UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Sistemas Informáticos y Computación

Detección de Variantes Genómicas utilizando Deep Learning

Trabajo Fin de Máster

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

AUTOR/A: Parres Montoya, Daniel

Tutor/a: Paredes Palacios, Roberto

CURSO ACADÉMICO: 2021/2022

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MIARFID Official Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València

Detección de Variantes Genómicas utilizando Deep Learning

TRABAJO DE FIN DE MÁSTER

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Autor: Daniel Parres Montoya
Tutor: Roberto Paredes Palacios

Curso 2021-2022

Resum

La detecció de variants d'ADN és una tasca important i difícil a la Genòmica. Les tecnologies de seqüenciació d'una sola molècula han aconseguit revolucionar la detecció de variants genètiques, l'acoblament de genomes complexos i la detecció de marques epigenètiques (marques químiques que s'afegeixen a l'ADN i permeten el funcionament correcte) en els últims anys. En el cas de la seqüenciació duna sola molècula, es té una taxa derror per nucleòtid dentre 5 i 15%. Per això, malgrat els ràpids avenços en les tecnologies de seqüenciació, és difícil detectar variants del genoma d'un individu a partir de milers de milions de lectures de seqüències curtes i errònies.

A causa de la massiva quantitat de lectures del genoma i l'alta taxa d'error de seqüenciació, aplicar *Deep Learning* per a l'aprenentatge de relacions entre les seqüències d'ADN, és l'enfocament que obté millors resultats, superant les eines d'última generació i mètodes tradicionals de detecció de variants.

La naturalesa de la tasca que es proposa emmarca aquest Treball Final de Màster dins de la Genòmica, concretament al problema de Detecció de Variants Genòmiques aplicant Tècniques d'Intel·ligència Artificial i del Reconeixement de Formes. Com a primer objectiu es realitza un estudi de les diferents eines de darrera generació, enfocaments i particularitats. El segon objectiu se centra en el tractament i la manipulació de les dades genòmiques per tal de desenvolupar una metodologia robusta. A la metodologia s'inclou l'anàlisi i el disseny d'arquitectures de Xarxes Neuronals Artificials per al problema que es tracta i els seus experiments corresponents. Finalment, es discuteixen els resultats obtinguts i es declaren futurs desenvolupaments per continuar noves investigacions en aquesta línia.

Paraules clau: Medicina de precisió, Detecció de Variants, Deep Learning, Genòmica, ADN

Resumen

La detección de variantes de ADN es una tarea importante y difícil en la Genómica. Las tecnologías de secuenciación de una sola molécula han conseguido revolucionar la detección de variantes genéticas, el ensamblaje de genomas complejos y la detección de marcas epigenéticas (marcas químicas que se añaden al ADN y permiten su correcto funcionamiento) en los últimos años. En el caso de la secuenciación de una sola molécula, se tiene una tasa de error por Nucleótido de entre 5 y 15 %. Por lo que a pesar de los rápidos avances en las tecnologías de secuenciación, es difícil detectar variantes del genoma de un individuo a partir de miles de millones de lecturas de secuencias cortas y erróneas.

Debido a la masiva cantidad de lecturas del genoma y la alta tasa de error de secuenciación, aplicar *Deep Learning* para el aprendizaje de relaciones entre las secuencias de ADN, es el enfoque que mejores resultados obtiene, superando a las herramientas de última generación y métodos tradicionales de detección de variantes.

La naturaleza de la tarea que se propone enmarca este Trabajo Final de Máster dentro de la Genómica, concretamente en el problema de Detección de Variantes Genómicas aplicando Técnicas de Inteligencia Artificial y del Reconocimiento de Formas. Como primer objetivo se realiza un estudio de las diferentes herramientas de última generación, enfoques y particularidades. El segundo objetivo se centra en el tratamiento y manipulación de los datos genómicos con el fin de desarrollar una metodología robusta. En la metodología se incluye el análisis y diseño de arquitecturas de Redes Neuronales Artificiales para el problema que se trata y sus correspondientes experimentos. Finalmente, se discuten los resultados obtenidos y se declaran futuros desarrollos para continuar nuevas investigaciones en esta línea.

Palabras clave: Medicina de precisión, Detección de Variantes, Deep Learning, Genómica, ADN

Abstract

DNA variant detection is an important and challenging task in Genomics. Single-molecule sequencing technologies have revolutionized the detection of genetic variants, the assembly of complex genomes and the detection of epigenetic marks (chemical marks that are added to DNA and allow it to function correctly) in recent years. In the case of single molecule sequencing, there is a per nucleotide error rate of between 5 and 15%. So despite rapid advances in sequencing technologies, it is difficult to detect variants of an individual's genome from billions of short and erroneous sequence reads.

Due to the massive amount of genome reads and the high sequencing error rate, applying *Deep Learning* for learning relationships between DNA sequences is the approach that yields the best results, outperforming state-of-the-art tools and traditional variant detection methods.

The nature of the proposed task frames this Master's Thesis within Genomics, specifically in the problem of Genomic Variant Calling by applying Artificial Intelligence and Shape Recognition Techniques. The first objective is a study of the different state-of-the-art tools, approaches and particularities. The second objective focuses on the treatment and manipulation of genomic data in order to develop a robust methodology. The methodology includes the analysis and design of Artificial Neural Network architectures for the problem at hand and their corresponding experiments. Finally, the results obtained are discussed and future developments are stated in order to continue new research in this line.

Key words: Precision Medicine, Variant Calling, Deep Learning, Genomics, DNA

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

En este primer capítulo se expone la motivación y se introduce el problema de Detección de Variantes Genómicas. También se presentan diferentes Fuentes de Datos y las principales características de cada una. Finalmente, se exponen los objetivos del Trabajo de Fin de Máster y su estructura.

1.1 Motivación

Este Trabajo de Fin de Máster se enmarca dentro de la Genómica. La **Genómica** es un campo de la biología que se centra en el estudio de todo el Ácido Desoxirribonucleico ADN de los organismos. Se denomina genoma a la secuencia completa de ADN de un individuo. Este campo de la biología tiene como objetivos principales la **identificación y caracterización de todos los genes y elementos funcionales de la secuencia genética de un organismo**. [1]

Gracias a la Genómica se puede **diagnosticar, predecir y prevenir enfermedades, así como desarrollar tratamientos personalizados y efectivos**, ahí es donde nace la Medicina de Precisión o Personalizada. [2]

La **Medicina de Precisión** es un nuevo enfoque de la medicina que se **centra en el individuo, sus genes, su estilo de vida y su entorno**. Debido a esto, uno de los principales objetivos es: el estudio y análisis de la variabilidad genética en el desarrollo de enfermedades. [3]

El ADN se almacena como un código compuesto por cuatro Bases químicas, Adenina (A), Guanina (G), Citosina (C) y Timina (T). En los humanos, el ADN consta de aproximadamente 3 mil millones de Bases, donde más del 99 % de esas Bases son iguales en todas las personas. Dependiendo del orden o secuencia de las Bases se determina la información disponible para construir y mantener un organismo. El estudio del genoma ayuda a entender el funcionamiento y desarrollo de los organismos, así como tratar y prevenir enfermedades. [4]

Debido a que el principal eje de la Medicina de Precisión es el estudio y análisis de la variabilidad genética del individuo, este trabajo se vertebra en el problema de Detección de Variantes Genómicas o en inglés *Variant Calling*. Ya que conocer las variaciones que sufre un individuo permite su posterior estudio, para el desarrollo de tratamientos personalizados y efectivos.

