <https://doi.org/10.1186/s12920-022-01190-3>
- Lunter, G., & Goodson, M. (2010). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6), 936-939. <https://doi.org/10.1101/gr.111120.110>
- Maestri, S., Maturo, M. G., Cosentino, E., Marcolungo, L., Iadarola, B., Fortunati, E., Rossato, M., & Delledonne, M. (2020). A Long-Read Sequencing Approach for Direct Haplotype Phasing in Clinical Settings. *International Journal Of Molecular Sciences*, 21(23), 9177. <https://doi.org/10.3390/ijms21239177>
- Mahboob M, Rout P, Bokhari SRA. Autosomal Dominant Polycystic Kidney Disease. [Updated 2023 Oct 18]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK532934/>
- Mecanismos moleculares de la replicación del ADN | Khan Academy. Accedido el 13 de abril, 2024, desde. <https://es.khanacademy.org/science/ap-biology/gene-expression-and-regulation/replication/a/molecular-mechanism-of-dna-replication>
- Nakamura, W., Hirata, M., Oda, S., Chiba, K., Okada, A., Mateos, R. N., Sugawa, M., Iida, N., Ushijima, M., Tanabe, N., Sanada, H., Sekine, S., Hirasawa, A., Kawai, Y., Tokunaga, K., Ishibashi-Ueda, H., Tomita, T., Noguchi, M., Takahashi, A., . . . Shiraishi, Y. (2024). Assessing the efficacy of target adaptive sampling long-read sequencing through hereditary cancer patient genomes. *Npj Genomic Medicine*, 9(1). <https://doi.org/10.1038/s41525-024-00394-z>
- Norri, T., Cazaux, B., Dönges, S., Valenzuela, D., & Mäkinen, V. (2021). Founder reconstruction enables scalable and seamless pangenomic analysis. *Bioinformatics*, 37(24), 4611-4619. <https://doi.org/10.1093/bioinformatics/btab516>
- Orsini, P., Minervini, C. F., Cumbo, C., Anelli, L., Zagaria, A., Minervini, A., Coccaro, N., Tota, G., Casieri, P., Impera, L., Parciante, E., Brunetti, C., Giordano, A., Specchia, G., & Albano, F. (2018). Design and MinION testing of a nanopore targeted gene sequencing panel for chronic lymphocytic leukemia. *Scientific Reports*, 8(1). <https://doi.org/10.1038/s41598-018-30330-y>

Pagnamenta, A. T., Camps, C., Giacopuzzi, E., Taylor, J. M., Hashim, M., Calpena, E., Kaisaki, P. J., Hashimoto, A., Yu, J., Sanders, E., Schweßinger, R., Hughes, J. R., Lunter, G., Dreau, H., Ferla, M. P., De Lange, L. F., Kesim, Y., Ragoussis, V., Vavoulis, D. V., . . . Taylor, J. C. (2023). Structural and non-coding variants increase the diagnostic yield of clinical whole genome sequencing for rare diseases. *Genome Medicine*, 15(1). <https://doi.org/10.1186/s13073-023-01240-0>

Paschal, C., Galey, M., Beck, A. E., Gillentine, M. A., Narayanan, J., Damaraju, N., Goffena, J., Storz, S. H. R., & Miller, D. E. (2024). Concordance of whole-genome long-read sequencing with standard clinical testing for Prader-Willi and Angelman syndromes. *medRxiv (Cold Spring Harbor Laboratory)*. <https://doi.org/10.1101/2024.04.02.24305233>

PKD1 gene | MedlinePlus Genetics. Accedido el 13 de abril, 2024, desde <https://medlineplus.gov/genetics/gene/pkd1>.

Poplin, R., Chang, P., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10), 983-987. <https://doi.org/10.1038/nbt.4235>

Proyecto GEMINI. Accedido el 10 de abril, 2024, desde <https://www.proyectogemini.cl/>

Ramachandran, A., Lumetta, S. S., Klee, E. W., & Chen, D. (2021). HELLO: improved neural network architectures and methodologies for small variant calling. *BMC Bioinformatics*, 22(1). <https://doi.org/10.1186/s12859-021-04311-4>

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics In Medicine*, 17(5), 405-424. <https://doi.org/10.1038/gim.2015.30>

Realizaron exitoso trasplante renal identificando previamente si existía riesgo de enfermedad renal hereditaria. (2024). Facultad de Medicina. <https://medicina.uach.cl/2024/03/19/realizaron-exitoso-trasplante-renal-identificando-previamente-si-existia-riesgo-de-enfermedad-renal-hereditaria/>

Rodríguez-Pérez, H., Ciuffreda, L., & Flores, C. (2022). NanoRTax, a real-time pipeline for taxonomic and diversity analysis of nanopore 16S rRNA amplicon sequencing data. *Computational And Structural Biotechnology Journal*, 20, 5350-5354. <https://doi.org/10.1016/j.csbj.2022.09.024>

- Salazar, C., Ferrés, I., Paz, M., Costáble, A., Moratorio, G., Moreno, P., & Iraola, G. (2023). Fast and cost-effective SARS-CoV-2 variant detection using Oxford Nanopore full-length spike gene sequencing. *Microbial Genomics*, 9(5). <https://doi.org/10.1099/mgen.0.001013>
- Samarakoon, H., Punchihewa, S., Senanayake, A., Hammond, J. M., Stevanovski, I., Ferguson, J. M., Ragel, R., Gamaarachchi, H., & Deveson, I. W. (2020). Genopo: a nanopore sequencing analysis toolkit for portable Android devices. *Communications Biology*, 3(1). <https://doi.org/10.1038/s42003-020-01270-z>
- Schmidt, J., Berghaus, S., Blessing, F., Herbeck, H., Blessing, J., Schierack, P., Rödiger, S., Roggenbuck, D., & Wenzel, F. (2022). Genotyping of familial Mediterranean fever gene (MEFV)—Single nucleotide polymorphism—Comparison of Nanopore with conventional Sanger sequencing. *PloS One*, 17(3), e0265622. <https://doi.org/10.1371/journal.pone.0265622>
- Shafin, K., Pesout, T., Chang, P., Nattestad, M., Kolesnikov, A., Goel, S., Baid, G., Kolmogorov, M., Eizenga, J., Miga, K. H., Carnevali, P., Jain, M., Carroll, A., & Paten, B. (2021). Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods*, 18(11), 1322-1332. <https://doi.org/10.1038/s41592-021-01299-w>
- Shishido, S. N., Welter, L., Rodríguez-Lee, M., Kolatkar, A., Xu, L., Ruiz, C. C., Gerdts, A. S., Restrepo-Vassalli, S., Carlsson, A., Larsen, J. F., Greenspan, E. J., Hwang, E. J., Waitman, K. R., Nieva, J., Bethel, K., Hicks, J., & Kühn, P. (2020). Preanalytical Variables for the Genomic Assessment of the Cellular and Acellular Fractions of the Liquid Biopsy in a Cohort of Breast Cancer Patients. *The Journal Of Molecular Diagnostics/The Journal Of Molecular Diagnostics*, 22(3), 319-337. <https://doi.org/10.1016/j.jmoldx.2019.11.006>
- Shum, B. O., Pretorius, C. J., Sng, L. M. F., Henner, I., Barahona, P., Başar, E., McGill, J., Wilgen, U., Zournazi, A., Downie, L., Taylor, N., Cheney, L., Wu, S., Twine, N. A., Bauer, D. C., Watts, G. F., Navilebasappa, A., Kumar, K. R., Ungerer, J. P., & Bennett, G. (2023). Feasibility of Targeted Next-Generation DNA Sequencing for Expanding Population Newborn Screening. *Clinical Chemistry*, 69(8), 890-900. <https://doi.org/10.1093/clinchem/hvad066>
- Siphy. (2010, 25 mayo). @Broadinstitute. <https://www.broadinstitute.org/scientific-community/software/siphy>
- SoRelle, J. A., Wachsmann, M. B., & Cantarel, B. L. (2020). Assembling and Validating Bioinformatic Pipelines for Next-Generation Sequencing Clinical Assays. *Archives Of Pathology & Laboratory Medicine*, 144(9), 1118-1130. <https://doi.org/10.5858/arpa.2019-0476-ra>

- Sorrentino, E., Albion, E., Modena, C., Daja, M., Cecchin, S., Paolacci, S., Miertuš, J., Bertelli, M., Maltese, P. E., Chiurazzi, P., Stuppa, L., Colombo, L., & Marceddu, G. (2022). PacMAGI: A pipeline including accurate indel detection for the analysis of PacBio sequencing data applied to RPE65. *Gene*, 832, 146554. <https://doi.org/10.1016/j.gene.2022.146554>
- Su, J., Lui, W. W., Lee, Y., Zheng, Z., Siu, G. K., Ng, T. T., Zhang, T., Lam, T. T., Lao, H., Yam, W., Tam, K., Leung, K., Lam, T., Leung, A. W., & Luo, R. (2023). Evaluation of *Mycobacterium tuberculosis* enrichment in metagenomic samples using ONT adaptive sequencing and amplicon sequencing for identification and variant calling. *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-32378-x>
- Svensson, D., Sjögren, R., Sundell, D., Sjödin, A., & Trygg, J. (2019). doepipeline: a systematic approach to optimizing multi-level and multi-step data processing workflows. *BMC Bioinformatics*, 20(1). <https://doi.org/10.1186/s12859-019-3091-z>
- Takashima, T., Brisset, S., Furukawa, A., Taniguchi, H., Takeyasu, R., Kawamura, A., & Tamura, Y. (2021). Case Report: BMPR2-Targeted MinION Sequencing as a Tool for Genetic Analysis in Patients With Pulmonary Arterial Hypertension. *Frontiers In Cardiovascular Medicine*, 8. <https://doi.org/10.3389/fcvm.2021.711694>
- Tan, Y., Michaeel, A., Blumenfeld, J., Donahue, S., Parker, T., Levine, D., & Rennert, H. (2012). A Novel Long-Range PCR Sequencing Method for Genetic Analysis of the Entire PKD1 Gene. *The Journal Of Molecular Diagnostics/The Journal Of Molecular Diagnostics*, 14(4), 305-313. <https://doi.org/10.1016/j.jmoldx.2012.02.007>
- Tang, H., Kirkness, E. F., Lippert, C., Biggs, W. H., Fabani, M. M., Guzmán, E., Ramakrishnan, S., Lavrenko, V., Kakaradov, B., Hou, C., Hicks, B., Heckerman, D., Och, F. J., Caskey, C. T., Venter, J. C., & Telenti, A. (2017). Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *American Journal Of Human Genetics*, 101(5), 700-715. <https://doi.org/10.1016/j.ajhg.2017.09.013>
- Tom, N., Tom, O., Malčíková, J., Pavlová, Š., Kubešová, B., Rausch, T., Kolařík, M., Beneš, V., Bystrý, V., & Posplíšilová, Š. (2018). ToTem: a tool for variant calling pipeline optimization. *BMC Bioinformatics*, 19(1). <https://doi.org/10.1186/s12859-018-2227-x>
- Torres MJ, Rodríguez JC, Hernández CR, Anabitarte A, Caballero A, Vázquez C, Fernández-Burriel M, Pérez P, Palop L. (2006). Diagnóstico molecular de la Poliquistosis Renal Autosómica Dominante en la Comunidad Autónoma de Canarias. *Nefrologia*, 26, 666-672. Disponible: <https://www.revistanefrologia.com/es-pdf-X0211699506020582>

Tunsjø, H. S., Ullmann, I. F., & Charnock, C. (2023). A preliminary study of the use of MinION sequencing to specifically detect Shiga toxin-producing Escherichia coli in culture swipes containing multiple serovars of this species. *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-35279-1>

Vogeley, C., Nguyen, T., Woeste, S., Krutmann, J., Haarmann-Stemmann, T., & Rossi, A. (2022). Rapid and simple analysis of short and long sequencing reads using DuesselporeTM. *Frontiers In Genetics*, 13. <https://doi.org/10.3389/fgene.2022.931996>

What are single nucleotide polymorphisms (SNPs)? | MedlinePlus Genetics. Accedido el 11 de abril del 2024, desde <https://medlineplus.gov/genetics/understanding/genomicresearch/snp/>

Weilguny, L., De Maio, N., Munro, R., Manser, C., Birney, E., Loose, M., & Goldman, N. (2023). Dynamic, adaptive sampling during nanopore sequencing using Bayesian experimental design. *Nature Biotechnology (Print)*, 41(7), 1018-1025. <https://doi.org/10.1038/s41587-022-01580-z>

Weißbach, S., Sys, S., Hewel, C., Todorov, H., Schweiger, S., Winter, J., Pfenninger, M., Torkamani, A., Evans, D., Bürger, J., Everschor-Sitte, K., May-Simera, H., & Gerber, S. (2021). Reliability of genomic variants across different next-generation sequencing platforms and bioinformatic processing pipelines. *BMC Genomics*, 22(1). <https://doi.org/10.1186/s12864-020-07362-8>

Wickland, D. P., Battu, G., Hudson, K. A., Diers, B. W., & Hudson, M. E. (2017). A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinformatics*, 18(1). <https://doi.org/10.1186/s12859-017-2000-6>

Wong, K., Pitcher, D., Braddon, F., Downward, L., Steenkamp, R., Annear, N., Barratt, J., Bingham, C., Chrysochou, C., Coward, R. J. M., Game, D., Griffin, S., Hall, M., Johnson, S., Kanigicherla, D., Frankl, F. K., Kavanagh, D., Kerecuk, L., Maher, E. R., . . . Wright, D. (2024). Effects of rare kidney diseases on kidney failure: a longitudinal analysis of the UK National Registry of Rare Kidney Diseases (RaDaR) cohort. *Lancet*. [https://doi.org/10.1016/s0140-6736\(23\)02843-x](https://doi.org/10.1016/s0140-6736(23)02843-x)

Xiao, T., & Zhou, W. (2020). The third generation sequencing: the advanced approach to genetic diseases. *Translational Pediatrics (Print)*, 9(2), 163-173. <https://doi.org/10.21037/tp.2020.03.06>

Xie, Y., Zhong, Y., Chang, J., & Kwan, H. S. (2021). Chromosome-level de novo assembly of Coprinopsis cinerea A43mut B43mut pab1-1 #326 and genetic variant identification of mutants using Nanopore MinION sequencing. *Fungal Genetics And Biology*, 146, 103485. <https://doi.org/10.1016/j.fgb.2020.103485>

Yeo, Z. X., Wong, J. C. L., Rozen, S., & Lee, A. S. G. (2014). Evaluation and optimisation of indel detection workflows for ion torrent sequencing of the BRCA1 and BRCA2 genes. *BMC Genomics*, 15(1). <https://doi.org/10.1186/1471-2164-15-516>

ANEXO

Anexo A

A continuación se presenta la extracción de la información según los campos mencionados en el punto 3.2.4 desde la Tabla 32 a la Tabla 83.

Tabla 32: Extracción de información de la Revisión Sistemática - Artículo 1

DOI y año	10.1038/s41525-024-00394-z, 2024
Nombre de la publicación	<i>Assessing the efficacy of target adaptive sampling long-read sequencing through hereditary cancer patient genomes</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Diagnóstico de enfermedad
Tipo de secuenciador que se usó	MinION, Illumina
Herramienta y usos	<ul style="list-style-type: none"> - <i>Basecalling</i>: Guppy - Mapeo: SAMtools, Minimap2, BWA - <i>Variant calling</i>: PEPPER-Margin-DeepVariant, GLIMPSE, GATK - <i>SV calling</i>: Nanomonsv - Clasificación de variantes: PLINK, WhatsHap, <i>Variant Effect Predictor</i> (VEP), LOFTEE, SpliceAI, CADD - Base de datos: ClinVar, SAVNet, gnomAD, ToMMo 14JPN
Hallazgos extras	<ul style="list-style-type: none"> - Se encontraron herramientas extras que sirven para la manipulación de los archivos. - SAMtools (Manipulación de archivos SAM y BAM) - bcftools (Manipulación de archivos VCF) - GWAS (Estadísticas) - f5c (methylation calling) - SAMtools (Convertir SAM a BAM)

Tabla 33: Extracción de información de la Revisión Sistemática - Artículo 2

DOI y año	10.1128/jcm.01576-23, 2024
Nombre de la publicación	<i>Closing the gap: Oxford Nanopore Technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Detección de bacteria causante de enfermedades
Tipo de secuenciador que se usó	MinION, Illumina
Herramienta y usos	<ul style="list-style-type: none"> - <i>Basecalling</i>: Dorado, Guppy - Control de calidad: Samtools - <i>Trimming</i> y filtrado: Trimmomatic, seqkit

	<ul style="list-style-type: none"> - Mapeo: Bowtie, Minimap2, BWA - <i>Variant calling</i>: Medaka, bcftools
Hallazgos extras	<ul style="list-style-type: none"> - Se comparó el rendimiento del secuenciador MinION con dos tipos de <i>flowcells</i> (R9 y R10) donde se concluyó la mejora que presentaba la nueva <i>flowcells</i> R10 en cuanto a la calidad de las lecturas, acercándose a la calidad generada por Illumina.

Tabla 34: Extracción de información de la Revisión Sistemática - Artículo 3

DOI y año	10.1093/jac/dkae060, 2024
Nombre de la publicación	<i>PHARE: a bioinformatics pipeline for compositional profiling of multiclonal Plasmodium falciparum infections from long-read Nanopore Sequencing data</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del <i>pipeline</i>	Otro
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - <i>Basecalling</i>: Dorado, Guppy - <i>Trimming</i> y filtrado: Filtlong - Mapeo: Minimap2 - <i>Variant calling</i>: Longshot, NanoSNP, Clair
Hallazgos extras	<ul style="list-style-type: none"> - Herramienta para la conversión de archivos FAST5 a POD5 (<i>POD5 package</i>) - Samtools (Manipulación de archivos SAM)

Tabla 35: Extracción de información de la Revisión Sistemática - Artículo 4

DOI y año	10.1038/s41598-023-35279-1, 2023
Nombre de la publicación	<i>A preliminary study of the use of MinION sequencing to specifically detect Shiga toxin-producing Escherichia coli in culture swipe containing multiple serovars of this species</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del <i>pipeline</i>	Detección de bacteria causante de enfermedades
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - <i>Basecalling</i>: Albacore - <i>Trimming</i> y filtrado: PoreChop

Tabla 36: Extracción de información de la Revisión Sistemática - Artículo 5

DOI y año	10.1099/mgen.0.001013, 2023
Nombre de la publicación	<i>Fast and cost-effective SARS-CoV-2 variant detection using OxfordNanopore full-length spike gene sequencing</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del <i>pipeline</i>	Detección de virus
Tipo de secuenciador que se usó	Nanoporo
Herramienta y usos	<ul style="list-style-type: none"> - <i>Basecalling</i>: Guppy - <i>Variant calling</i>: Nanopolish

Tabla 37: Extracción de información de la Revisión Sistemática - Artículo 6

DOI y año	10.1186/s13073-023-01240-0, 2023
Nombre de la publicación	<i>Structural and non-coding variants increase the diagnostic yield of clinical whole genome sequencing for rare diseases</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del <i>pipeline</i>	Diagnóstico de enfermedad
Tipo de secuenciador que se usó	Illumina Hiseq
Herramienta y usos	<ul style="list-style-type: none"> - <i>Variant calling</i>: DeepVariant, bcftools - SV y CNV <i>calling</i>: Lumpy, CNVnator, svtools, SVRare, ExpansionHunter - Clasificación de variantes: SnpEFF, CADD, DANN, REVEL, FATHMM, ncER, ReMM, spliceAI, MaxEntScan, vcfanno - Base de datos: gnomad, 1000Genome, ClinVar, dbVAR, DECIPHER, GREEN-VARAN v1.0, GREEN-DB v.2.5
Hallazgos extras	GLNexus (v1.2.6) (fusión de archivos VCF)

Tabla 38: Extracción de información de la Revisión Sistemática - Artículo 7

DOI y año	10.1186/s13059-023-02923-y, 2023
Nombre de la publicación	<i>Transformation of alignment files improves performance of variant callers for long-read RNA sequencing data</i>
Tipo de artículo	Comparación de herramientas
Objetivo del pipeline	Detección de variantes
Tipo de secuenciador que se usó	Illumina, PacBio, Nanoporo
Herramienta y usos	<ul style="list-style-type: none"> - <i>Trimming</i> y filtrado: samtools - Alineamiento: Minimap2, STAR - <i>Variant calling</i>: GATK, bcftools, FreeBayes, Platypus, DeepVariant, Clair3, NanoCaller, Longshot, WhatsHap - <i>SV calling</i>: pbsv
Hallazgos extras	El artículo hace una diferencia entre herramientas de <i>variant calling</i> comúnmente utilizadas para NGS (GATK, bcftools, FreeBayes, Platypus) y utilizadas por secuenciadores de nanoporos (DeepVariant, Clair3, NanoCaller, Longshot)

Tabla 39: Extracción de información de la Revisión Sistemática - Artículo 8

DOI y año	10.1038/s41587-022-01580-z, 2023
Nombre de la publicación	<i>Dynamic, adaptive sampling during nanopore sequencing using Bayesian Experimental design</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Detección de variantes
Tipo de secuenciador que se usó	Nanoporo (GridION)
Herramienta y usos	<ul style="list-style-type: none"> - <i>Trimming</i> y filtrado: Picard - Mapeo: Minimap2, Samtools - <i>Variant calling</i>: Freebayes, Medaka - Clasificación de variantes: VCFLib
Hallazgos extras	bcftools se utilizó para manipular archivos VCF.

Tabla 40: Extracción de información de la Revisión Sistemática - Artículo 9

DOI y año	10.1101/s10096-023-04590-0, 2023
Nombre de la publicación	<i>A comparison of five Illumina, Ion Torrent, and nanopore sequencing technology-based approaches for whole genome sequencing of SARS-CoV-2</i>
Tipo de artículo	Comparación de <i>pipelines</i>
Objetivo del pipeline	Detección de virus
Tipo de secuenciador que se usó	Ion Torrent, NGS, MinION
Herramienta y usos	<ul style="list-style-type: none"> - <i>Trimming</i> y filtrado: Trimmomatic, Cutadapt - Mapeo: BWA, Bowtie2, SAMtools, Minimap2 - <i>Variant calling</i>: BCFtools
Hallazgos extras	Minimap2 está diseñado para el mapeo de secuencias, de dispositivos con una alta tasa de error.

Tabla 41: Extracción de información de la Revisión Sistemática - Artículo 10

DOI y año	10.1101/clinchem/hvad066, 2023
Nombre de la publicación	<i>Feasibility of Targeted Next-Generation DNA Sequencing for Expanding Population Newborn Screening</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Diagnóstico de enfermedades
Tipo de secuenciador que se usó	NGS
Herramienta y usos	<ul style="list-style-type: none"> - <i>Variant calling</i>: GATK - Clasificación de variantes: Plink - Base de datos: ClinVar, GnomAD
Hallazgos extras	En el artículo se menciona que el American College of Medical Genetics and Genomics (ACMG) describe ciertos criterios para la clasificación manual de variantes genéticas.

Tabla 42: Extracción de información de la Revisión Sistemática - Artículo 11

DOI y año	10.1038/s41598-023-32378-x, 2023
Nombre de la publicación	<i>Evaluation of Mycobacterium tuberculosis enrichment in metagenomic samples using ONT adaptive sequencing and amplicon sequencing for identification and variant calling</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Detección de bacteria, causante de enfermedades
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - Control de calidad: FastQC - <i>Trimming</i> y filtrado: Nanofilt - Alineamiento: Minimap2 - <i>Variant calling</i>: Clair3
Hallazgos extras	La precisión y el f1-score del llamado de variantes fueron relativamente altas, pero aun así el estudio muestra cómo el uso de la herramienta “ <i>readfish</i> ”, para la etapa de <i>basecalling</i> , puede mejorar significativamente el resultado de las etapas posteriores

Tabla 43: Extracción de información de la Revisión Sistemática - Artículo 12

DOI y año	10.1371/journal.pone.0265622, 2022
Nombre de la publicación	<i>Genotyping of familial Mediterranean fever gene (MEFV)-Polynucleotide polymorphism-Comparison of Nanopore with conventional Sanger Sequencing.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Diagnóstico de enfermedad
Tipo de secuenciador que se usó	MinION, Sanger
Herramienta y usos	<ul style="list-style-type: none"> - <i>Basecalling</i>: Guppy - Control de calidad: pycoQC - <i>Trimming</i> y filtrado: NanoFilt - Mapeo: Minimap2, Samtools - <i>Variant calling</i>: bcftools - Clasificación de variantes: bedtools, ANNOVAR
Hallazgos extras	<p>Se detectaron 433 SNPs de los cuales 284 se consideraron heterocigotas y 149 homocigotas. Estas variantes fueron confirmadas por Sanger. Además ONT encontró variantes en 2 pacientes en la zona UTR que no fueron detectadas inicialmente por Sanger.</p> <p>Se utilizan herramientas para visualizar los datos como <i>Integrative Genomics Viewer</i>.</p>

Tabla 44: Extracción de información de la Revisión Sistemática - Artículo 13

DOI y año	10.1016/j.csbj.2022.09.024, 2022
Nombre de la publicación	<i>NanoRTax, a real-time pipeline for taxonomic and diversity analysis of nanopore 16S rRNA amplicon sequencing data.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Otro
Tipo de secuenciador que se usó	Nanoporo
Herramienta y usos	<ul style="list-style-type: none"> - Control de calidad: Fastp

Tabla 45: Extracción de información de la Revisión Sistemática - Artículo 14

DOI y año	10.3390/plants11162103, 2022
Nombre de la publicación	<i>A Pipeline NanoTRF as a New Tool for De Novo Satellite DNA Identification in the Raw Nanopore Sequencing Reads of Plant Genomes.</i>
Tipo de artículo	Presentación de herramienta(s)
Objetivo del pipeline	Otro
Tipo de secuenciador que se usó	MinION, Illumina
Herramienta y usos	<ul style="list-style-type: none"> - <i>Basecalling</i>: Guppy - Control de calidad: FastQC - <i>Trimming</i> y filtrado: Trimmomatic
Hallazgos extras	NanoTRF (identificación repetida de TR)

Tabla 46: Extracción de información de la Revisión Sistemática - Artículo 15

DOI y año	10.1038/s41587-022-01221-5, 2022
Nombre de la publicación	<i>Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del <i>pipeline</i>	Diagnóstico de enfermedad
Tipo de secuenciador que se usó	Nanoporo
Herramienta y usos	<ul style="list-style-type: none"> - <i>Basecalling</i>: Guppy - Mapeo: Minimap2 - <i>Variant calling</i>: PEPPER-Margin-DeepVariant - <i>SV calling</i>: Sveval, Sniffles, truvari - Clasificación de variantes: Gencode - Bases de datos: Database of Genomic Variants, dbVar, ClinVar, HGMD, OMIM, gnomAD
Hallazgos extras	Samtools (división en fragmentos de contig)

Tabla 47: Extracción de información de la Revisión Sistemática - Artículo 16

DOI y año	10.1093/bioinformatics/btac043, 2022
Nombre de la publicación	<i>ONTdeCIPHER: an amplicon-based nanopore sequencing pipeline for tracking pathogenic variants.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del <i>pipeline</i>	Detección de variantes
Tipo de secuenciador que se usó	Nanoporo
Herramienta y usos	<ul style="list-style-type: none"> - <i>Trimming</i> y filtrado: SeqKit - Control de calidad: pycoQC, MultiQC, DeepTools - Mapeo: Minimap2 - <i>Variant calling</i>: Medaka - <i>SV calling</i>: Sniffles - Clasificación de variantes: SnpEff

Tabla 48: Extracción de información de la Revisión Sistemática - Artículo 17

DOI y año	10.1186/s12920-022-01190-3, 2022
Nombre de la publicación	<i>ECNano: A cost-effective workflow for target enrichment sequencing and accurate variant calling on 4800 clinically significant genes using a single MinION flowcell.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Diagnóstico de enfermedades
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - Trimming y filtrado: Porechop, SAMtools - Mapeo: Minimap2 - Clasificación de variantes: VariantBAM - Variant calling: LongShot, Medaka, Clair, Clair-ensemble - Base de datos: ClinVar
Hallazgos extras	<ul style="list-style-type: none"> - Clair-ensemble se optimizó para el <i>variant calling</i> de secuencias <i>long-reads</i>. - Clair-ensemble obtuvo una mayor precisión y <i>f1-score</i> que las demás herramientas de <i>variant calling</i>.

Tabla 49: Extracción de información de la Revisión Sistemática - Artículo 18

DOI y año	10.1016/j.gene.2022.146554, 2022
Nombre de la publicación	<i>PacMAGI: A pipeline including accurate indel detection for the analysis of PacBio sequencing data applied to RPE65.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	General
Tipo de secuenciador que se usó	PacBio
Herramienta y usos	<ul style="list-style-type: none"> - Control de calidad: FastQC, MultiQC, Qualimap, Picard - Mapeo: Pbmm2 - Variant calling: DeepVariant, Longshot - Clasificación de variantes: VEP, VarSome. - Base de datos: APPRIS
Hallazgos extras	<ul style="list-style-type: none"> - Pbmm2 está basada en Minimap2, pero optimizada para secuenciadores PacBio. - LongShot utiliza modelo de Markov oculto, para detectar variantes.