1.2 Detección de Variantes Genómicas

Los agentes esenciales para la mayoría de procesos celulares son conocidos como **proteínas**. En la Figura 1.1 se presenta el **Dogma Central de la Biología**, donde se aprecia que una molécula de Ácido Desoxirribonucleico (ADN) codifica las instrucciones básicas para la síntesis de proteínas.

Para los biólogos y médicos, el proceso que se describe en el Dogma Central de la Biología (Figura 1.1) es de gran interés. Ya que, este se encarga de la codificación y decodificación de las instrucciones básicas de la vida. Aunque dicho proceso puede verse afectado en diferentes etapas, la etapa clave es la secuencia de ADN. Debido a que en la secuencia de ADN se codifica la información de un organismo utilizando cuatro tipos de componentes, conocidos como **Bases**.

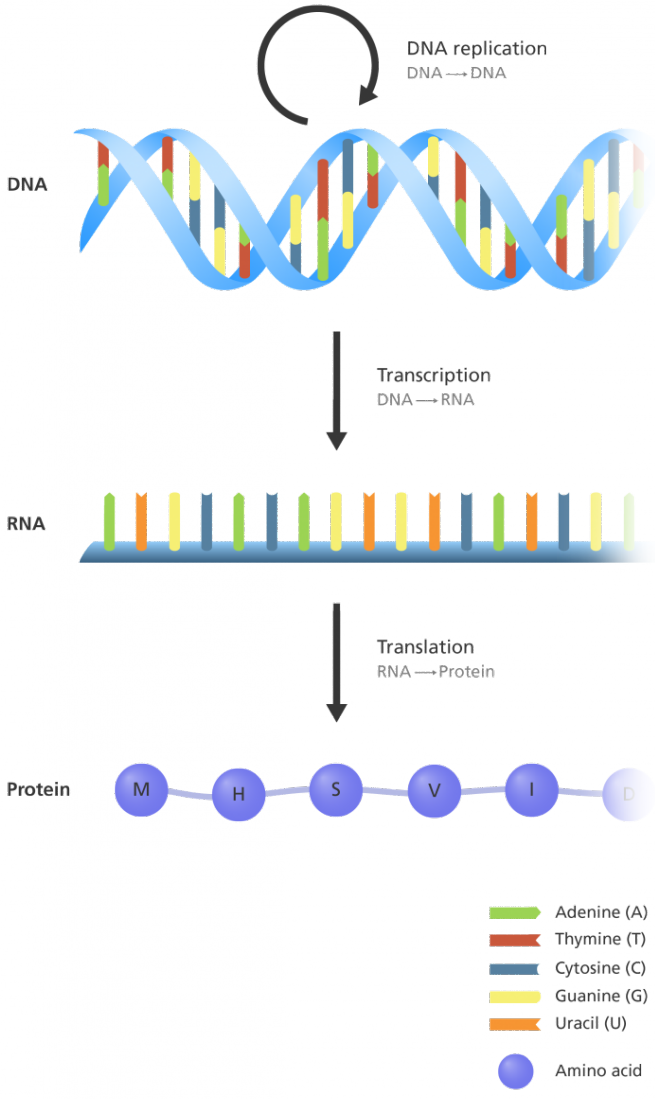

Figura 1.1: Dogma Central de la Biología. [5]

El **ácido desoxirribonucleico** o **ADN** es el material que contiene la información hereditaria en los organismos. Casi todas las células del cuerpo humano contienen el mismo ADN y este se encuentra en el núcleo celular. Cabe destacar que también existe una pequeña cantidad de ADN en las mitocondrias, pero este trabajo se centra en el ADN

nuclear y no en el mitocondrial. A**l conjunto completo de ADN de un organismo se le denomina Genoma**.

Tal y como se ha comentado anteriormente, el ADN se almacena como un código compuesto por cuatro Bases químicas o Nucleótidos: Adenina (A), Guanina (G), Citosina (C) y Timina (T). En los humanos, el ADN consta de aproximadamente 3 mil millones de Bases, donde más del 99 % de esas Bases son iguales en todas las personas. Dependiendo del orden o secuencia de las Bases se determina la información disponible para construir y mantener un organismo. El estudio del genoma ayuda a entender el funcionamiento y desarrollo de los organismos, así como tratar y prevenir enfermedades.

El genoma de un organismo se puede obtener mediante el **proceso de secuenciación del ADN**. Las tecnologías de secuenciación han ido evolucionando en los últimos años dando lugar a datos de alta resolución y bajo coste. Estas tecnologías se conocen como *Next Generation Sequencing* (NGS) y *Third Generation Sequencing* (TGS). A la hora de secuenciar el ADN de un individuo con las tecnologías NGS o TGS se obtiene cada base varias veces, por lo que las principales ventajas son: una mayor redundancia y mayor relación señal/ruido.

El proceso de secuenciación consta de varias lecturas de ADN en una misma localización para posteriormente alinearlas con un Genoma de referencia. Esto se realiza así debido a que actualmente no se puede secuenciar el genoma entero con una única lectura.

Para analizar los datos secuenciados de ADN y detectar variaciones, se necesita comparar con una secuencia de referencia estándar o también conocido como Genoma de referencia. En los últimos años se han desarrollado diferentes técnicas para el análisis de secuencias, que permiten caracterizar secuencias del genoma de un organismo o una población en función de sus variaciones.

Existen diferentes tipos de variaciones, cambios en posiciones aisladas o en la estructura de miles de Bases. La detección de variantes tiene un gran abanico de aplicaciones, como el estudio de enfermedades [6] o mejora de cultivos [7]. [8]

Como se ha comentado anteriormente, hay diversos tipos de variaciones, pero las más comunes son: Polimorfismo Nucleótido Único o *Single Nucleotide Polymorphism* (SNP) y las Inserciones-Deleciones o *Insertions-Deletions*(Indels). En la Tabla 1.1 se muestran diferentes ejemplos de los tipos de variaciones, donde en la segunda columna de la tabla se muestra la secuencia de referencia y en la tercera columna la secuencia alternativa donde ocurre una variación determinada, marcada de color azul.

Tipo	Secuencia de Referencia	Secuencia Alternativa
SNP	T	G
Insertion	AGT	A C GT
	ATCGGG	ATC TGA GGG
Deletion	A C GT	AGT
	ATC TGA GGG	ATCGGG

Tabla 1.1: Tipos de variaciones. [47]

Las lecturas de ADN que producen las tecnologías *Next Generation Sequencing* (NGS) son de corta longitud, concretamente entre 50 y 250 Bases. Además, tienen tasas de error bajas y se caracterizan también por ser más económicas que las *Third Generation Sequencing* (TGS). Debido a esto, son el estándar para la secuenciación de genomas. Por otro lado, las tecnologías TGS proporcionan lecturas más largas que las NGS, entre 1 y 50 kilobases.