Tabla 50: Extracción de información de la Revisión Sistemática - Artículo 19

DOI y año	10.3390/ijms232415845, 2022
Nombre de la publicación	<i>Application of Long-Read Nanopore Sequencing to the Search for Mutations in Hypertrophic Cardiomyopathy.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Diagnóstico de enfermedades
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: Minimap2 - <i>Variant calling</i>: Clair3, Medaka. - Bases de datos: <i>1000Genome</i>, <i>Exome Aggregation Consortium</i>, dbSNP, HGMD - Clasificación de variantes: Annovar, CADD, SIFT, <i>Mutation Taster</i>, MutPred
Hallazgos extras	Nanopolish: Corrección de errores en las lecturas y <i>variant calling</i>

Tabla 51: Extracción de información de la Revisión Sistemática - Artículo 20

DOI y año	10.1016/j.jmoldx.2022.06.006, 2022
Nombre de la publicación	<i>NGS4THAL, a One-Stop Molecular Diagnosis and Carrier Screening Tool for Thalassemia and Other Hemoglobinopathies by Next-Generation Sequencing.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Diagnóstico de enfermedades
Tipo de secuenciador que se usó	NGS
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: BWA - <i>Variant calling</i>: GATK, Pindel - SV/CNV <i>calling</i>: BreakDancer, CoNIFER
Hallazgos extras	<ul style="list-style-type: none"> - <i>BreakDancer</i> está optimizada para trabajar con datos de secuenciadores de nueva generación (NGS). - Pysam: Realizar operaciones en archivos BAM

Tabla 52: Extracción de información de la Revisión Sistemática - Artículo 21

DOI y año	10.3389/fgene.2022.931996, 2022
Nombre de la publicación	<i>Rapid and simple analysis of short and long sequencing reads using DuesselporeTM.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del <i>pipeline</i>	Otro
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - Control de calidad: FastQC - Mapeo: Minimap2

Tabla 53: Extracción de información de la Revisión Sistemática - Artículo 22

DOI y año	10.3390/genes13091583, 2022
Nombre de la publicación	<i>Evaluation of the Available Variant Calling Tools for Oxford Nanopore Sequencing in Breast Cancer.</i>
Tipo de artículo	Comparación de herramientas
Objetivo del <i>pipeline</i>	Diagnóstico de enfermedades
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - Control de calidad: Bedtools - Mapeo: Minimap2 - Trimming y filtrado: Samtools - Variant calling: Human-SNP-wf, Clair3, Clair, NanoCaller, Longshot, Medaka - Clasificación de variantes: SnpEff - Base de datos: ClinVar
Hallazgos extras	<ul style="list-style-type: none"> - wf-Human-SNP obtuvo en promedio un mayor <i>f1-score</i> en la detección de SNPs e indels - Clair y Medaka utilizan deep learning para el llamado de variantes. - Longshot utiliza un modelo de Markov oculto para la búsqueda de variantes. - Nanocaller utiliza una red neuronal convolucional para la búsqueda de variantes

Tabla 54: Extracción de información de la Revisión Sistemática - Artículo 23

DOI y año	10.1093/gigascience/giab063, 2021
Nombre de la publicación	<i>Vulcan: Improved long-read mapping and structural variant calling via dual-mode alignment.</i>
Tipo de artículo	Presentación de herramienta(s)
Objetivo del pipeline	Detección de variantes
Tipo de secuenciador que se usó	Nanoporos (PromethION), PacBio
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: NGMLR, Minimap2, Vulcan, WinnowMap, MashMap, LAST, LRA, Duplomap, GraphMap - SV calling: Sniffles
Hallazgos extras	<ul style="list-style-type: none"> - El artículo menciona herramientas para la simulación de variantes estructurales y cobertura de las lecturas con el fin de validar el <i>pipeline</i> propuesto (SURVIVOR, simSV y Nanosim-h)

Tabla 55: Extracción de información de la Revisión Sistemática - Artículo 24

DOI y año	10.1016/j.jgar.2021.04.004. 2021
Nombre de la publicación	<i>Genomic and phylogenetic analysis of a multidrug-resistant Burkholderia contaminans strain isolated from a patient with ocular infection.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Ensamblaje de genoma
Tipo de secuenciador que se usó	Illumina, Nanoporos
Herramienta y usos	<ul style="list-style-type: none"> - Control de calidad: FastQC, MultiQC - Trimming y filtrado: Fastp

Tabla 56: Extracción de información de la Revisión Sistemática - Artículo 25

DOI y año	10.1261/rna.078800.121, 2021
Nombre de la publicación	<i>FLAME: long-read bioinformatics tool for comprehensive spliceosome characterization.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del <i>pipeline</i>	Otro
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - <i>Basecalling</i>: Guppy - Mapeo: Minimap2, Samtools

Tabla 57: Extracción de información de la Revisión Sistemática - Artículo 26

DOI y año	10.1038/s41592-021-01299-w, 2021
Nombre de la publicación	<i>Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enable high accuracy in nanopore long-reads</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del <i>pipeline</i>	Detección de variantes
Tipo de secuenciador que se usó	Nanoporos, PacBio, Illumina
Herramienta y usos	<ul style="list-style-type: none"> - <i>Basecalling</i>: Guppy - Mapeo: Minimap2, pbmm2 - <i>Variant calling</i>: Medaka, Clair, longshot, PEPPER-Margin-DeepVariant, DeepVariant, WhatsHap
Hallazgos extras	<ul style="list-style-type: none"> - Presenta una combinación de herramientas (PEPPER-Margin-DeepVariant) para la llamada de variantes para mejorar los resultados de esta etapa. - Samtools (manipulación de archivos SAM)

Tabla 58: Extracción de información de la Revisión Sistemática - Artículo 27

DOI y año	10.1016/j.fgb.2020.103485, 2021
Nombre de la publicación	<i>Chromosome-level de novo assembly of Coprinopsis cinerea A43mut B43mut pab1-1 #326 and genetic variant identification of mutants using Nanopore MinION sequencing.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Ensamblaje de genoma
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - Control de calidad: MinIONQC - Mapeo: Minimap2, MUMmer - Variant calling: BcfTools - SV calling: Sniffles
Hallazgos extras	MinIONQC : Paquete de R, para evaluar la calidad de las lecturas

Tabla 59: Extracción de información de la Revisión Sistemática - Artículo 28

DOI y año	10.1093/bioinformatics/btab516, 2021
Nombre de la publicación	<i>Founder reconstruction enables scalable and seamless pangenomic analysis</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Detección de variantes
Tipo de secuenciador que se usó	No se especifica
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: BWE, Bowtie2 - Variant calling: GATK, BCFtools
Hallazgos extras	En el artículo se presenta un pipeline con alineamiento múltiple para la detección de variantes, aun así los resultados son mínimamente inferiores a los obtenidos con un alineamiento normal.

Tabla 60: Extracción de información de la Revisión Sistemática - Artículo 29

DOI y año	10.1186/s12859-021-04311-4, 2021
Nombre de la publicación	<i>HELLO: improved neural network architectures and methodologies for small variant calling</i>
Tipo de artículo	Comparación de herramientas
Objetivo del pipeline	Detección de variantes
Tipo de secuenciador que se usó	NGS, PacBio
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: BWA - <i>Variant calling</i>: HELLO, DeepVariant, GATK
Hallazgos extras	<ul style="list-style-type: none"> - DeepVariant funciona mediante una CNN, que viene pre entrenada, pero también se puede re-entrenar con datos más personalizados. - HELLO, desarrollada en el propio artículo, se basa en una red neuronal profunda, para la predicción de variantes. además fue la herramienta que obtuvo los mejores resultados en cuanto a precisión y <i>recall</i>

Tabla 61: Extracción de información de la Revisión Sistemática - Artículo 30

DOI y año	10.3389/fcvm.2021.711694, 2021
Nombre de la publicación	<i>Case Report: BMPR2-Targeted MinION Sequencing as a Tool for Genetic Analysis in Patients With Pulmonary Arterial Hypertension.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Diagnóstico de enfermedades
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - Control de calidad: FASTQC. - <i>Trimming</i> y filtrado: Nanofilt - Mapeo: Minimap2 - <i>Variant calling</i>: Varscan - Base de datos: ClinVar
Hallazgos extras	Samtools: Conversión de formato

Tabla 62: Extracción de información de la Revisión Sistemática - Artículo 31

DOI y año	10.1186/s12864-020-07362-8, 2021
Nombre de la publicación	<i>Reliability of genomic variants across different next-generation sequencing platforms and bioinformatic processing pipelines.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Detección de variantes
Tipo de secuenciador que se usó	NGS
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: BWA. - Variant calling: GATK

Tabla 63: Extracción de información de la Revisión Sistemática - Artículo 32

DOI y año	10.1038/s42003-020-01270-z, 2020
Nombre de la publicación	<i>Genopo: a nanopore sequencing analysis toolkit for portableAndroid devices.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Detección de virus
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: Minimap2 - Variant calling: Bcftools, Nanopolish
Hallazgos extras	<ul style="list-style-type: none"> - Bioawk (manipulacion de archivos) - Samtools (Manipular archivos SAM) - Bedtools (Manipular archivos BED) - f5c (<i>Methylation calling</i>)

Tabla 64: Extracción de información de la Revisión Sistemática - Artículo 33

DOI y año	10.3390/ijms21239177, 2020
Nombre de la publicación	<i>A Long-Read Sequencing Approach for Direct Haplotype Phasing in Clinical Settings</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del <i>pipeline</i>	Detección de variantes
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - <i>Trimming</i> y filtrado: NanoFilt - Mapeo: BWA, Minimap2 - <i>Variant calling</i>: Medaka

Tabla 65: Extracción de información de la Revisión Sistemática - Artículo 34

DOI y año	10.1016/j.jmoldx.2019.12.006, 2020
Nombre de la publicación	<i>Rapid Next-Generation Sequencing-Based Diagnostics of Bacteremia in Septic Patients.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del <i>pipeline</i>	Detección de bacteria, causante de enfermedades
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - Control de calidad: NanoPlot - <i>Trimming</i> y filtrado: Porechop, BBMap - Mapeo: BWA
Hallazgos extras	SAMtools: Conversión de formato SAM a BAM

Tabla 66: Extracción de información de la Revisión Sistemática - Artículo 35

DOI y año	10.1016/j.jmoldx.2019.11.006, 2020
Nombre de la publicación	<i>Preanalytical Variables for the Genomic Assessment of the Cellular and Acellular Fractions of the Liquid Biopsy in a Cohort of Breast Cancer Patients</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Detección de variantes
Tipo de secuenciador que se usó	NGS
Herramienta y usos	<ul style="list-style-type: none"> - <i>Trimming</i> y filtrado / <i>Variant calling</i>: SAMtools - Mapeo: Bowtie, BWA - Clasificación de variantes: GENCODE - Base de datos: dbSNP

Tabla 67: Extracción de información de la Revisión Sistemática - Artículo 36

DOI y año	10.3389/fmicb.2020.01883, 2020
Nombre de la publicación	<i>Analytical Performance Validation of Next-Generation Sequencing Based Clinical Microbiology Assays Using a K-mer Analysis Workflow.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	General
Tipo de secuenciador que se usó	NGS
Herramienta y usos	<ul style="list-style-type: none"> - <i>Trimming</i> y filtrado: Trimmomatic - Mapeo: BWA, Minimap2 - <i>Variant calling</i>: Bcftool, Htslib, Snippy, Freebayes - Clasificación de variantes: Vcftools
Hallazgos extras	Picard: Manipulación de archivos SAM, BAM y VCF

Tabla 68: Extracción de información de la Revisión Sistemática - Artículo 37

DOI y año	10.5858/arpa.2019-0476-RA, 2020
Nombre de la publicación	<i>Assembling and Validating Bioinformatic Pipelines for Next-Generation Sequencing Clinical Assays.</i>
Tipo de artículo	Presentación de herramienta(s)
Objetivo del pipeline	Diagnóstico de enfermedades
Tipo de secuenciador que se usó	NGS
Herramienta y usos	<ul style="list-style-type: none"> - <i>Trimming and filtering:</i> Picard, Samtools - <i>Mapeo:</i> Bowtie2, BWA - <i>Variant calling:</i> GATK, Platypus, Strelka, Freebayes, MuTect2, VarScan, EBCall, Virmid, Shimmer - <i>SV / CNV calling:</i> VarScan2, CNVKit, Control-FREEC, OncoCNV, Pindel, Lumpy, Delly - Clasificación de variantes: vcftools, SnpSift, FATHMM, KGGSeq - Bases de datos: OMIM, ClinVar, GnomAD
Hallazgos extras	<ul style="list-style-type: none"> - En el artículo se menciona que BWA tiene una mayor precisión en el alineamiento que Bowtie2. - Según el artículo EBCall, Mutect2, Virmid y Strelka han obtenido mejores resultados en detección de SNPs, mientras que EBCall en la detección de indels. - Cerebro: Detección de errores en las variantes llamadas

Tabla 69: Extracción de información de la Revisión Sistemática - Artículo 38

DOI y año	10.1101/gr.244939.118, 2019
Nombre de la publicación	<i>Structural variants identified by Oxford Nanopore PromethION Sequencing of the human genome.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Detección de variantes
Tipo de secuenciador que se usó	Nanoporos
Herramienta y usos	<ul style="list-style-type: none"> - <i>Basecalling</i>: Guppy - Control de calidad: NanoPack, Mosdepth - Mapeo: NGMLR, LAST, Minimap2, BWA - Clasificación de variantes: vcfanno - SV <i>calling</i>: Sniffles, NanoSV, pbsv, SVIM, npInv, Manta, LUMPY
Hallazgos extras	<ul style="list-style-type: none"> - VCFtools, BCFtools (procesar archivos VCF) - SURVIVOR (combinar archivos con las llamadas de variantes estructurales) - Se distinguen entre los SV <i>calling</i>, para lecturas cortas (Manta, LUMPY) y los especializados en lecturas largas (Sniffles, NanoSV, pbsv, SVIM, npInv)

Tabla 70: Extracción de información de la Revisión Sistemática - Artículo 39

DOI y año	10.1093/gigascience/giz104, 2019
Nombre de la publicación	<i>Comparison of single-nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin-producing Escherichia coli.</i>
Tipo de artículo	Comparación de <i>pipelines</i>
Objetivo del pipeline	Detección de bacteria causante de enfermedades
Tipo de secuenciador que se usó	MinION, NGS
Herramienta y usos	<ul style="list-style-type: none"> - Control de calidad: Nanoplot - <i>Trimming</i> y filtrado: Trimmomatic, Porechop, Filtlong - Mapeo: BWA, Minimap2 - <i>Variant calling</i>: GATK

Tabla 71: Extracción de información de la Revisión Sistemática - Artículo 40

DOI y año	10.3389/fmicb.2019.00179, 2019
Nombre de la publicación	<i>Cross-Border Transmission of <i>Salmonella Choleraesuis</i> var. <i>Kunzendorf</i> in European Pigs and Wild Boar: Infection, Genetics, and Evolution</i>
Tipo de artículo	<i>Presentación de pipeline.</i>
Objetivo del pipeline	Otro
Tipo de secuenciador que se usó	NGS
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: BWA - Variant calling: SAMtools

Tabla 72: Extracción de información de la Revisión Sistemática - Artículo 41

DOI y año	10.1186/s12859-019-3091-z, 2019
Nombre de la publicación	<i>Doepipeline: a systematic approach to optimizing multi-level and multi-step data processing workflows</i>
Tipo de artículo	Optimización de parámetros
Objetivo del pipeline	Otro
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: BWA - Variant calling: GATK
Hallazgos extras	PICARD: Conversión de formatos

Tabla 73: Extracción de información de la Revisión Sistemática - Artículo 42

DOI y año	10.1186/s12859-019-2965-4, 2019
Nombre de la publicación	<i>iMAP: an integrated bioinformatics and visualization pipeline for microbiome data analysis</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Otro
Tipo de secuenciador que se usó	NGS
Herramienta y usos	<ul style="list-style-type: none"> - Control de calidad: FASTQC - Mapeo: MAFFT
Hallazgos extras	MAFFT es una herramienta para alineamiento múltiple, pero también puede ser ocupada para realizar alineamiento con una secuencia de referencia

Tabla 74: Extracción de información de la Revisión Sistemática - Artículo 43

DOI y año	10.1186/s12864-019-5559-7, 2019
Nombre de la publicación	<i>mGAP: the macaque genotype and phenotype resource, a framework for accessing and interpreting macaque variant data, and identifying new models of human disease</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Otro
Tipo de secuenciador que se usó	NGS
Herramienta y usos	<ul style="list-style-type: none"> - Clasificación de variantes: PolyPhen2, SnpEff - Base de datos: dbSNP, dbVar, ClinVar
Hallazgos extras	PolyPhen2 predice el efecto que pueda tener una variante en el funcionamiento de las proteínas

Tabla 75: Extracción de información de la Revisión Sistemática - Artículo 44

DOI y año	10.1186/s12859-018-2227-x, 2018
Nombre de la publicación	<i>ToTem: a tool for variant calling pipeline optimization</i>
Tipo de artículo	Presentación de herramienta
Objetivo del pipeline	No aplica
Tipo de secuenciador que se usó	Illumina
Herramienta y usos	<ul style="list-style-type: none"> - Alineamiento: BWA - Variant calling: DeepSNV, Mutect2, VarDict, VarScan2, GATK
Hallazgos extras	<ul style="list-style-type: none"> - El artículo menciona bcbio la cual es una herramienta para automatizar <i>pipelines</i> - RTG Tools, Hap.py: generan metricas para realizar <i>benchmarks</i> entre herramientas - vcflib: procesar archivos VCF

Tabla 76: Extracción de información de la Revisión Sistemática - Artículo 45

DOI y año	10.1038/s41598-018-30330-y, 2018
Nombre de la publicación	<i>Design and MinION testing of a nanopore targeted gene sequencing panel for chronic lymphocytic leukemia.</i>
Tipo de artículo	Comparación de herramientas
Objetivo del pipeline	Diagnóstico de enfermedades
Tipo de secuenciador que se usó	MinION
Herramienta y usos	<ul style="list-style-type: none"> - Control de calidad: NanoOK - Mapeo: BWA - Variant calling: Nanopolish, VarScan, FreeBayes - Clasificación de variantes: ANNOVAR
Hallazgos extras	<ul style="list-style-type: none"> - Nanopolish solo detectó 3 de las 8 variantes previamente llamadas por VarScan, esto dado que Nanopolish es más estricto para la detección de variantes y menos sensible que VarScan - Poretools toolkit: Convertir archivo FAST5 a FASTQ

Tabla 77: Extracción de información de la Revisión Sistemática - Artículo 46

DOI y año	10.1016/j.ajhg.2017.09.013, 2017
Nombre de la publicación	<i>Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human WholeGenomes.</i>
Tipo de artículo	Presentación de herramienta(s)
Objetivo del pipeline	Detección de variantes
Tipo de secuenciador que se usó	Illumina
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: BWA - Variant calling: Manta, Isaac, GATK, lobSTR, TREDPARSE
Hallazgos extras	POD5 package (Herramienta para la conversión de archivos FAST5 a POD5)

Tabla 78: Extracción de información de la Revisión Sistemática - Artículo 47

DOI y año	10.1186/s12859-017-2000-6, 2017
Nombre de la publicación	<i>A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy.</i>
Tipo de artículo	Presentación de pipeline
Objetivo del pipeline	Otro
Tipo de secuenciador que se usó	NGS
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: BWA - Trimming y filtrado: Cutadapt, Picard - Variant calling: Platypus, SAMtools, BCFtools - Clasificación de variantes: VCFtools
Hallazgos extras	Platypus está optimizado para trabajar con datos de secuenciadores NGS

Tabla 79: Extracción de información de la Revisión Sistemática - Artículo 48

DOI y año	10.1186/s13053-016-0058-1, 2016
Nombre de la publicación	<i>Pedigree based DNA sequencing pipeline for germline genomes of cancer families.</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del <i>pipeline</i>	Diagnóstico de enfermedades
Tipo de secuenciador que se usó	NGS
Herramienta y usos	<ul style="list-style-type: none"> - <i>Variant calling</i>: Platypus, SAMtools - Clasificación de variantes: CADD, ANNOVAR, HaploReg, Regulome - Base de datos: 1000 Genomes, dbSNP, Exome Aggregation Consortium (ExAC)

Tabla 80: Extracción de información de la Revisión Sistemática - Artículo 49

DOI y año	10.7717/peerj.2074, 2016
Nombre de la publicación	<i>DeepSNVMiner: a sequence analysis tool to detect emergent, rare mutations in subsets of cell populations.</i>
Tipo de artículo	Comparación de herramientas
Objetivo del <i>pipeline</i>	Detección de variantes
Tipo de secuenciador que se usó	NGS
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: BWA. - <i>Variant calling</i>: FreeBayes, GATK, BCFtools, LoFreq, DeepSNVMiner
Hallazgos extras	Se compararon los falsos positivos y falsos negativos de 5 herramientas de <i>variant calling</i> , donde DeepSNVMiner obtuvo los mejores resultados.

Tabla 81: Extracción de información de la Revisión Sistemática - Artículo 50

DOI y año	10.1186/1471-2164-15-516, 2014
Nombre de la publicación	<i>Evaluation and optimisation of indel detection workflows for ion torrent sequencing of the BRCA1 and BRCA2 genes</i>
Tipo de artículo	Optimización de parámetros
Objetivo del pipeline	Detección de variantes
Tipo de secuenciador que se usó	Ion Torrent
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: BWA - <i>Variant calling</i>: Samtools, GATK
Hallazgos extras	Tanto la herramienta Samtools como GATK obtuvieron resultados semejantes.

Tabla 82: Extracción de información de la Revisión Sistemática - Artículo 51

DOI y año	10.1155/2013/730210, 2013
Nombre de la publicación	<i>Computational and Bioinformatics Frameworks for Next-Generation Whole Exome and Genome Sequencing</i>
Tipo de artículo	Presentación de herramienta(s)
Objetivo del pipeline	Detección de variantes
Tipo de secuenciador que se usó	General
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: Bowtie, Bowtie2, SEAL, SOAP3, BWA, mrFAST, Novoalign, SHRIMP, MAQ, Stampy, ELAND, LAST Aligner, SARUMAN - <i>Variant calling</i>: GATK, SOAPSnp, VarScan/VarScan2, ATLAS 2, Pindel, Dindel - Clasificación de variantes: SIFT, PolyPhen, VariBench, snpEFF, SeattleSeq, ANNOVAR, VAAST, VAT, VARIANT, VAR-MD - Base de datos: SNPeffect Database

Tabla 83: Extracción de información de la Revisión Sistemática - Artículo 52

DOI y año	10.1159/000342770, 2012
Nombre de la publicación	<i>Fighting Outbreaks with Bacterial Genomics: Case Review and Workflow Proposal</i>
Tipo de artículo	Presentación de <i>pipeline</i>
Objetivo del pipeline	Detección de bacteria causante de enfermedades
Tipo de secuenciador que se usó	NGS
Herramienta y usos	<ul style="list-style-type: none"> - Mapeo: BWA, SOAP2, Bowtie - Variant calling: Samtools, GATK, SOAPSnp
Hallazgos extras	SOAP2 y SOAPSnp son herramientas optimizadas para el análisis de secuencias, generadas por secuenciadores NGS

Anexo B

A continuación se presenta desde la Tabla 84 a la Tabla 90, la descripción del *input*, *output* y estadísticos generados de las herramientas, para el control de calidad de secuencias, previo al mapeo. De igual manera, desde la Tabla 91 a la Tabla 98, se presentan las características y el enlace a su repositorio o sitio oficial, de las herramientas para el control de calidad del mapeo.

Tabla 84: Características de FastQC para control de calidad.

Herramienta	FastQC		
URL	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/		
Input	FASTQ	Output	HTML
Estadísticas generadas	<ul style="list-style-type: none"> - Lecturas totales generadas. - Bases totales generadas. - Largo minimo y maximo de las lecturas. - Porcentaje de GC promedio. - Box-Plot de calidad por base. - Gráfico de la distribución de la calidad de los reads. - Gráfico de la composición de bases por posición de los reads - Gráfico de la distribución del porcentaje de GC de los reads - Gráfico de la proporción de bases 'N' en cada posición a lo largo de los reads - Gráfico de la distribución del largo de las lecturas. - Gráfico con el nivel de duplicación de las secuencias - Secuencias sobrerepresentadas - Gráfico del porcentaje del contenido de adaptadores a lo largo de los reads 		

Tabla 85: Características de Fastp para control de calidad.

Herramienta	Fastp		
URL	https://github.com/OpenGene/fastp		
Input	FASTQ	Output	HTML + JSON
Estadísticas generadas	<ul style="list-style-type: none"> - Largo promedio de las lecturas. - Tasa de duplicación de las lecturas. - Número de bases con calidad Q20/Q30. - Contenido de GC promedio.. - Lecturas totales generadas. - Bases totales generadas. - Adaptadores encontrados. - Gráfico de la distribución del tamaño de los insertos. - Gráfico de la calidad media a lo largo de las lecturas por cada base. - Gráfico del contenido de bases a lo largo de las lecturas. - KMERS - Secuencias sobrerrepresentadas 		

Tabla 86: Características de MinIONQC para control de calidad.

Herramienta	MinIONQC		
URL	https://github.com/roblanf/minion_qc		
Input	sequencing_summar y.txt	Output	PNG + YML
Estadísticas generadas	<ul style="list-style-type: none"> - Gráfico de la distribución del largo promedio de las lecturas por canal. - Gráfico de la distribución de la mediana del largo de las lecturas por canal. - Gráfico con el números de bases leídas por canal. - Gráfico con el número de lecturas por canal. - Gráfico con la distribución del largo de las lecturas por hora por canal. - Gráfico con la distribución de gigabases secuenciadas por cada canal de la <i>flowcell</i>. - Gráfico con el largo promedio de las lecturas por hora. - Gráfico distribución del largo de las lecturas. - Gráfico de la relación entre calidad y largo de los reads - Gráfico de la calidad promedio de las lecturas por hora - Gráfico del número de reads generados por hora - Gráfico de gigabases totales generadas, según longitud mínima - Gráfico de gibabases totales acumuladas por hora 		

Tabla 87: Características de NanoPlot para control de calidad.

Herramienta	NanoPlot		
URL	https://github.com/wdecoster/NanoPlot		
Input	FASTQ / FASTA / FASTQ (con información adicional sobre el canal y la hora)	Output	HTML + PNG
Estadísticas generadas	<ul style="list-style-type: none"> - Largo promedio - Calidad promedio - Largo medio - Calidad media - Número de lecturas generadas. - Número y porcentaje de lecturas con calidad mayor a Q5 / Q7 / Q10 / Q12 / Q15. - Histograma del largo de los reads - Gráfico del rendimiento acumulado a lo largo de las bases generadas. - Gráfico del largo de los reads versus la calidad de estos 		

Tabla 88: Características de NanoQC para control de calidad.

Herramienta	NanoQC		
URL	https://github.com/wdecoster/nanoQC		
Input	FASTQ	Output	HTML
Estadísticas generadas	<ul style="list-style-type: none"> - Gráfico de distribución del largo de los reads - Gráfico de frecuencia de cada nucleótido en las 100 primeras bases de las lecturas - Frecuencia de cada nucleótido en las 100 últimas bases de las lecturas - Gráfico de calidad media en las 100 primeras bases de las lecturas. - Gráfico de calidad media en las 100 últimas bases de las lecturas. 		

Tabla 89: Características de PycoQC para control de calidad.

Herramienta	PycoQC		
URL	https://github.com/a-slide/pycoQC		
Input	sequencing_summar y.txt	Output	HTML
Estadísticas generadas	<ul style="list-style-type: none"> - Duración de la secuenciación - Canales activos - Número de lecturas generadas. - Número de bases - N50 - Largo promedio de las lecturas - Calidad promedio promedio de las lecturas. - Gráfico de la distribución del largo de los reads - Gráfico de la distribución de la calidad de los reads - Gráfico de la relación entre calidad y largo de los reads - Gráfico de las lecturas generadas por hora. - Gráfico de las bases generadas por hora - Gráfico con el largo promedio, mínimo y máximo de las lecturas en cada momento - Gráfico con la calidad promedio, mínima y máxima de las lecturas en cada momento 		

Tabla 90: Características de Seqkit para control de calidad.

Herramienta	Seqkit		
URL	https://github.com/shenwei356/seqkit		
Input	FASTA / FASTQ	Output	TSV
Estadísticas generadas	<ul style="list-style-type: none"> - Número de secuencias - Suma de los largos de las secuencias - Lectura más corta - Lectura más larga - Largo promedio de los reads - Cuartiles del largo de los reads - Número de gaps - N50 - N50_num (cuántas secuencias suman para llegar al valor N50) - Porcentaje de reads con calidad igual o sobre Q20 - Porcentaje de reads con calidad igual o sobre Q30 - Calidad promedio de las lecturas - Porcentaje de GC de las lecturas 		

Tabla 91: Características de Bedtools para analizar la calidad del mapeo.

Herramienta	Bedtools		
URL	https://bedtools.readthedocs.io/en/latest		
Input	SAM / BAM	Output	BED
Características	<ul style="list-style-type: none"> - Análisis de cobertura con `coverage` - Número de veces se ha secuenciado cada posición en el genoma: `genomcov` y que porcentaje representa. 		

Tabla 92: Características de DeepTools para analizar la calidad del mapeo.

Herramienta	DeepTools		
URL	https://deeptools.readthedocs.io/en/latest		
Input	BAM (indexado)	Output	PNG + TXT
Características	<ul style="list-style-type: none"> - Cuartiles de la profundidad de lectura alcanzada. - Gráficos del porcentaje de cobertura en cada base - Gráfico del porcentaje de bases con X cobertura - Gráfico de distribución promedio de la cobertura entre las regiones objetivo. 		

Tabla 93: Características de Qualimap para analizar la calidad del mapeo.

Herramienta	Qualimap		
URL	http://qualimap.conesalab.org		
Input	SAM / BAM	Output	HTML + TXT
Características	<ul style="list-style-type: none"> - Lecturas mapeadas y no mapeadas - Largo máximo, mínimo y promedio de las lecturas. - Porcentaje de error (discrepancias entre la secuencia de referencia y las lecturas alineadas). - Inserciones, delecciones y <i>missmatches</i>. - Número de lecturas recortadas. - MAPQ promedio. - Número de lecturas mapeadas en una región objetivo. 		

Tabla 94: Características de Mosdepth para analizar la calidad del mapeo.