En los últimos años se han diseñado diferentes tecnologías de detección de variantes. Las herramientas tradicionales más destacadas son Samtools [9] y The Genome Analysis Toolkit (GATK) [10]. Estas herramientas tienen métricas de rendimiento altas cuando se promedian con todo el genoma, pero existen regiones específicas del genoma donde el rendimiento es inferior [11]. [8]

Debido a esas zonas del genoma donde el rendimiento de las herramientas tradicionales es inferior, es una necesidad desarrollar tecnologías capaces de obtener buenas métricas en todo el genoma y sobre todo en las regiones difíciles de identificar. En el caso de las tecnologías tradicionales, se debe de analizar las suposiciones simplistas y las heurísticas implementadas para desarrollar herramientas capaces de ir más allá en el entendimiento de la naturaleza de los datos y el reconocimiento de patrones. [8]

A la hora de identificar variantes, se trata de delinear los errores que proceden de la secuenciación del ADN y el preprocesamiento de los datos. Por lo que juega un rol importante la tecnología de secuenciación. Con la mejora de las tecnologías de secuenciación pueden surgir nuevos marcos de trabajo, como podría ser el uso de múltiples tecnologías de secuenciación al mismo tiempo con el fin de aumentar la fiabilidad de la identificación. [8]

Se debe destacar que el estudio y análisis en este campo, genera conocimientos sobre las variantes importantes que desarrollan diferentes enfermedades. En el caso de cáncer, la principal causa es un defecto en el genoma, que puede ser heredado [13] o desarrollado a lo largo de la vida de un organismo [12]. [8]

Para la detección de variantes se utilizan secuencias de ADN obtenidas de diferentes tecnologías, donde cada una tiene sus propias características y diversas tasas de error. Muchos métodos utilizan el conocimiento experto y heurísticas para resolver esta tarea, y a pesar de ello, existen regiones del genoma difíciles donde estas herramientas no consiguen obtener buenas métricas. Ahí es donde entra el campo del *Machine Learning* y la Redes Neuronales Artificiales, capaces de abarcar tareas complejas sin necesidad de implementar explícitamente conocimiento experto.

En 2018, de la mano de Google, surge la primera propuesta de *Deep Learning* para la detección de variantes llamada DeepVariant [14]. La tecnología desarrollada por Google, comparada con herramientas tradicionales como GATK obtiene mejores métricas en las evaluaciones del estándar de oro (WGS y WES) y también en comparaciones clínicas como se demuestra en [15].

Los datos de secuenciación de tecnologías NGS se estructuran en forma de lecturas cortas de entre 50 y 250 Bases, debido a que estas lecturas son muy cortas en comparación con las 3.000 millones de Bases que tiene el genoma humano, estas pequeñas secuencias tienen que ser mapeadas. Las lecturas secuenciadas se mapean respecto a un Genoma de referencia [17]. A las herramientas que mapean las lecturas de ADN secuenciado con un Genoma de referencia se las denomina alineadores o *aligners*, uno de los más extendidos es Minimap2 [20].

De una forma simplificada, se puede resumir la detección de variantes cómo: a partir de las lecturas que se han mapeado con el Genoma de referencia se puede listar cada posición específica del genoma y ver todas las lecturas mapeadas en dicha posición. De esta forma, se analiza para asumir un consenso entre las Bases mapeadas para determinar cuál es la base real del organismo y compararla con la base del Genoma de referencia. De esta forma, se determina si se ha producido una variante o no. En la Figura 1.2 se presenta un ejemplo donde se tienen cuatro lecturas mapeadas al Genoma de referencia. Y en cierta posición, la referencia tiene una “C” y las lecturas mapeadas tienen “T, T, T y C”, debido a que hay más “Ts” que “Cs” se asume que la base real del individuo es “T”, y se declara que en esa posición hay una variación.

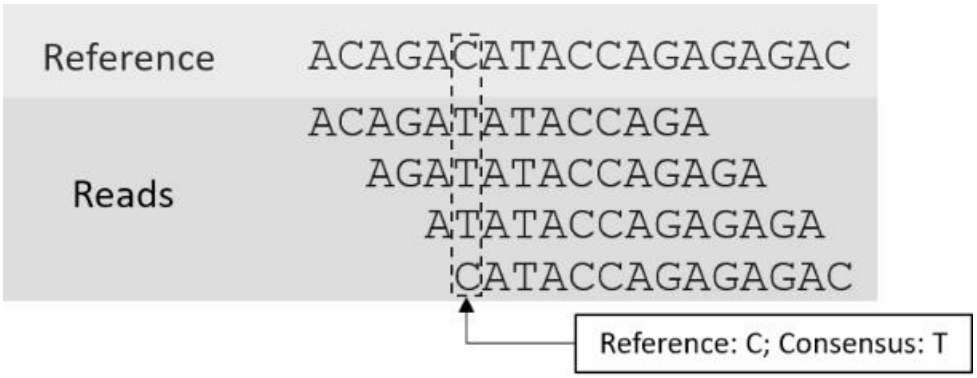

Figura 1.2: Ejemplo simplificado de detección de variantes. [8]

Pero desde un punto de vista práctico, en los métodos de última generación para la detección de variantes, existen dos enfoques principales, tal y como se propone en [18]. El primero es el uso de **modelos estadísticos**, cuyo objetivo es medir la intensidad de la señal que declara una variante. Y el segundo es el uso de **filtros heurísticos que se desarrollan mediante conocimiento experto**, que trata de localizar patrones de errores de secuenciación y alineamiento que pasan desapercibidos por los modelos estadísticos. Este flujo de trabajo obtiene métricas de rendimiento altas en general, pero como se ha mencionado anteriormente, en regiones específicas del genoma el rendimiento es inferior. En [8] se propone un ejemplo sencillo que explica este fenómeno: en regiones del genoma donde la distribución es uniforme para las Bases A, C, G y T es sencillo detectar variantes, pero existen regiones interesantes donde el contenido de G y C es alto o muy bajo y el número medio de lecturas (cobertura) se ve afectado y esta deficiencia puede causar problemas (ya que la cobertura de lecturas es el punto clave para asumir un consenso de la base real). A parte de las regiones difíciles, en [11] se discute que la mayoría de tecnologías no son capaces de analizar eficientemente las regiones de baja complejidad debido a su poca entropía. Como por ejemplo regiones donde surgen las repeticiones de la misma base. Debido a estos casos, el uso de técnicas de *Deep Learning*, proporciona resultados con mejores métricas en el ámbito global y local.

Finalmente, cabe mencionar que existe una división en la detección de variantes: ***Germline Variant Calling*** o ***Somatic Variant Calling***.

En el ***Germline Variant Calling*** se tiene que el Genoma de referencia es el estándar para la especie que se desea estudiar, a partir de este se identifican las variaciones con las lecturas del ADN del individuo a analizar. Dado un *locus* o región específica del genoma, se espera ver uno de los dos siguientes casos:

- **Homocigosidad**: Todas las lecturas tienen la misma base.
- **Heterocigosidad**: Aproximadamente la mitad de todas las lecturas tienen una base y la otra mitad tienen otra.

Mientras que en el ***Somatic Variant Calling*** el Genoma de referencia es un tejido del mismo individuo. En este caso se espera ver "mosaicismo.^{en}tre las células. El mosaicismo se produce cuando un individuo tiene dos o más conjuntos de células genéticamente diferentes en su cuerpo, si el conjunto de células anormales supera al conjunto de células normales, puede dar lugar a enfermedades.

Este trabajo se centra en el *Germline Variant Calling*, por lo que cuando se habla de secuencia de referencia, se está nombrando al genoma humano de referencia completo y no a una secuencia de algún tejido en concreto.

1.3 Fuentes de Datos

Existen diferentes tecnologías de secuenciación y cada una de estas tiene sus propias características y tasas de error. La principal división que se puede realizar entre tecnologías de secuenciación en la actualidad es en *Next Generation Sequencing* (NGS) o *Third Generation Sequencing* (TGS).