Herramienta	Mosdepth		
URL	https://github.com/brentp/mosdepth		
Input	BAM (indexado) / CRAM	Output	TXT / BED
Características	<ul style="list-style-type: none"> - Profundidad de lectura de cada base. - Profundidad promedio en la región objetivo. - Porcentaje de bases obtuvieron X profundidad en una región objetivo. 		

Tabla 95: Características de NanoPlot para analizar la calidad del mapeo.

Herramienta	NanoPlot		
URL	https://github.com/wdecoster/NanoPlot		
Input	SAM (ordenado) / CRAM	Output	HTML
Características	<ul style="list-style-type: none"> - Porcentaje de identidad promedio. - Porcentaje de lecturas mapeadas. - Gráfico del largo de las lecturas mapeadas versus el largo de las lecturas secuenciadas. - Gráfico de la calidad de las lecturas versus la calidad del mapeo de las lecturas. - Gráfico del largo de las lecturas versus la calidad del mapeo. - Gráfico del porcentaje de identidad de las lecturas versus la calidad del mapeo. - Gráfico del porcentaje de identidad versus el largo de las lecturas mapeadas. - Histograma de la distribución del porcentaje de identidad de las lecturas. 		

Tabla 96: Características de Picard para analizar la calidad del mapeo.

Herramienta	Picard		
URL	https://broadinstitute.github.io/picard/command-line-overview.html		
Input	BAM	Output	TXT
Características	<ul style="list-style-type: none"> - Lecturas mapeadas por cromosomas. - Número total de lecturas mapeadas. - Número de lecturas con calidad de mapeo sobre Q20. - Tasa de bases que no coinciden con la referencia. - Número de inserciones y delecciones cada 100 bases. 		

Tabla 97: Características de Samtools para analizar la calidad del mapeo.

Herramienta	Samtools		
URL	https://github.com/samtools/samtools		
Input	SAM / BAM	Output	SAM / BAM
Características	<ul style="list-style-type: none"> - Número de bases y lecturas mapeadas por región. - Número de veces que se mapea cada base. - Resumen de los Samflags. - Número de mismatches - Distribución de inserciones y delecciones. 		

Tabla 98: Características de SeqKit para analizar la calidad del mapeo.

Herramienta	SeqKit		
URL	https://github.com/shenwei356/seqkit		
Input	BAM	Output	TXT
Características	<ul style="list-style-type: none"> - Número y porcentaje de lecturas mapeadas. 		

Anexo C

Descripción del *input*, *output* y características de las herramientas, para la etapa de *trimming* y filtrado (Tabla 99 a Tabla 109).

Tabla 99: Características de BBduk para *trimming* y filtrado.

Herramienta	BBduk		
URL	https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbduk-guide/		
Input	FASTQ	Output	FASTQ
Características	<ul style="list-style-type: none"> - Podar los N primeros y/o últimas bases - Podar las bases más a la izquierda que no cumplan con cierto valor de calidad - Podar las bases más a la derecha que no cumplan con cierto valor de calidad - Filtrar lecturas por calidad promedio - Filtrar lecturas que posean adaptadores - Filtrar lecturas por largo mínimo 		

Tabla 100: Características de Chopper para *trimming* y filtrado.

Herramienta	Chopper (sustituto de Nanofilt)		
URL	https://github.com/wdecoster/chopper		
Input	FASTQ	Output	FASTQ
Características	<ul style="list-style-type: none"> - Podar N nucleótidos del comienzo de los lecturas - Podar N nucleótidos del final de las lecturas - Filtrar lecturas que no cumplen con un largo máximo - Filtrar lecturas que no cumplen con un largo mínimo - Filtrar lecturas que no cumplen con una calidad promedio - Filtrar lecturas contaminadas, de acuerdo a secuencias de referencia 		

Tabla 101: Características de Cutadapt para *trimming* y filtrado.

Herramienta	Cutadapt		
URL	https://github.com/marcelm/cutadapt		
Input	FASTA / FASTQ / uBAM ⁴⁰	Output	FASTA / FASTQ / uBAM
Características	<ul style="list-style-type: none"> - Podar adaptadores - Podar las N primeras bases de las lecturas - Podar las N últimas bases de las lecturas - Podar los extremos de las lecturas que no cumplen con una calidad mínima - Establecer un tamaño máximo a las lecturas - Filtrar lecturas por largo mínimo - Filtrar lecturas por largo máximo - Filtrar las lecturas en las que se encontraron adaptadores - Filtrar lecturas con X bases “N” o porcentaje de bases “N” - Demultiplexar lecturas - Generar reporte 		

Tabla 102: Características de DeepTools para *trimming* y filtrado.

Herramienta	DeepTools		
URL	https://deeptools.readthedocs.io/en/latest/		
Input	BAM	Output	BAM
Características	<ul style="list-style-type: none"> - Filtrar lecturas por mapping quality - Filtrar lecturas por samflags - Filtrar duplicados 		

⁴⁰ **uBAM:** unaligned BAM

Tabla 103: Características de Fastp para *trimming* y filtrado.

Herramienta	Fastp		
URL	https://github.com/OpenGene/fastp		
Input	FASTQ	Output	FASTQ
Características	<ul style="list-style-type: none"> - Filtrar lecturas con X número de bases N - Filtrar lecturas con más del X% de bases no cualificadas - Filtrar lecturas por calidad promedio mínima - Filtrar lecturas por largo mínimo - Filtrar lecturas por largo máximo - Filtrar lecturas por porcentaje de complejidad mínima - Podar adaptadores - Podar lecturas mediante una ventana deslizante que evalúa la calidad de las lecturas y las poda si no se cumple con el mínimo, mientras que si se cumple, se detiene (válido en ambas direcciones de la secuencia) - Podar lectura mediante ventana deslizante que recorre toda la secuencia y cuando no se cumple con la calidad mínima, poda la lectura desde la posición de la ventana, hacia la derecha. - Poda X bases del principio de la lectura - Poda X bases del final de la lectura - Fija las lecturas a un largo máximo, podando desde el final de esta. 		

Tabla 104: Características de Filtlong para *trimming* y filtrado.

Herramienta	Filtlong		
URL	https://github.com/rrwick/Filtlong		
Input	FASTQ	Output	FASTQ
Características	<ul style="list-style-type: none"> - Conservar sólo las N mejores bases - Conservar sólo el X% de las mejores lecturas - Filtrar las lecturas por largo máximo - Filtrar las lecturas por largo mínimo - Filtrar las lecturas por calidad mínima - Realiza una evaluación de calidad mediante ventana deslizante, filtrando la lectura, si la calidad promedio dentro de la ventana no cumple con el mínimo establecido - Filtrar las lecturas, de acuerdo a lecturas de referencias de secuenciadores de Illumina o ensambles 		

Tabla 105: Características de Picard para *trimming* y filtrado.

Herramienta	Picard		
URL	https://broadinstitute.github.io/picard/		
Input	SAM / BAM	Output	SAM / BAM
Características	<ul style="list-style-type: none"> - Marca y remueve lecturas duplicadas 		

Tabla 106: Características de Porechop para *trimming* y filtrado.

Herramienta	Porechop		
URL	https://github.com/rrwick/Porechop		
Input	FASTA / FASTQ	Output	FASTA / FASTQ
Características	<ul style="list-style-type: none"> - Demultiplexar las lecturas - Podar adaptadores - Dividir lecturas con adaptadores en el medio - Filtrar lecturas con adaptadores en el medio 		

Tabla 107: Características de Samtools para *trimming* y filtrado.

Herramienta	Samtools		
URL	https://github.com/samtools/samtools		
Input	SAM	Output	SAM
Características	<ul style="list-style-type: none"> - Marcar y filtrar duplicados - Filtrar secuencias de acuerdo a samflags - Filtrar secuencias respecto a la calidad del mapeo 		

Tabla 108: Características de Seqkit para *trimming* y filtrado.

Herramienta	Seqkit		
URL	https://github.com/shenwei356/seqkit		
Input	FASTQ	Output	FASTQ
Características	<ul style="list-style-type: none"> - Filtrar lecturas por largo máximo - Filtrar lecturas por largo minima - Filtrar lecturas por calidad máxima - Filtrar lecturas por calidad mínima - Filtrar lecturas duplicadas 		

Tabla 109: Características de Trimmomatic para *trimming* y filtrado.

Herramienta	Trimmomatic		
URL	https://github.com/usadellab/Trimmomatic		
Input	FASTQ	Output	FASTQ
Características	<ul style="list-style-type: none"> - Realiza un recorte mediante ventana deslizante, cortando una vez que la calidad promedio dentro de la ventana caiga por debajo de un umbral - Poda el inicio de las lecturas que no cumplen con una calidad mínima - Poda al final de las lecturas que no cumplen con una calidad mínima - Poda las lecturas de un largo específico - Poda las n primeras bases al comienzo de las lecturas - Filtrar las lecturas que no cumplen con un largo mínimo 		

Anexo D

La Tabla 110 presenta los motivos de exclusión de las herramientas de *variant calling*.

Tabla 110: Motivo de exclusión de herramientas de *variant calling*.

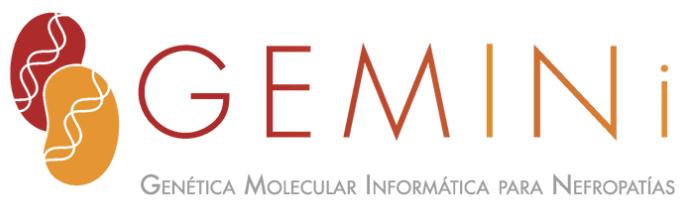
Herramienta	Motivo
ATLAS 2	No ha recibido actualizaciones desde 2013.
Clair	Versión anterior de Clair3 .
Clair-ensemble	Versión optimizada de Clair , que no ha tenido nuevas actualizaciones desde su salida.
DeepSNV	No se garantiza su funcionamiento en un futuro.
DeepSNVMiner	No devuelve resultados en formato VCF .
Dindel	Ya no está disponible.
EBCall	Herramienta para el llamado de variantes somáticas.
GLIMPSE	Herramienta para la imputación ⁴¹ de variantes genéticas a partir de datos con baja cobertura.
Hello	Herramienta de <i>variant calling</i> entrenada específicamente para muestras provenientes de Illumina y PacBio .

⁴¹ **Imputación de variantes genéticas:** Proceso mediante el cual se predicen variantes genéticas no observadas directamente en los datos de secuenciación.

Herramienta	Motivo
Htslib	Librería de BcfTools.
Human-SNP-wf	Forma parte del pipeline wf-human-variation . Esta herramienta utiliza Clair3 , y como se menciona en el artículo de Helal et al. (2022), los resultados obtenidos son bastante similares.
Isaac	Ya no recibe soporte.
LobSTR	Ya no recibe soporte.
Manta	Herramienta para el llamado de variantes estructurales.
Metaka	Ya no recibe soporte para SNPs ni INDELS .
MuTect2	Herramienta para el llamado de variantes somáticas.
Nanopolish	No soporta las nuevas <i>flowcells</i> R10.4, las cuales fueron utilizadas
Pindel	Ya no recibe soporte.
Platypus	No ha recibido actualizaciones en los últimos seis años y presenta problemas de compatibilidad con las herramientas necesarias.
Samtools	Herramienta para manipular archivos SAM y BAM .
Shimmer	Herramienta para el llamado de variantes somáticas.
Strelka	La herramienta presenta tiempos de ejecución elevados que no generaron resultados, además de ser una herramienta desarrollada por Illumina para datos provenientes de secuenciadores de nueva generación (NGS).
SOAPsnp	No recibe soporte desde 2015.
TREDPARSE	<i>Short Tandem Repeat caller</i> .
VarDict	Ya no recibe soporte.
VarScan	Versión anterior de VarScan2 .
Virimd	Herramienta para el llamado de variantes somáticas.
WhatsHap	Herramienta para <i>phasing</i> de haplotipos.

Anexo E

El siguiente documento es el manual de usuario generado como resultado de este trabajo.



Manual de Usuario Pipeline análisis de variantes genéticas

Autor: Diego Millar Coronado

v1.0, 10 de Noviembre de 2024

ÍNDICE

ÍNDICE.....	2
1. PRESENTACIÓN.....	3
2. CONSIDERACIONES DE USO.....	4
3. REQUERIMIENTOS DE HARDWARE.....	5
4. REQUERIMIENTOS DE SOFTWARE.....	6
4.1 Preparación de ambiente de Conda (Miniconda).....	6
4.2 Configuración de ANNOVAR: Herramientas y Bases de Datos.....	7
5. GUIA DE USO.....	8
5.1 Preparación del entorno de trabajo.....	8
5.1.1 Creación de los directorios.....	8
5.1.2 Variables utilizadas.....	9
5.2 Ejecución del pipeline mediante comandos.....	10
5.2.1 Basecalling y demultiplexación.....	10
5.2.2 Control de calidad.....	11
5.2.3 Trimming y filtrado.....	13
5.2.4 Mapeo.....	14
5.2.5 Control de calidad del mapeo, filtrado y preprocesamiento de la salida.....	14
5.2.6 Variant calling.....	16
5.2.7 Anotación de variantes.....	17
5.2.8 Sistema de ranking de variantes.....	18
5.3 Ejecución automatizada mediante script.....	19
ANEXO.....	22
Anexo A.....	22

1. PRESENTACIÓN

El presente *pipeline* de análisis de variantes genéticas es un flujo de trabajo bioinformático diseñado para analizar variantes genéticas a partir de secuencias de ADN generadas con el secuenciador **Oxford Nanopore MinION**. Este *pipeline* ha sido desarrollado y validado específicamente para el diagnóstico de la enfermedad hereditaria **ADPKD**, utilizando muestras de ADN provenientes de pacientes en quienes se ha confirmado previamente la presencia de una variante en el gen **PKD1**.

La solución propuesta fue desarrollada como parte del trabajo de titulación de Diego Millar, en su camino para obtener el título de Ingeniero Civil en Informática. Este trabajo, llevado a cabo en el marco del proyecto **GEMINI-2**, es el resultado de nueve meses de trabajo.

Este *pipeline* puede ser implementado siguiendo una guía estructurada, que detalla paso a paso cada etapa del análisis, o bien ejecutado de forma automatizada mediante un *script* que integra y automatiza todas las fases del proceso.

2. CONSIDERACIONES DE USO

El *pipeline* presentado fue desarrollado y validado en un entorno **Linux**, por lo que podrían surgir diferencias en los comandos si se ejecutan en otros sistemas operativos. Además, es posible que se requieran mayores recursos de *hardware* si se analizan muestras de tamaño significativamente superior al conjunto de datos utilizado para la validación. Como referencia, la suma de los archivos **POD5** empleados tenía un tamaño aproximado de **77 GB**, mientras que el archivo **FASTQ** de mayor peso mapeado alcanzaba alrededor de **1,42 GB**.

Por otro lado, aunque este *pipeline* fue desarrollado y validado con muestras del gen **PKD1**, los comandos y herramientas presentados son aplicables a cualquier otro gen o región genómica. Sin embargo, es importante tener en cuenta que el *script ranking_system.py* incorpora información de variantes del gen **PKD1** almacenadas en la **Clinica Mayo**, lo cual puede resultar no aplicable para el análisis de otros genes.