En [19] se tratan las principales características de las NGS y TGS, de ahí se puede derivar la Tabla 1.2, donde se presentan las principales ventajas y desventajas de cada tipo. Por lo que a la hora de decidir qué tecnología utilizar, se debe de evaluar y priorizar el entorno y las necesidades.

	NGS	TGS
4*Ventajas	**Alta precisión**	**Genera lecturas de secuenciación muy largas**
	Más barato	Permite comenzar con fragmentos de ADN más largos
	Permite secuencias ADN fragmentado	Marcadores epigenéticos estables
		Facilita la preparación de bibliotecas y tecnologías portátiles
3*Desventajas	**Produce lecturas de secuenciación cortas**	Señales de fragmentos individuales pueden ser débiles
	No distingue ciertos tipos de variaciones	**Precisión general más baja**
	No adecuada en regiones con muchas repeticiones	

Tabla 1.2: Comparativa NGS y TGS.

Una vez contrastadas las diferencias entre NGS y TGS, es interesante mencionar las principales plataformas de secuenciación.

Illumina es la principal plataforma de NGS y comenzó comprando a Solexa, su primera plataforma llamada Genome Analyzer. La puso a la venta en 2007 y fue mejorando sus características a lo largo del tiempo. La segunda máquina de NGS de Illumina fue HiSeq y se comercializó en 2016, a la evolución de HiSeq se la denominó HiSeq X10. Esta conseguía aumentar el número de fragmentos que se podrían analizar. En la actualidad, Illumina proporciona las máquinas NextSeq y NovaSeq junto a un amplio abanico de secuenciadores como iSeq o MiniSeq. [19]

En cuanto a las TGS **Pacific Biosciences** y **Oxford Nanopore Technologies** (ONT) son las más reconocidas. Pacific Biosciences comercializó su primer producto en 2010 llamado PacBio RS. En 2018, Illumina trata de comprar Pacific Biosciences pero se declara como competencia desleal y no se realiza la compra. Por lo que la compañía toma un nuevo rumbo centrándose en comprar la empresa Circulomics que fabrica kits de extracción de ADN de alto peso molecular, que permite la mejora del flujo de trabajo de secuenciación de lecturas largas. [19]

Por otro lado, Oxford Nanopore Technologies es propietario de una tecnología de secuenciación novedosa, que trata de terminar la secuencia de ADN al pasar por un nanoporo. Esta tecnología utiliza corriente iónica y mide los cambios de carga eléctrica a

medida que los Nucleótidos pasan por el nanoporo para determinar las secuencias de Bases que pasan. [19]

Este trabajo se centra en las fuentes de datos de Oxford Nanopore Technologies debido a sus múltiples ventajas y la gran acogida que está obtenido esta tecnología tanto en el ámbito científico como en el clínico. La principal ventaja que destaca la tecnología ONT es su portabilidad y bajo coste. Y muchos expertos califican que es el método más rentable para cartografiar isoformas de genomas complejos, secuenciar directamente el ARN y caracterizar el ADN [19].

Las fuentes de datos empleadas en este trabajo son tres, primero las lecturas de la secuencia de ADN del individuo del cual se tienen que detectar las variaciones, segundo el Genoma de referencia con el que se tiene que comparar y finalmente un archivo con las variaciones del individuo etiquetadas. Estas tres fuentes de datos son:

1. **Oxford Nanopore (ONT) Sequencing Data** Los datos del organismo del que se deben detectar las variaciones es la muestra de referencia estándar HG001 (o NA12878) secuenciada por la tecnología ONT:

   - HG001 Circulomics (NA12878), GRCh38_no_alt, 58.06-fold:
     `https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=NHGRI_UCSC_panel/HG001/nanopore/Guppy_4.2.2/HG001_Circulomics_Guppy_4.2.2.fastq.gz`

2. **Genoma de referencia** La secuencia estándar para detectar si el individuo tiene variantes en este trabajo es:

   - GRCh38_no_alt:
     `https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz`

3. **Genome In A Bottle (GIAB) Truth Variants** El conjunto de datos GIAB proporciona para la muestra de referencia estándar HG001 (o NA12878) los SNPs e Indels de alta confianza que han sido etiquetados mediante diferentes tecnologías de secuenciación y genotipado, distintos alineadores y algoritmos de detección de variantes. [37]

   En este trabajo concretamente se utiliza la siguiente muestra de variaciones etiquetadas:

   - HG001 (NA12878), GRCh38, v3.3.2:
     `https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.2/GRCh38/`

1.4 Objetivos del trabajo

Este Trabajo Final de Máster comprende diferentes aspectos relativos a la Genómica y en específico a la Medicina de Precisión abarcando el problema de la detección de variantes. Por lo que en este apartado se presentan los objetivos.

El primer objetivo para abarcar el problema de la detección de variantes utilizando *Deep Learning*, es el estudio del Estado del Arte para dicha tarea. Conocer el enfoque tradicional que se ha estado utilizando a lo largo de los años y las nuevas propuestas

que están surgiendo. Este objetivo trata de enmarcar y entender por qué el uso de *Deep Learning* es una buena alternativa para la detección de variantes.

El segundo objetivo que se propone es el tratamiento y manipulación de datos genómicos. Conocer cómo se almacenan los datos secuenciados de ADN, formatos que se utilizan y herramientas necesarias para manipularlos.

El tercer objetivo se centra en la propuesta de arquitecturas neuronales para resolver la tarea. A partir del estudio del Estado del Arte y del análisis de tratamiento y manipulación de los datos genómicos, es interesante proponer topologías neuronales que puedan ser interesantes y competentes para la detección de variantes.

Como último objetivo, se propone una comparativa o *benchmarking* entre las redes neuronales diseñadas en el tercer objetivo y el Estado del Arte en la actualidad. Con esta comparativa se desea abrir la correspondiente discusión y proponer desarrollos futuros para posteriores investigaciones.

1.5 Estructura del trabajo

La estructura del trabajo se enmarca con el fin de cumplir con los objetivos propuestos en el punto anterior y establecer los siguientes pasos para estudios posteriores.

Por lo que primero se cubre el Estado del Arte, estudiando las diferentes propuestas tradicionales y profundizando en las más recientes relativas al *Deep Learning*.

Establecer la metodología de investigación es el siguiente paso, donde se expone y detalla el tratamiento de los datos genómicos y su correspondientes extracción de características. Además, en este capítulo se proponen diferentes arquitecturas de Redes Neuronales Artificiales para abarcar la detección de variantes y se establece el diseño experimental.

El capítulo de Experimentos y Resultados, presenta los estudios realizados y las métricas obtenidas en estos, para finalmente, en el capítulo de Conclusión y Trabajos Futuros discutir los resultados y expresar las líneas de investigación que se abren con este Trabajo Final de Máster.

CAPÍTULO 2
Estado del Arte

En este segundo capítulo se presenta el Estado del Arte desde dos perspectivas, la primera desde el enfoque tradicional y técnicas se han ido utilizando para la detección de variantes genómicas y el segundo enfoque trata la aplicación del *Deep Learning* y cómo este ha conseguido sobrepasar los resultados de las técnicas tradicionales.

2.1 Enfoques tradicionales

Las técnicas para la detección de variantes genómicas se enfrentan a diferentes problemas: mutaciones aleatorias, errores de homopolímero, errores de alineamiento, inserciones y deleciones (Indels), etc. Por lo que es necesario sistemas capaces de obtener buenas prestaciones ante diferentes situaciones. Tal y como se describe en [21] existen dos principales tipos de paradigmas para la detección de variantes: **Enfoque basado en el Ensamblaje** y **Enfoque Bayesiano**. [20]

El **Enfoque basado en el Ensamblaje** tiene como objetivo construir un genoma a partir de un gran número de lecturas secuenciadas de ADN sin ningún conocimiento a priori de la secuencia correcta ni del orden de los fragmentos de ADN. Existen gran cantidad de algoritmos [49] y programas (ABySS [22], DNASTAR [23], Newbler [24]) desarrollados con este enfoque en la actualidad.

El proceso que sigue el Enfoque basado en el Ensamblaje es: primero realiza un ensamblaje de *novo* utilizando ventanas fijas de lecturas secuenciadas, con el fin de construir haplotipos candidatos. Un haplotipo es una combinación de distintas formas alternativas de genes. El siguiente paso es calcular sus probabilidades con la secuencia de referencia. En cada ventana, el haplotipo con mayor probabilidad se considera la secuencia verdadera y de ahí deben de identificarse las variantes. [20]

Por otro lado, el **Enfoque Bayesiano** mapea las lecturas secuenciadas sobre el Genoma de referencia utilizando algoritmos de alineamiento para posteriormente identificar variantes candidatas. Las técnicas básicas de este enfoque detectan variantes a partir del número de Bases de las lecturas que tienen alta confianza que son diferentes con las Bases del Genoma de referencia. Mientras que los métodos más avanzados utilizan técnicas de Machine Learning que factorizan diferentes parámetros y estudian sus relaciones entre lecturas y Genoma de referencia para identificar variantes.

El proceso que siguen los Enfoques Bayesianos es: primero se mapean las lecturas secuenciadas con el Genoma de referencia para generar variantes candidatas. Y una vez realizado el alineamiento, se modelan los errores de secuenciación y se identifican variantes aplicando técnicas estadísticas o de *Machine Learning*. [20]

En [48] se comparan los dos enfoques y se abre una discusión detallada acerca de las ventajas y desventajas de cada técnica. Pero en resumidas cuentas, el Enfoque Bayesiano es muy potente para detectar variantes de un sólo Nucleótido (SNV), pero cuando se alinean las lecturas secuenciadas a regiones de Indels candidatos pueden confundirse. Mientras que el Enfoque basado en el Ensamblaje soluciona el problema de alineaciones incorrectas alrededor de Indels, por lo que mejora la precisión en comparación con el Enfoque Bayesiano. Pero la gran desventaja del Ensamblaje es su altísima complejidad computacional y el enorme número de haplotipos candidatos.

En la actualidad, los programas más extendidos y utilizados que aplican el Enfoque Bayesiano son: Samtools [9] y GATK-UnifiedGenotyper [10]. Para el Enfoque basado en el Ensamblaje se suele utilizar GATK-HaplotypeCaller [10]. Además, cabe destacar, que están surgiendo tecnologías que combinan los dos enfoques tradicionales, considerados **métodos bayesianos de haplotipos** como FreeBayes [50], PyroHMMvar [25] y Platypus [26]. [20]

2.2 Deep Learning

Tradicionalmente, el enfoque más extendido, asequible y eficaz ha sido el Enfoque Bayesiano. Este se basa en el uso de técnicas probabilísticas para la modelización de errores de secuenciación y de alineamiento de las lecturas. Por lo que la aplicación de *Deep Learning* a la detección de variantes es una evolución que se tenía que dar debido a la extensa amplitud y aplicación de los métodos neuronales a diferentes áreas de conocimiento. Este apartado trata los métodos de detección de variantes que mejores resultados obtienen en la actualidad.

2.2.1. DeepVariant

Google es la primera en utilizar *Deep Learning* en la tarea de detección de variantes. Su propuesta se denomina DeepVariant [14]. Principalmente DeepVariant está diseñado para aplicarse a las *Next Generation Sequencing* (NGS), que son las tecnologías de secuenciación que generan lecturas cortas con errores de entre el 0,1 y 10 %. Dependiendo de la tecnología de secuenciación utilizada entran en juego características que afectan a las tasas de error de secuenciación.

El Estado del Arte para la tarea de detección de variantes antes del nacimiento de DeepVariant estaba compuesto por una gran variedad de técnicas estadísticas cuyo objetivo era modelar los errores de secuenciación y alineamiento. Este es el caso de la herramienta más extendida, GATK [10]. GATK usa *Logistic Regression* para modelar los errores de Nucleótidos, un *Hidden Markov Model* para calcular las verosimilitudes entre las lecturas secuenciadas, *Naive Bayes* para identificar variantes y *Gaussian Mixtures Models* con conocimiento experto para eliminar falsos positivos. En concreto GATK obtiene métricas de rendimiento altas pero con errores significativos para la tecnología Illumina. Además, extender estos modelos a otras tecnologías de secuenciación es muy difícil debido a la necesidad de ajustes por parte de los expertos de las técnicas estadísticas. Esto es un gran problema en un área donde las tecnologías evolucionan tan rápidamente. [14]

Originalmente, DeepVariant se propone como una tecnología de detección de variantes para NGS que reemplaza todos los componentes y conocimiento experto de las técnicas estadísticas por un único modelo de *Deep Learning*. Esta herramienta propuesta por Google resuelve la tarea en dos pasos: primero busca variantes candidatas y codifica la información de dicha región en una imagen, denominada Pileup. En segundo paso, se

encarga de introducir dicha imagen en un modelo neuronal llamado InceptionV2 [27] que detecta si hay variante o no, este proceso se presenta en la Figura 2.1 y 2.2. [14]

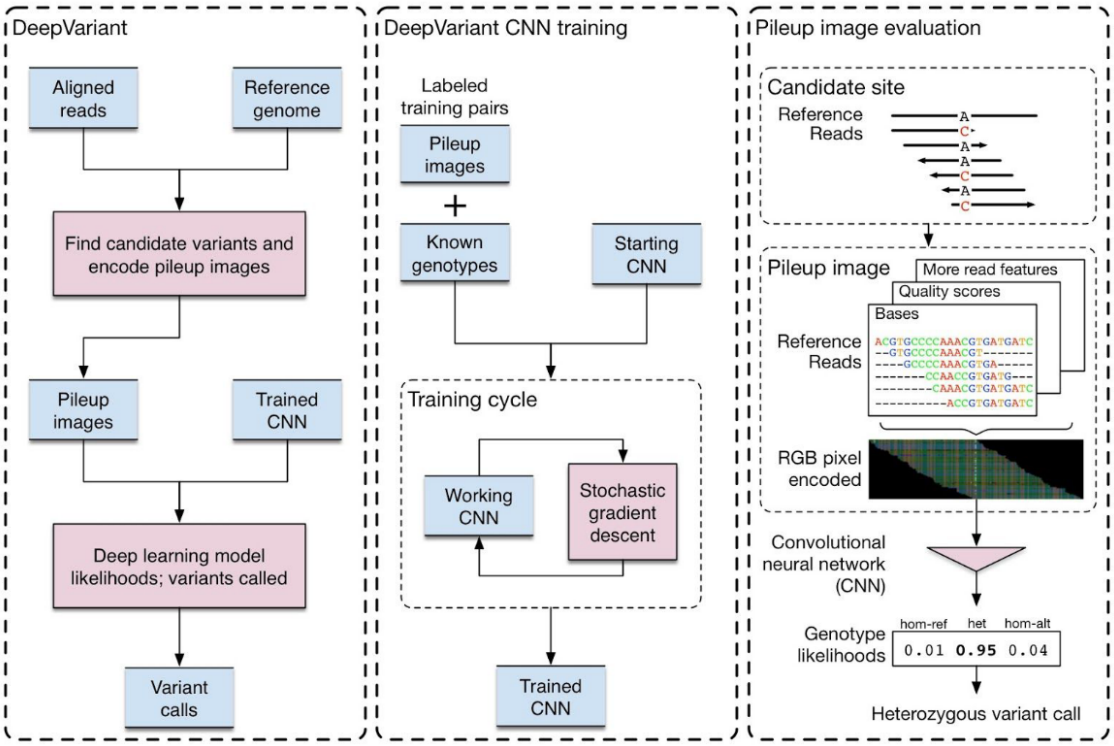

Figura 2.1: Flujo de trabajo de DeepVariant. [14]