Para ejecutar el *pipeline*, es necesario contar con los programas requeridos instalados (los cuales se detallarán más adelante) y tener conocimientos básicos para ejecutar comandos en la terminal. Solo en caso de querer actualizar los *scripts* diseñados para este *pipeline*, sería necesario disponer de conocimientos adicionales en **Python** y **Bash**.

3. REQUERIMIENTOS DE HARDWARE

Las diferentes herramientas utilizadas en este *pipeline* tienen distintos requisitos de *hardware*. Aunque existen ciertos parámetros básicos que deben cumplirse en la mayoría de los casos, los cuales se describen en la Tabla 1, algunas herramientas requieren configuraciones más avanzadas.

Tabla 1: Requerimientos básicos de hardware.

Componente	Especificación mínima recomendada
CPU	AMD Ryzen 5 o Intel Core i5 (o equivalente)
RAM	8 GB
Almacenamiento	512 GB

En particular, la herramienta **Dorado**, utilizada en la etapa de *basecalling* y demultiplexión, que necesita una **tarjeta gráfica NVIDIA con un mínimo de 8 GB de VRAM** o bien un computador **Mac con procesador M1 o M2**.

Es importante tener en cuenta que, durante la etapa de mapeo, los requisitos de memoria **RAM** pueden incrementarse significativamente si se utiliza el genoma humano completo como referencia.

4. REQUERIMIENTOS DE SOFTWARE

A continuación se presentan el *software* y librerías de **Python** requeridas para ejecutar cada etapa del *pipeline* (Tabla 2).

Tabla 2: *Software* requerido por el *pipeline*.

Software	Etapa	URL
Dorado	<i>Basecalling</i>	https://github.com/nanoporetech/dorado/releases/
Miniconda	Control de calidad - Sistema de <i>ranking</i> de variantes	https://docs.anaconda.com/miniconda/miniconda-install/
FastQC	Control de calidad	https://anaconda.org/bioconda/fastqc
FastP	<i>Trimming</i>	https://anaconda.org/bioconda/fastp
Minimap2	Mapeo	https://anaconda.org/bioconda/minimap2
Samtools	Control de calidad del mapeo	https://anaconda.org/bioconda/samtools
Bedtools		https://anaconda.org/bioconda/bedtools
Bcftools	<i>Variant calling</i>	https://anaconda.org/bioconda/bcftools
Docker		https://docs.docker.com/engine/install/
ANNOVAR	Anotación y clasificación de variantes	https://annovar.openbioinformatics.org/en/latest/user-guide/download/
<i>variant_analysis_pipeline</i>	Sistema de <i>ranking</i> de variantes	https://github.com/GEMINI-2/variant_analysis_pipeline (Se requiere solicitar acceso)
Numpy		https://anaconda.org/anaconda(numpy
Pandas		https://anaconda.org/anaconda/pandas
Requests		https://anaconda.org/anaconda/requests
BioPython		https://anaconda.org/anaconda/biopython

4.1 Preparación de ambiente de Conda (Miniconda)

Conda es un gestor de entornos y paquetes que facilita la instalación y administración de *software* en diferentes ambientes. Varios de los programas y bibliotecas necesarios para ejecutar el *pipeline* y el sistema de *ranking* se pueden instalar utilizando **Conda**, excepto **Dorado**, **ANNOVAR** y **Docker**. Para ello, se empleará el archivo “**environment.yml**”, ubicado en la carpeta “**variant_analysis_pipeline**”. Este archivo contiene la información

sobre los programas y bibliotecas requeridos, junto con sus versiones específicas. Para crear el entorno de **Conda**, se debe ejecutar el siguiente comando:

```
$ conda env create -f <PIP_PATH>/environment.yml
```

Donde la variable **<PIP_PATH>** representa la dirección a la carpeta **“variant_analysis_pipeline”**.

4.2 Configuración de ANNOVAR: Herramientas y Bases de Datos

ANNOVAR requiere una serie de bases de datos y herramientas para su funcionamiento depende de una serie de bases de datos y herramientas que, en conjunto, **utilizan aproximadamente 200 GB de espacio de almacenamiento**. Entre estas se incluyen bases de datos de clasificación de variantes genéticas, como **ClinVar**; bases de datos de frecuencia alélica, como **GnomAD**; y bases de datos que integran predicciones de diversas herramientas de patogenicidad, como **DBNSFP**. A continuación, se presenta la lista completa de bases de datos necesarias:

- avsnp150
- dbnsfp42c
- clinvar_20240611
- intervar_20180118
- abraom
- gnomad41_genome
- gnomad41_exome
- esp6500siv2_all
- 1000g2015aug

Para descargar las bases de datos proporcionadas por **ANNOVAR**, es necesario especificar la versión del genoma que se utilizará. En este caso, se empleará la versión **hg38 (GRCh38)**. Esta versión del genoma humano, lanzada en 2013, es la más reciente y reemplaza a la versión anterior, **hg19 (GRCh37)**. La versión **hg38** se distingue por mejorar la calidad de las secuencias mediante la corrección de errores y la incorporación de nuevas secuencias, incluyendo regiones complejas que no estaban presentes en versiones anteriores. Para descargar este genoma de referencia, se debe ejecutar el siguiente comando:

Una de las bases de datos clave que proporciona **ANNOVAR** es **refGene**, la cual contiene anotaciones genéticas detalladas. Estas incluyen información como el gen asociado a una variante, la ubicación precisa de la variante (por ejemplo, intrón, exón, upstream, etc.), y su efecto funcional (como *synonymous*, *nonsynonymous*, *stopgain*, entre otros). Para descargar la base de datos **refGene** correspondiente a la versión del genoma **hg38**, se debe ejecutar el siguiente comando:

```
$ annotate_variation.pl -buildver hg38 -downdb -webfrom annovar refGene <DB>
```

Donde **<DB>** representa la ruta de la carpeta donde se almacenará la base de datos. Posteriormente, para descargar las herramientas y bases de datos requeridas, se debe ejecutar el mismo comando, reemplazando el parámetro **refGene** por el nombre de cada herramienta o base de datos que se desee obtener.

5. GUIA DE USO

A continuación se detalla el paso a paso del *pipeline* para el análisis bioinformático de las muestras provenientes del secuenciador *Oxford Nanopore MinION* para la detección y clasificación de variantes.

5.1 Preparación del entorno de trabajo

5.1.1 Creación de los directorios

Para desarrollar el *pipeline* de manera más eficiente, es recomendable mantener un entorno de trabajo organizado, con directorios estructurados para almacenar las salidas generadas en cada etapa del análisis. Esto facilita el acceso y manejo de los resultados, mejorando la fluidez y la claridad durante el desarrollo y ejecución del *pipeline*.

Al iniciar el análisis, es fundamental contar con las muestras en formato **POD5**, las cuales deben estar ubicadas en una carpeta exclusiva para estos archivos. Además, es aconsejable que dicha carpeta se aloje dentro de un directorio de trabajo donde se ejecutará todo el *pipeline*. La estructura sugerida de los directorios se muestra en la Figura 1.

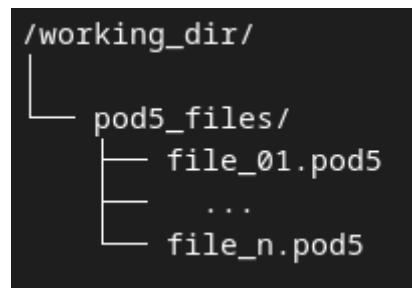


Figura 1: Estructura del directorio de trabajo inicial.

Una vez creado el directorio de trabajo y ubicado el directorio con las muestras correspondientes, se recomienda crear un nuevo directorio llamado “**output**”. Dentro de este, se deberán crear subdirectorios específicos para cada etapa del *pipeline*, donde se almacenarán los resultados generados en cada fase (ver Figura 2).

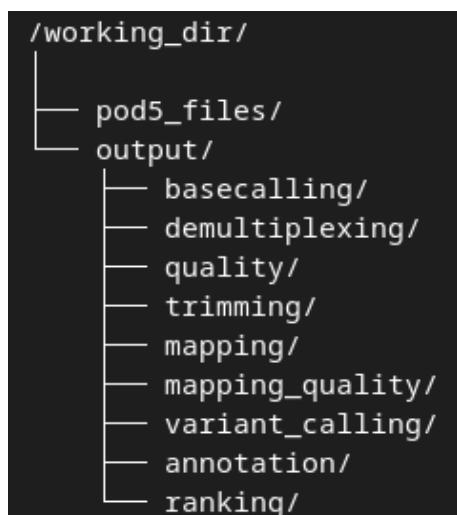


Figura 2: Estructura del directorio de resultados.

Finalmente, se debe crear un directorio llamado “***data***”, donde se almacenará un archivo **BED** denominado **regions.bed**, que contiene las regiones cubiertas por los amplicones. Este archivo se encuentra disponible en la carpeta “***data***” del directorio **variant_analysis_pipeline** y define la ubicación de las regiones genómicas representadas en la Figura 3 con respecto al cromosoma 16. De esta manera, la estructura final de las carpetas debería verse como se muestra en la Figura 4.

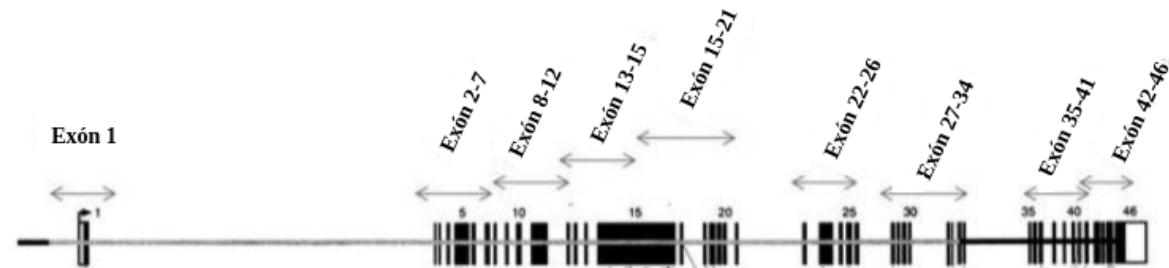


Figura 3: Regiones del gen *PKD1* cubiertas con amplicones⁴².

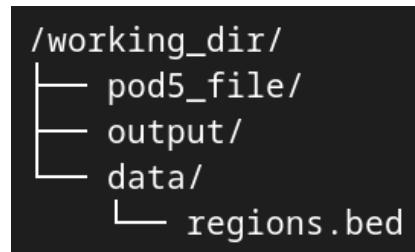


Figura 4: Estructura del directorio “*data*”.

En las siguientes etapas, se considerará la configuración de directorios previamente mencionada para la ejecución de cada paso. No obstante, es importante señalar que el genoma de referencia no se incluyó dentro del directorio de trabajo. Esto se debe a que dicho archivo puede ser utilizado en múltiples análisis, por lo que se recomienda que el usuario lo almacene en una ubicación de su preferencia para facilitar su acceso y reutilización.

5.1.2 Variables utilizadas

A lo largo del desarrollo del *pipeline*, se presentarán comandos que utilizan ciertos parámetros cuyos valores pueden repetirse o ser variables que necesitan ser definidas por el usuario que está ejecutando el *pipeline*. Por lo tanto, en la Tabla 3, se proporciona una lista de variables importantes que serán utilizadas en una o más etapas del *pipeline*.

⁴² <https://pmc.ncbi.nlm.nih.gov/articles/PMC3391417/>

Tabla 3: Variables generales.

Variable	Descripción
DB	Ubicación del directorio donde se almacenan las base de datos descargadas de ANNOVAR .
LIBRARY_KIT	Nombre del kit de preparación de librerías utilizado (Ejemplo: SQK-NBD114-24).
PIP_PATH	Ruta al directorio donde están ubicados los <i>scripts</i> de variant_analysis_pipeline
REF	Nombre del archivo FASTA , con el genoma de referencia (Ejemplo: GRCh38.fasta)
REF_DIR	Ruta absoluta ⁴³ al directorio donde se almacena el genoma de referencia.
THREADS	Número de núcleos disponibles para la ejecución en paralelo de las tareas.

5.2 Ejecución del pipeline mediante comandos

A continuación, se explica paso a paso la ejecución del *pipeline*. Para comenzar, es importante activar el entorno de **Conda** creado previamente, ya que este se utilizará en la mayoría de las etapas, excepto en las fases de *basecalling*, *variant calling* y anotación de variantes. Sin embargo, se recomienda mantenerlo activado en todo momento para asegurar continuidad en el proceso. Para activar el ambiente de **Conda** se debe ejecutar el siguiente comando:

```
$ conda activate pipeline_PKD1
```

Para las próximas etapas, todos los comandos deben ejecutarse desde el directorio de trabajo.

5.2.1 Basecalling y demultiplexación

En la etapa de *basecalling* se utilizará la herramienta **Dorado**. Esta herramienta ofrece diversos modelos que varían en precisión, lo cual impacta directamente en el tiempo de ejecución. Para este *pipeline* se empleará el modelo más preciso disponible hasta la fecha (**SUP**). La ejecución se realiza con el siguiente comando::

```
$ dorado duplex sup pod5_files > output/basecalling/samples.bam
```

Una vez completada la etapa de *basecalling*, si se han secuenciado muestras de más de un paciente simultáneamente, es necesario llevar a cabo el proceso de demultiplexación. Para realizar la demultiplexación, se debe ejecutar el siguiente comando, el cual generará un archivo **FASTQ** para cada paciente, identificado por un *barcode*, así como un archivo

⁴³ **Ruta absoluta:** La ruta completa desde el directorio raíz hasta el archivo o directorio. Ejemplo en **Linux**: /home/<USER>/.../<TARGET_DIRECTORY>; en **Mac**: /Users/<USER>/.../<TARGET_DIRECTORY>.

adicional que contendrá las muestras en las que no se pudo distinguir los *barcodes*. Es importante especificar en el comando el código del kit de preparación de librerías.

```
$ dorado demux -o ./output/demultiplexing \
    -kit-name <LIBRARY_KIT> \
    --emit-summary --emit-fastq \
    output/basecalling/samples.bam
```

A partir de las siguientes etapas, es fundamental tener en cuenta que cada comando debe ser ejecutado y adaptado para cada una de las muestras disponibles. Por lo tanto, para referirse a una muestra específica, se utilizará el término <**SAMPLE**>, seguido de un sufijo que indique el caso correspondiente y, finalmente, el formato de la muestra.

Si no se han secuenciado muestras de más de un paciente, se puede continuar con la siguiente etapa. Sin embargo, es necesario convertir el archivo **BAM** obtenido en el paso anterior a un formato **FASTQ**, lo cual se puede lograr ejecutando el siguiente comando de la herramienta **Samtools**:

```
$ samtools fastq output/basecalling/samples.bam \
    -o output/demultiplexing/<SAMPLE>.fastq
```

5.2.2 Control de calidad

Para analizar la calidad de los datos, se utilizará la herramienta **FastQC**. Esta genera un informe en formato **HTML** que se puede visualizar directamente en el ordenador. El comando para ejecutar **FastQC** es el siguiente:

```
$ fastqc --memory 4096 output/demultiplexing/<SAMPLE>.fastq \
    -o ./output/quality
```

Notar que este comando debe ser ejecutado para cada muestra obtenida en la demultiplexación (<**SAMPLE**>), modificando el nombre del archivo de entrada.