Figura 2.2: Flujo de trabajo de DeepVariant simplificado. [30]

La propuesta de Google se basa en transformar el problema de detección de variantes, en un problema de *Computer Vision*. En el GitHub oficial de DeepVariant [30] se muestra el flujo de trabajo simplificado de esta tecnología (Figura 2.2). Las entradas que recibe es el genoma del individuo que se desea analizar (BAM o CRAM) y el Genoma de referencia (FASTA). Después ejecuta "*make_examples*" que se encarga de buscar variantes candidatas con heurísticas simples y codificar la información en imágenes de 6 canales diferentes, tal y como se muestra en la Figura 2.3 obtenida de [28] llamada Pileup. Los Pileups calculados se pasan a un modelo neuronal InceptionV4 [29] y las probabilidades

calculadas por el modelo se postprocesan y se obtiene un fichero de salida VCF o gVCF con las variantes detectadas del individuo.

Figura 2.3: Pileup de variante candidata. [28]

Con este flujo de trabajo DeepVariant es capaz de superar todos los métodos tradicionales en gran variedad de entornos experimentales, tanto para la detección de sustituciones (SNPs) como de inserciones y deleciones (Indels). Esta tecnología demuestra que un algoritmo de reconocimiento de patrones de propósito general sin conocimiento experto como es una Red Neuronal Artificial, es capaz de obtener mejores resultados y más robustez a los errores de secuenciación y alineamiento que las técnicas estadísticas tradicionales que se centran en el conocimiento experto. [31]

2.2.2. P.E.P.P.E.R.

DeepVariant es la primera tecnología *Deep Learning* en aplicarse al problema de detección de variantes en *Next Generation Sequencing* (NGS) y obtiene el mejor rendimiento de todo el estado del arte. Pero si se aplica DeepVariant directamente a genomas secuenciados por *Third Generation Sequencing* (TGS), que produce lecturas mucho más largas que las NGS pero con mayores tasas de error, se obtienen demasiadas variantes candidatas, lo que limita la aplicación de DeepVariant a las TGS. Debido a este problema, nace PEPPER-Margin-DeepVariant [32], un flujo de trabajo que produce resultados del estado del arte para TGS empleando *Deep Learning*.

Como se ha comentado en el capítulo de Introducción, las TGS tienen un gran potencial para llevar la detección de variantes a otro nivel, ya que gracias a las largas lecturas que producen se pueden mapear de mejor forma regiones difíciles del ADN. El problema de las TGS es su alta tasa de error de secuenciación de entre el 5 y 15 %.

PEPPER-Margin-DeepVariant es un flujo de trabajo propuesto en [32] para la detección de variantes en TGS. En la publicación se demuestra que el uso de TGS con la tecnología PEPPER-Margin-DeepVariant supera a los métodos de NGS en la identificación de SNPs e Indels, además, en zonas difíciles o con pocas lecturas secuenciadas es capaz de obtener resultados de alta calidad. [32]

El flujo de trabajo de PEPPER-Margin-DeepVariant se expone en la Figura 2.4 y se puede dividir en 3 etapas: como entrada se le pasa el genoma secuenciado de un individuo (en formato BAM) y el Genoma de referencia (en formato FASTA) y comienza el primer módulo llamado PEPPER. PEPPER busca variantes candidatas y las codifica en imágenes Pileups, tal y como hace DeepVariant, solo que el Pileup que genera no es tan detallado y únicamente genera una imágen de 1 canal (Figura 2.5). Los Pileups se introducen a una Red Neuronal Recurrente y esta calcula si existe o no variante y dichas probabilidades se pasan al siguiente módulo. El segundo módulo es Margin Haplotag, que utilizan *Hidden Markov Models* para asignar haplotipos a las variantes y lecturas de baja calidad. Finalmente, el módulo DeepVariant, vuelve a calcular los candidatos propuestos por PEPPER usando las lecturas de Margin. Cabe destacar que el Pileup que genera DeepVariant no es el mismo que PEPPER. Ya que la imagen que se crea tiene en cuenta toda la información de los alineamiento de la región candidata, en la Figura 2.6 se presenta un ejemplo que corresponde con una imágen de 7 canales bastante parecida a la expuesta en la Figura 2.3.

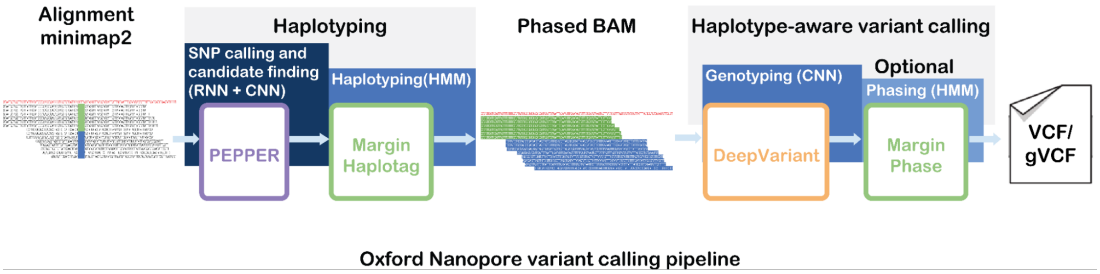

Figura 2.4: Flujo de Trabajo de PEPPER-Margin-DeepVariant. [33]

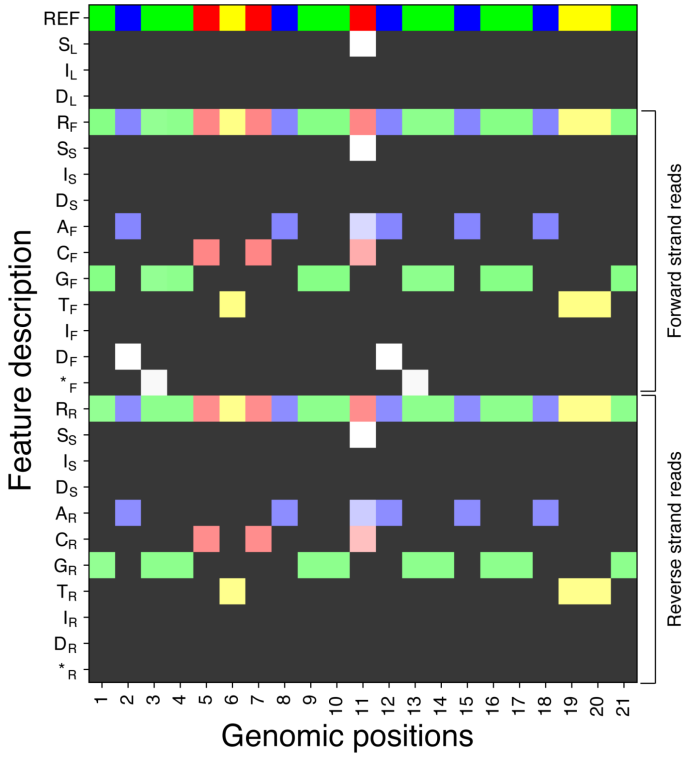

Figura 2.5: Pileup generado por PEPPER. Cabe destacar que la imagen muestra colores para que se entienda mejor lo que se representa, pero realmente la imagen únicamente contiene un canal. [33]