El comando anterior genera dos archivos: un archivo **HTML** que contiene el reporte de calidad de las lecturas, y un directorio comprimido en formato **ZIP**, que incluye toda la información presentada en el primer archivo, como imágenes, íconos e información de los reportes.

El archivo **HTML** se divide en diez secciones, cada una de las cuales analiza la calidad de diferentes aspectos de los datos (ver Tabla 4). Cada sección tiene un ícono que indica el estado de calidad (Figura 5): un ícono verde señala que la calidad es adecuada, un ícono amarillo con un signo de interrogación indica posibles problemas en la calidad, y un ícono rojo con una cruz sugiere que la calidad es deficiente para esa sección.

-  Calidad Óptima
-  Calidad Subóptima
-  Calidad Insuficiente

Figura 5: Indicadores de calidad de FastQC

Tabla 4: Descripción de las secciones del reporte de calidad.

Sección	Descripción
Estadísticas básicas	Contiene información básica del archivo, así como también estadísticas básicas como el número de lecturas, el número de bases y el porcentaje de GC .
Calidad de las secuencias por base	Muestra gráficos de caja que representan la distribución de la calidad de las bases a lo largo de cada posición en la lectura.
Puntuaciones de calidad por secuencia	Gráfico, que representa la distribución de calidad promedio de las lecturas.
Contenido de secuencia de base por base	Gráfico que muestra el porcentaje de cada base en cada posición de las lecturas.
Contenido de GC por secuencia	Gráfico que muestra la distribución de GC .
Contenido de bases N por base	Gráfico que muestra el contenido de bases N a lo largo de la lectura.
Distribución de la longitud de la secuencia	Gráfico que muestra la distribución del largo de las lecturas.
Niveles de duplicación de secuencias	Gráfico que muestra el porcentaje de lecturas duplicadas.
Secuencias sobrerepresentadas	Tabla que muestra secuencias sobrerepresentadas.
Contenido de adaptadores	Gráfico que muestra el contenido de adaptadores en cada posición a lo largo de las lecturas.

Al interpretar el reporte⁴⁴, es importante considerar el origen de los datos. Por ejemplo, al evaluar el contenido de **GC** (guanina y citosina), **FastQC** asume que la distribución ideal es normal, con un promedio de **GC** del 50%. Sin embargo, el gen **PKD1** tiene un alto porcentaje de **GC**, lo cual debe tenerse en cuenta al analizar los resultados.

⁴⁴ <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/>

5.2.3 Trimming y filtrado

Una vez analizada la calidad de las muestras, se recomienda aplicar los siguientes filtros:

- Filtrar las lecturas con una longitud menor a 30 bases.
- Filtrar las lecturas con una longitud mayor a la del amplicón más largo (**READ_MAX_LEN**).
- Filtrar las lecturas que contengan un 40% o más de bases con calidad inferior a 7 en la escala Phred.
- Utilizar una ventana deslizante de largo de 10 bases, que vaya de 3' a 5', y que recorte las lecturas con una calidad promedio de 7 (en escala **phred**).
- Utilizar una ventana deslizante de largo de 10 bases, que vaya de 5' a 3', y que recorte las lecturas con una calidad promedio de 7 (en escala **phred**).

Estos filtros se llevarán a cabo utilizando la herramienta **Fastp**, mediante el siguiente comando:

```
$ fastp -i output/demultiplexing/<SAMPLE>.fastq \
-o output/trimming/<SAMPLE>_trimmed.fastq \
--length_required 30 \
--length_limit <READ_MAX_LEN> \
--qualified_quality_phred 7 \
--unqualified_percent_limit 40 \
--cut_front \
--cut_tail \
--cut_window_size 10 \
--cut_mean_quality 7 \
--thread <THREADS> \
--dont_eval_duplication \
--disable_trim_poly_g \
--disable_adapter_trimming \
--disable_quality_filtering \
--dont_overwrite

$ mv "./fastp.json" output/trimming/fastp_0<ID>.json
$ mv "./fastp.html" output/trimming/fastp_0<ID>.html
```

Es importante notar que la variable **<ID>** debe permitir identificar la muestra que se acaba de procesar.

Una vez que se ha ejecutado el comando, se puede realizar nuevamente un control de calidad para evaluar cómo han mejorado los resultados. Sin embargo, hay que tener en cuenta la nueva ruta de los archivos.

```
$ fastqc --memory 4096 output/trimming/<SAMPLE>_trimmed.fastq \
-o ./output/quality
```

Sin embargo, si sólo se desean analizar parámetros como la longitud promedio, el número de lecturas o la distribución de la calidad, se puede consultar el archivo **fastp.html**, el cual se genera durante el proceso de *trimming*. Este archivo contiene información sobre la calidad de las muestras antes y después del *trimming*. No obstante, a diferencia de **FastQC**, no presenta indicadores del nivel de calidad de los datos ni la distribución de la longitud de las lecturas.

5.2.4 Mapeo

Para mapear las lecturas, se utilizará el genoma de referencia correspondiente al cromosoma 16, utilizando la versión **hg38**, el cual puede ser descargado desde la página del **NCBI**⁴⁵ (National Library of Medicine). Es fundamental asegurarse de que el encabezado del archivo **FASTA** del cromosoma 16 comience con "**chr16**". Si este no es el caso, será necesario editararlo, ya sea manualmente con un editor de texto. En la Figura 6 se muestra un ejemplo donde el "*accession number*" del genoma del cromosoma 16 fue reemplazado por "**chr16**".

```
Header original:  
>NC_000016.10 Homo sapiens chromosome 16, GRCh38.p14 Primary Assembly  
  
Header modificado:  
>chr16 Homo sapiens chromosome 16, GRCh38.p14 Primary Assembly
```

Figura 6: Ejemplo de modificación del encabezado.

Las lecturas serán mapeadas utilizando el software **Minimap2**, habilitando el parámetro **lr:hq**, que es específico para muestras provenientes de secuenciadores de nanoporos

```
$ minimap2 -t <THREADS> -ax lr:hq \  
<REF_DIR>/<REF> \  
output/trimming/<SAMPLE>_trimmed.fastq > output/mapping/<SAMPLE>.sam
```

5.2.5 Control de calidad del mapeo, filtrado y preprocesamiento de la salida

Una vez mapeadas las lecturas, se obtendrá un archivo en formato **SAM** como resultado. El primer paso será analizar el porcentaje de lecturas que fueron correctamente mapeadas. Esto se puede lograr utilizando la herramienta **Samtools**, ejecutando el siguiente comando:

```
$ samtools flagstats ./output/mapping/<SAMPLE>.sam
```

Este comando proporcionará un resumen estadístico, incluyendo el porcentaje de lecturas mapeadas en el genoma de referencia.

Una vez analizado el porcentaje de lecturas que se lograron mapear, se procederá a analizar el número de lecturas que se mapean en cada región objetivo, así como el nivel de profundidad de lectura obtenido. Sin embargo, antes de esto, es necesario realizar un preprocesamiento de las lecturas. En primer lugar, se debe transformar el archivo **SAM** a un formato **BAM** utilizando el siguiente comando:

```
$ samtools view -S -b output/mapping/<SAMPLE>.sam > \  
output/mapping/<SAMPLE>.bam
```

A continuación, se ordenan las muestras según su posición en el genoma de referencia:

```
$ samtools sort -o output/mapping/<SAMPLE>.bam output/mapping/<SAMPLE>.bam
```

⁴⁵ [https://www.ncbi.nlm.nih.gov/nuccore/NC_000016.10?report=fasta&log\\$=seqview](https://www.ncbi.nlm.nih.gov/nuccore/NC_000016.10?report=fasta&log$=seqview)

A continuación, se conservarán únicamente las lecturas mapeadas como primarias, que representan la mejor alineación posible para cada lectura.

```
$ samtools view -b -h -F 0x100 -F 0x800 \
    output/mapping/<SAMPLE>.bam > aux.bam
```

Después, se eliminarán las lecturas que no fueron mapeadas en las regiones objetivo:

```
$ samtools view -L data/regions.bed -b \
    aux.bam > output/mapping/<SAMPLE>.bam
```

Finalmente, se indexará el archivo **BAM** para mejorar la velocidad de procesamiento en las siguientes etapas:

```
$ samtools index output/mapping/<SAMPLE>.bam
```

Volviendo al análisis de las lecturas, se utilizará la herramienta **Bedtools** para evaluar la cobertura alcanzada por las lecturas en cada región. Primero, se convertirá el archivo **BAM** a formato **BED** y luego se ordenará con los siguientes comandos:

```
$ bedtools bamtobed -i output/mapping/<SAMPLE>.bam > \
    output/mapping_qc/<SAMPLE>.bed
$ bedtools sort -i output/mapping_qc/<SAMPLE>.bed > \
    output/mapping_qc/<SAMPLE>_sorted.bed
```

Una vez que se haya ordenado el archivo, se puede calcular la cobertura utilizando el siguiente comando:

```
$ bedtools coverage -a data/regions.bed \
    -b output/mapping_qc/<SAMPLE>_sorted.bed > \
    output/mapping_qc/<SAMPLE>_coverage.bed
```

El archivo **BED** generado por este comando puede ser visualizado en **Excel**. Este archivo presenta estadísticas importantes, como el número de lecturas mapeadas y el porcentaje de bases cubiertas en cada región, lo que resulta útil para evaluar si las regiones secuenciadas han sido adecuadamente cubiertas. A continuación, se detalla el orden de las columnas:

1. Cromosoma
2. Posición inicial
3. Posición final
4. Número de lecturas mapeadas en la región
5. Número de bases cubiertas
6. Número de bases de la región
7. Porcentaje de bases cubiertas

Por último, se evaluará la profundidad de lectura obtenida. Para ello, se utilizará nuevamente la herramienta **Samtools** junto con un *script* de **Python** de **variant_analysis_pipeline**. Primero, se generará un archivo que contiene el número de veces que se ha leído cada posición en el genoma:

```
$ samtools depth output/mapping/<SAMPLE>.bam > <SAMPLE>_depth_report.tsv
```

A continuación, se pasará como argumento al *script* de **Python** el archivo generado, el archivo con las regiones objetivo y la ruta donde se almacenará el archivo con el reporte:

```
$ python <PIP_PATH>/depth_report.py <SAMPLE>_depth_report.tsv \
    data/regions.bed \
    output/mapping_quality/<SAMPLE>_depth_report.tsv
```

Este comando generará un reporte que, al igual que en el caso anterior, puede visualizarse en **Excel**. Este reporte muestra el porcentaje de muestras que han sido leídas un número determinado de veces para cada una de las regiones secuenciadas. A continuación, se detallan las columnas que aparecen en el archivo de salida:

1. Cromosoma
2. Posición inicial de la región.
3. Posición final de la región.
4. Porcentaje de bases leídas más de una vez
5. Porcentaje de bases leídas más de 30 veces
6. Porcentaje de bases leídas más de 50 veces
7. Porcentaje de bases leídas más de 100 veces
8. Porcentaje de bases leídas más de 500 veces
9. Porcentaje de bases leídas más de 1000 veces

Se recomienda que aproximadamente el 100% de las bases sean leídas con una profundidad de al menos 500x. Esto asegura una mayor confiabilidad en el análisis y mejora la precisión de las etapas posteriores del proceso.

Finalmente, una vez completados estos pasos, se pueden eliminar el archivo **BAM** temporal **aux.bam**, así como los archivos **BED** y **TSV** temporales utilizados.

```
$ rm aux.bam
$ rm output/mapping_qc/<SAMPLE>_sorted.bed
$ rm output/mapping_qc/<SAMPLE>.bed
$ rm <SAMPLE>_depth_report.tsv
```

5.2.6 Variant calling

Una vez confirmada la cobertura de las regiones de interés, se procederá a la etapa de *variant calling*. Sin embargo, antes es necesario indexar el genoma de referencia. Para ello, se utilizará la herramienta **Samtools** con el siguiente comando:

```
$ samtools faidx <REF_DIR>/<REF>
```

Para llevar a cabo esta etapa, se utilizará la herramienta **DeepVariant**. Esta herramienta se ejecuta a través de **Docker** con el comando que se puede ver a continuación. En este comando, hay varios parámetros nuevos que deben ser modificados por el usuario. Estos son:

- **INPUT_DIR**: ruta absoluta de la carpeta donde se encuentran las muestras mapeadas.
- **OUTPUT_DIR**: ruta absoluta de la carpeta donde se almacenarán los resultados (la carpeta debe ser única para cada muestra y no debe estar creada previamente; por ejemplo: /home/user/working_dir/output/variant_calling/sample_x).

```
$ docker run \
-v <INPUT_DIR>/input \
-v <REF_DIR>/ref \
-v <OUTPUT_DIR>/output \
google/deepvariant:1.6.1 \
/opt/deepvariant/bin/run_deepvariant \
--model_type=ONT_R104 \
--ref="/ref/<REF>" \
--reads="/input/<SAMPLE>.bam" \
--regions="/ref/regions.bed" \
--output_vcf=/output/output.vcf.gz \
--output_gvcf=/output/output.g.vcf.gz \
--intermediate_results_dir /output/intermediate_results_dir \
--num_shards=<THREADS>
```

Una vez completada la etapa de variant calling, **DeepVariant** generará múltiples archivos en la carpeta especificada como **OUTPUT_DIR**. Para obtener el archivo que contiene las variantes y almacenarlo en la carpeta **variant_calling**, primero se debe descomprimir el archivo con el siguiente comando:

```
$ zcat <OUTPUT_DIR>/output.vcf.gz > \
output/variant_calling/<SAMPLE>.vcf
```

Finalmente, **DeepVariant** incluye un filtro denominado "**RefGene**", que permite evaluar la relevancia clínica de las variantes en función de su ubicación en el genoma. Sin embargo, **DeepVariant** no elimina automáticamente las variantes que no cumplen con este filtro, por lo que es necesario filtrarlas antes de continuar. Para ello, se utilizará la herramienta **Bcftools** con el siguiente comando:

```
$ bcftools view -f PASS -O v \
-o output/variant_calling/<SAMPLE>_pass.vcf \
output/variant_calling/<SAMPLE>.vcf
```

5.2.7 Anotación de variantes

Una vez obtenidas las variantes, es momento de recopilar la evidencia y predecir sus posibles efectos. Esto se logra mediante la herramienta **ANNOVAR**, utilizando el siguiente comando:

```
$ table_annovar.pl output/variant_calling/<SAMPLE>_pass.vcf <DB> \
-buildver hg38 \
-out output/annotation/<SAMPLE> \
-remove \
-protocol \
refGene,avsnp150,dbnsfp42c,clinvar_20240611,intervar_20180118,abraom,g \
nomad41_genome,gnomad41_exome,esp6500siv2_all,1000g2015aug_all \
-operation g,f,f,f,f,f,f,f,f \
-nastring . \
-vcfinput \
-polish
```

Este comando genera tres tipos de archivo. El primero es uno con la extensión **.avinput**, que sirve como entrada para **ANNOVAR** y contiene la información básica de cada variante recopilada por la herramienta de *variant calling*. El segundo archivo tiene la extensión **.txt**, pero en realidad cumple con el formato **TSV (Tab-Separated Values)**. Este archivo contiene la información recopilada por **ANNOVAR** para cada una de las variantes, así como la

información del archivo **.avinput**. Por último, el tercer archivo tiene la extensión **VCF**, que es el archivo original utilizado como entrada, al cual se le ha incorporado la información recopilada por **ANNOVAR**.

5.2.8 Sistema de *ranking* de variantes

Una vez anotadas las variantes, es el momento de puntuarlas según la evidencia recopilada. Esto se llevará a cabo mediante el *script* de **Python ranking_system.py**, que integra la evidencia con las variantes reportadas por la **Clinica Mayo**. Posteriormente, el *script* calificará cada variante y generará un archivo en formato **TSV** que contiene la lista de variantes junto con su respectiva puntuación y evidencia.

```
$ python <PIP_PATH>/ranking_system.py \
    output/annotation/<SAMPLE>.hg38_multianno.txt \
    output/ranking/<SAMPLE>.tsv
```

Las variantes estarán ordenadas de acuerdo a varios criterios: si cumplen con la frecuencia alélica mínima requerida, si cumplen con la frecuencia alélica de variante establecida, su puntaje asignado y su posición en el genoma. El archivo generado por este comando puede abrirse en **Excel**. En él, se visualizan 16 columnas, cuyas descripciones se detallan en la Tabla 5.

Tabla 5: Descripción de las columnas del archivo del sistema de *ranking*.

Columna	Descripción
Start	Posición de inicio de la variante en el cromosoma.
Ref	Base de referencia.
Alt	Variante encontrada.
Region	Intrón o exón del gen PKD1 , donde se encuentra la variante.
Score	Puntaje obtenido en el sistema de <i>ranking</i> de variantes (Véase Anexo A).
Clin_Mayo_Clas	Clasificación de la variante en Clinica Mayo .
CLNSIG	Clasificación de la variante en ClinVar .
InterVar_automated	Clasificación de la variante según InterVar .
SIFT_pred	Clasificación de la variante según SIFT .
LRT_pred	Clasificación de la variante según LRT .
MutationTaster_pred	Clasificación de la variante según MutationTaster .
ExonicFunc.refGene	Efecto de la variante (Ejemplo: <i>Stopgain</i> , <i>Frameshift</i> , etc).
Gene.refGene	Gen donde se localiza la variante.
AF_Valid	Booleano que indica si la variante está presente en menos del 5%

	de la población.
VAF_Valid	Booleano que indica si la variante cumple con una frecuencia alélica mayor o igual a 0.3 y menor o igual a 0.7.
FORMAT	Valores de la columna “FORMAT” del archivo VCF original.
default	Valores de la columna “default” del archivo VCF original.

5.3 Ejecución automatizada mediante script

Para simplificar la ejecución del *pipeline*, se desarrolló un *script* en **Bash** que automatiza todas las etapas del proceso. Este *script* permite ejecutar el *pipeline* desde cualquier etapa, lo que brinda mayor versatilidad.

Antes de ejecutar el *script*, es fundamental activar el entorno de **Conda** previamente configurado. Para ello, debe ejecutarse el siguiente comando:

```
$ conda activate pipeline_PKD1
```

Con el entorno activado, el *pipeline* puede ejecutarse con el siguiente comando:

```
$ bash <PIP_PATH>/pipeline_PKD1.sh <CONFIG_FILE>
```

En el comando anterior, se observa que el *script* recibe el parámetro **<CONFIG_FILE>**, el cual corresponde a un archivo de texto que contiene los parámetros necesarios para la ejecución de cada etapa del *pipeline*. Para iniciar, es necesario proporcionar un parámetro que especifique la ubicación de los archivos que servirán como punto de partida del *pipeline*. El nombre de este parámetro dependerá del punto de partida seleccionado para el *pipeline* (ver Tabla 6).

Tabla 6: Parámetros de los archivos de entrada del archivo de configuración.

Parámetro	Descripción
POD5_DIR	Ruta absoluta del directorio donde se encuentran los archivos POD5 (Ejecuta el <i>pipeline</i> completo).
FASTQ_DIR	Ruta absoluta del directorio donde se encuentran las muestras en formato FastQ (El <i>pipeline</i> se ejecutará desde el control de calidad)
SAM_DIR	Ruta absoluta del directorio donde se encuentran las muestras en formato SAM (El <i>pipeline</i> se ejecutará desde el control de calidad del mapeo)
BAM_DIR	Ruta absoluta del directorio donde se encuentran las muestras en formato BAM (El <i>pipeline</i> se ejecutará desde el <i>variant calling</i>)
VCF_DIR	Ruta absoluta del directorio donde se encuentran las muestras

	en formato VCF (El <i>pipeline</i> se ejecutará desde el el proceso de anotación y clasificación)
--	--

Una vez definido el punto de inicio del *pipeline*, es necesario proporcionar algunos parámetros requeridos según el punto de partida. Estos parámetros se detallan en la Tabla 7.

Tabla 7: Parámetros del archivo de configuración.

Parámetro	Descripción
LIBRARY_KIT	Nombre del kit de preparación de librerías utilizado (Ejemplo: SQK-NBD114-24). Requerido cuando se ejecuta el <i>pipeline</i> desde el basecalling .
REF	Nombre del archivo FASTA , con el genoma de referencia (Ejemplo: GRCh38.fasta). Requerido si se ejecuta el <i>pipeline</i> desde el variant calling o antes.
REF_DIR	Dirección absoluta de la carpeta donde está ubicado el genoma de referencia (Ejemplo: /home/user/working_dir/ref_seq. Requerido si se ejecuta el <i>pipeline</i> desde el variant calling o antes.
REGIONS	Archivo en formato BED que incluye las regiones amplificadas durante la PCR . Requerido si se ejecuta el <i>pipeline</i> desde el variant calling o antes.
REGIONS_DIR	Dirección absoluta de la carpeta donde está ubicado el archivo con las regiones secuenciadas (Ejemplo: /home/user/working_dir/ref_seq). Requerido si se ejecuta el <i>pipeline</i> desde el variant calling o antes.
DB	Dirección del directorio donde se almacenan las base de datos descargadas de ANNOVAR. Requerido si se ejecuta el <i>pipeline</i> desde la etapa de anotaciones y clasificación o antes.
THREADS	Número de núcleos disponibles para la ejecución en paralelo de las tareas. Parámetro obligatorio.
READ_MAX_LEN	Largo máximo de las lecturas (Para la etapa de <i>trimming</i>). Requerido si se ejecuta desde el control de calidad .
DEMULTIPLEXING	Booleano (true o false), que representa si se realiza o no el proceso de demultiplexación (Requerido si se define POD5_DIR).
N_SAMPLES	Número de muestras que se multiplexaron (Requerido si <DEMULTIPLEXING> está definido como “true”)
LAST_STAGE	Valor numérico que indica la última etapa que se ejecutará en el <i>pipeline</i> . Las etapas disponibles son: 1 para <i>basecalling</i> y demultiplexación, 2 para control de calidad, 3 para <i>trimming</i> y filtrado, 4 para mapeo, 5 para control de calidad del mapeo, 6 para

	<i>variant calling</i> , 7 para anotación y 8 para el sistema de ranking. Este parámetro es opcional; si no se proporciona, el <i>pipeline</i> se ejecutará completamente.
--	--

Por último, en la Figura 7 se muestra un ejemplo de cómo se vería un archivo de configuración que comienza en la etapa de *basecalling* y requiere demultiplexación. En esta figura, se omite el parámetro <LAST_STAGE>, por lo que el *pipeline* se ejecutará de forma completa.

```
LIBRARY_KIT="SQK-NBD114-24"
REF_DIR="/home/user/Documentos/working_dir/ref_seq"
REF="GRCh38.fasta"
REGIONS_DIR="/home/user/Documentos/working_dir/ref_seq"
REGIONS="regions.bed"
DB="/home/user/Documents/humandb"
THREADS=4
READ_MAX_LEN=6000
POD5_DIR="/home/user/Documents/pod5"
DEMULTIPLEXING=true
N_SAMPLES=8
```

Figura 7: Ejemplo de archivo de configuración.

El resultado de cada etapa del análisis se encuentra en la carpeta “*output*”, la cual sigue la misma estructura presentada en la Figura 2.

ANEXO

Anexo A

El sistema de *ranking* de variantes es el método empleado por el equipo de **GEMINI** para **evaluar y priorizar las variantes según su potencial patogénico**. Esta etapa es parte del proceso de **clasificación, anotación y filtrado de variantes**. Este sistema fue diseñado originalmente para clasificar variantes de muestras generadas por el secuenciador **Illumina MiSeq** y fue desarrollado por el bioinformático Cristian Yañez en colaboración con Paola Krall. Posteriormente, el sistema fue implementado mediante un *script* en **Python** y se ajustaron algunos parámetros para adaptarlo al análisis de datos generados por el secuenciador *Oxford Nanopore MinION*, como parte del trabajo de título de Diego Millar.

Las variantes que se busca clasificar en esta sección son aquellas identificadas previamente por la herramienta de *variant calling* (**DeepVariant**) y que cuentan con anotaciones generadas a partir de *software* de anotación de variantes (**ANNOVAR**).

El método de evaluación de variantes se fundamenta en diversos factores, como la clasificación de las variantes en distintas bases de datos, las predicciones de patogenicidad proporcionadas por programas especializados, el tipo de variante y su ubicación en el gen. Además, se considera la clasificación final de la variante tras aplicar la guía propuesta por la **ACMG/AMP** mediante el *software* **InterVar**.

Este sistema de *ranking* asigna puntaje a las variantes con un algoritmo personalizado. El cual califica las variantes con un puntaje que va de 0 a 100 puntos, donde un puntaje de 100 es considerado como una variante patogénica, mientras que un puntaje igual a cero se interpreta como una variante benigna.

Se descartan las variantes con una frecuencia alélica en la población igual o superior al 5%, así como aquellas con una frecuencia alélica de la variante (VAF) menor o igual a 0.3 o mayor a 0.7. En el caso del *script* del sistema de *ranking* no elimina estas variantes, sino que las clasifica en un nivel inferior, por debajo de aquellas que cumplen con estos criterios.

Finalmente, el sistema de *ranking* propone los siguientes criterios de evaluación recopilados en la Tabla 8. Tras evaluar cada variante en cada uno de los criterios, se le asigna puntaje a la variante, siguiendo el siguiente diagrama de flujo presentado en la Figura 8.

Siguiendo estas instrucciones para cada una de las variantes encontradas, se obtiene una lista de las variantes, la cual puede ser ordenada según la probabilidad de ser responsables de la enfermedad, en base a la evidencia presentada. Sin embargo, solo se puede tener un alto grado de certeza de que son variantes patogénicas aquellas que han sido calificadas con 80 puntos o más (Códigos C1 a C4). Si no se obtiene ninguna variante que supere ese puntaje, entonces se deberá evaluar la posibilidad de realizar un nuevo análisis que incluya a los familiares afectados y sanos, con el fin de conseguir más evidencia que pueda respaldar la patogenicidad de alguna de las variantes encontradas.

Tabla 8: Criterios para clasificación de variantes sistema de *ranking* Yáñez-Krall.

Código	Criterion
C1	RefGene clasifica la variante como “ <i>Stopgain</i> ”.
C2	Clinvar , Intervar o Clínica Mayo clasifican la variante como “ Patogénica ”.
C3	Mutation Taster la clasifica como “A” (“ <i>disease_causing_automatic</i> ”).
C4	RefGene la clasifica como “ <i>splicing</i> ”, “ <i>frameshift deletion</i> ”, “ <i>frameshift insertion</i> ”.
C5	Mutation Taster la clasificada como “D” (“ <i>disease_causing</i> ”).
C6	LRT la clasifica como “D” (<i>Deleterious</i>).
C7	Sift la clasifica como “D” (<i>Deleterious</i>).
C8	Variante clasificada como “ <i>frameshift deletion</i> ” o “ <i>frameshift insertion</i> ”.

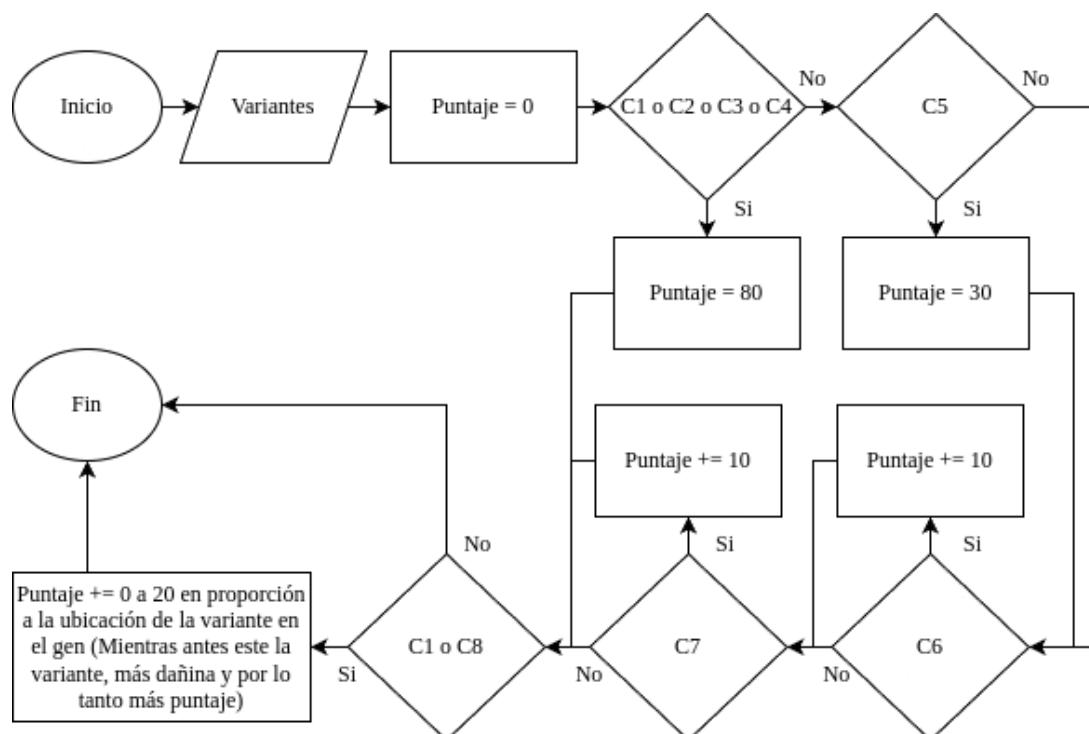


Figura 8: Diagrama de flujo para la calificación de variantes.