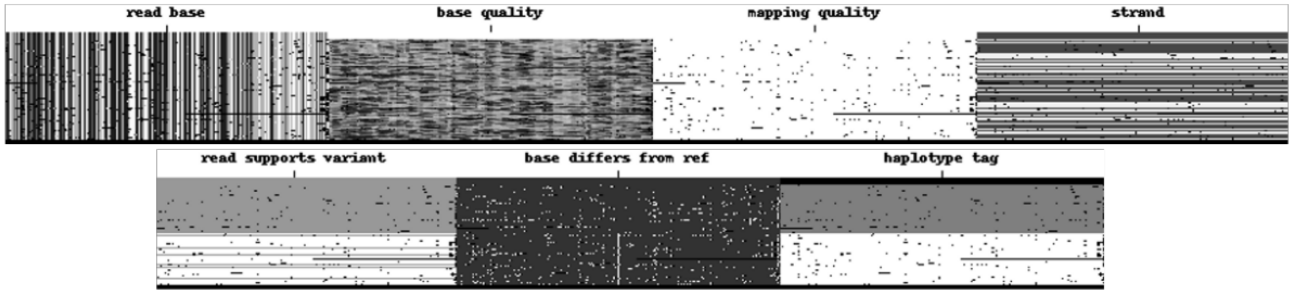

Figura 2.6: Pileup generado por el módulo DeepVariant de PEPPER-Margin-DeepVariant. [33]

2.2.3. Clair3

Clair3 [34] es la tecnología que junto a PEPPER-Margin-DeepVariant lideran la detección de variantes para *Third Generation Sequencing* (TGS). Clair3 la tercera generación de métodos neuronales para la tarea de detección de variantes propuestos por The University of Hong Kong, sucesor de Clair [35], el cual es descendiente de Clairvoyante [36].

Desde la publicación de DeepVariant, gran cantidad de métodos basados en *Deep Learning* se han aplicado a la detección de variantes, superando a las técnicas estadísticas tradicionales. En la actualidad, la tecnología que mejor funciona para la detección de variantes de lecturas cortas (NGS) es DeepVariant. Debido a esto, en las TGS el dominio de los métodos neuronales han sido desde el comienzo los mejores, debido a que la dificultad de las lecturas largas y el alto error de secuenciación que estas conllevan. [34]

Teniendo en cuenta que el problema de detección de variantes tiene similitudes entre NGS y TGS en [34] se expone el flujo de trabajo que compone Clair3. Este consta de dos enfoques; el uso de Pileups de variantes candidatas por un lado y de información completa de los alineamientos por otro. El uso de imágenes Pileups tiene la particularidad de ser más eficiente temporalmente, mientras que el uso de la información completa de los alineamientos proporciona mayor *precision* y *recall*, pero es más costoso computacionalmente. Por lo que en la publicación, deciden que los diseños no son exclusivos y no se han realizado estudios combinando estos dos enfoques, de ahí nace Clair3. [34]

El flujo de trabajo de Clair3 se presenta en la Figura 2.7. La filosofía del diseño se resume en dos redes neuronales, una Red Neuronal Recurrente (RNN) (Figura 2.8 a) que tiene como entrada los Pileups con resúmenes de la variante candidata y la segunda una Red Neuronal Convolucional (CNN) (Figura 2.8 b) que su entrada en la información completa de alineamientos. Según la Figura 2.7, Clair3 confía en la detección de variantes de la CNN a no ser que la RNN pueda tomar una decisión rápida y fiable. Esto es debido a que las exigencias computacionales del manejo de la entrada de la CNN y sus datos son más altas que la RNN con sus Pileups. [34]

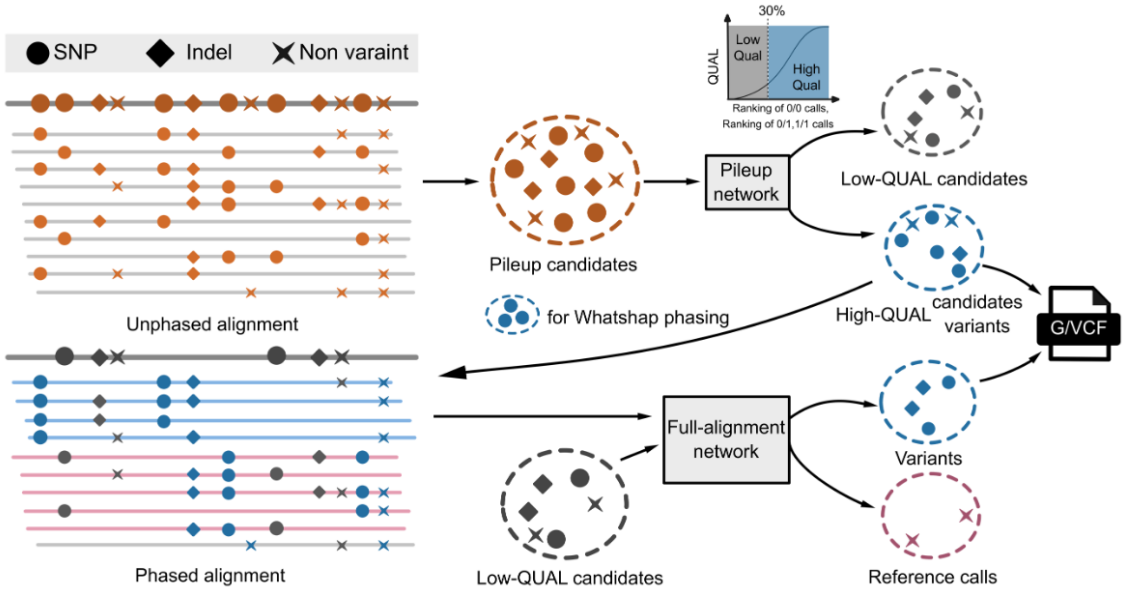

Figura 2.7: Flujo de trabajo de Clair3. [34]

2.3 Resumen General

Con el estudio del Estado del Arte para el problema de la detección de variantes genómicas se han repasado los dos principales enfoques tradicionales:

- **Enfoque basado en el Ensamblaje**: tiene como objetivo construir un genoma a partir de un gran número de lecturas secuenciadas de ADN sin ningún conocimiento a priori de la secuencia correcta ni del orden de los fragmentos de ADN.

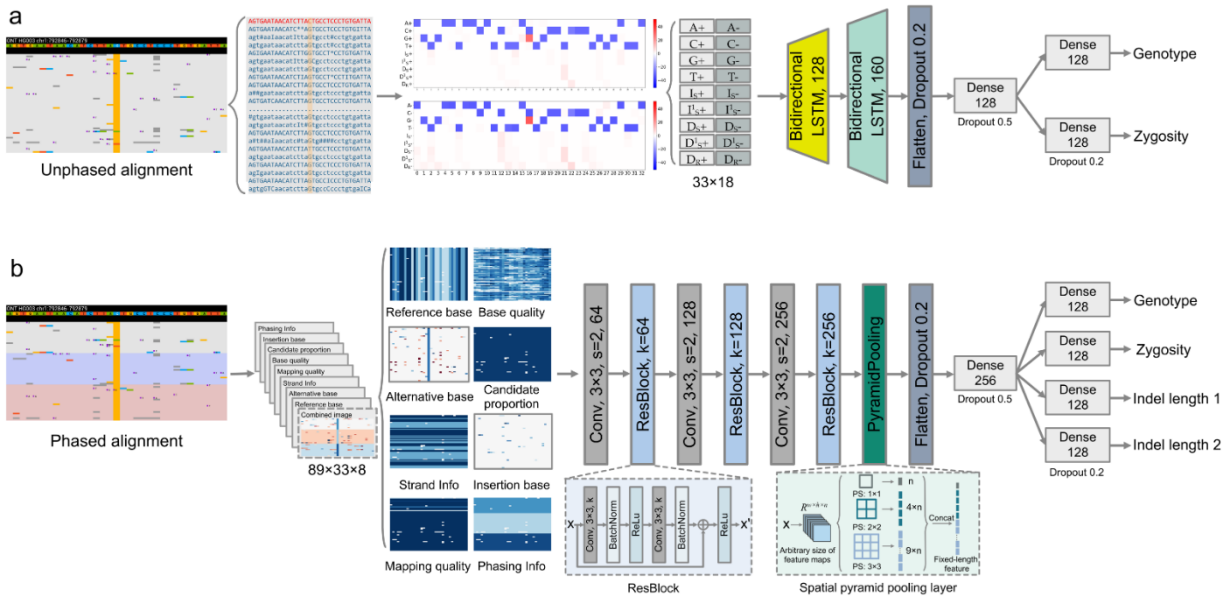

Figura 2.8: Arquitecturas neuronales de Clair3. (a) Corresponde al modelo de Pileups. (b) Corresponde al modelo de información completa de alineamientos. [34]

- **Enfoque Bayesiano**: mapea las lecturas secuenciadas sobre el Genoma de referencia utilizando algoritmos de alineamiento para posteriormente identificar variantes candidatas.

El **Enfoque Bayesiano** es el más extendido debido a su potencia para detectar variaciones de un sólo Nucleótido (SNV), pero este empeora cuando las lecturas secuenciadas se alinean a regiones de Indels. Por otro lado, el **Enfoque basado en el Ensamblaje** se comporta mejor con las alineaciones incorrectas alrededor de Indels, pero tiene un altísimo coste computacional.

En 2018, con la aparición de **DeepVariant** de la mano de Google, el paradigma cambia y el nuevo método basado en *Deep Learning* supera a las herramientas más extendidas como GATK [10] en las NGS, es decir, lecturas cortas.

Esto da lugar a que con la aparición de las TGS, lecturas largas, el paradigma más exitoso desde el principio sea el uso de modelos neuronales. Donde en la actualidad **PEPPER-Margin-DeepVariant** y **Clair3** son las tecnologías que mejores resultados consiguen en la detección de variantes.

Tal y como se ha comentado en el capítulo Introducción, en este trabajo se ha decidido utilizar los datos de Oxford Nanopore Technologies (ONT). Los datos de ONT pertenecen a las TGS, por lo que este Estado del Arte cubre todo lo necesario para desarrollar la Metodología y Experimentación necesarias para el estudio y análisis de métodos neuronales para la detección de variantes genómicas.

CAPÍTULO 3
Metodología

El tercer capítulo presenta la Metodología, donde se describe de forma detallada el procedimiento llevado a cabo para realizar el presente trabajo.

3.1 Preparación de los datos

En este apartado se listan de nuevo las fuentes de datos utilizadas y se explica su correspondiente preprocesado, ya que estos datos no están listos para su uso.

Las tres fuentes de datos comentadas anteriormente son:

1. **Lecturas secuenciadas de un individuo Oxford Nanopore (ONT) Sequencing Data** Los datos del organismo del que se deben detectar las variaciones es la muestra de referencia estándar HG001 (o NA12878) secuenciada por la tecnología ONT:

   - HG001 Circulomics (NA12878), GRCh38_no_alt, 58.06-fold:
     `https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=NHGRI_UCSC_panel/HG001/nanopore/Guppy_4.2.2/HG001_Circulomics_Guppy_4.2.2.fastq.gz`

2. **Genoma de referencia** La secuencia estándar utilizada en este trabajo es:

   - GRCh38_no_alt:
     `https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz`

3. **Variaciones etiquetadas del individuo Genome In A Bottle (GIAB) Truth Variants** El conjunto de datos GIAB proporciona para la muestra de referencia estándar HG001 (o NA12878) los SNPs e Indels de alta confianza que han sido etiquetados mediante diferentes tecnologías de secuenciación y genotipado, distintos alineadores y algoritmos de detección de variantes. [37]

   En este trabajo concretamente se utiliza la siguiente muestra de variaciones etiquetadas:

   - HG001 (NA12878), GRCh38, v3.3.2:
     `https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.2/GRCh38/`

Los datos HG001 del individuo secuenciados con la tecnología ONT, se tienen que alinear con el Genoma de referencia. En este trabajo, el algoritmo de alineamiento utilizado es Minimap2 (v2.17-r941) [20]. Minimap2 es un software de alineamiento de propósito general, diseñado para mapear secuencias largas en genomas de referencia. Esta herramienta realiza la alineación de las lecturas divididas y utiliza diferentes heurísticas para llevar a cabo el mapeo. Se ha decidido utilizar esta tecnología debido a que es un mínimo de 4 veces más rápido que el resto de alineadores y porque está diseñado para soportar grandes bases de datos y lecturas largas.

Después de alinear los datos utilizando Minimap2, se utilizan las herramientas Samtools (v1.10) [38] para la manipulación de datos genómicos. Con Samtools se realiza la conversión de los datos secuenciados del individuo al formato BAM, que es un formato para representar lecturas alineadas. Se ordena el archivo BAM utilizando Samtools y se obtiene su correspondiente archivo de índices, que permite el acceso rápido a diferentes regiones de las lecturas alineadas. Con estos pasos, se tienen las fuentes de datos listas para su uso. La forma de referirse a cada uno de los tres archivos de ahora en adelante es:

- BAM: archivo que contiene las lecturas del individuo alineadas.
- FASTA: genoma de referencia.
- VCF: archivo que contiene todas las variantes etiquetadas.

3.2 Extracción de características

En el capítulo correspondiente al Estado del Arte, se ha estudiado que los métodos que utilizan *Deep Learning* para la detección de variantes transforman el problema clásico en una tarea de *Computer Vision*. Primero detectan con unas heurísticas simples variantes candidatas en el archivo BAM utilizando el archivo FASTA como Genoma de referencia. El siguiente paso convierte en una imagen resumen denominada Pileup la información de la región seleccionada, tal y como se presenta en la Figura 2.5.

Los dos métodos que conforman el Estado del Arte para las *Third Generation Sequencing* (TGS), o lecturas largas, son P.E.P.P.E.R.-Margin-DeepVariant y Clair3. Los Pileups o imágenes que generan estas tecnologías son casi idénticos, el Pileup de P.E.P.P.E.R. se presenta en la Figura 2.5 y el de Clair3 en la Figura 3.1. La principal diferencia es que los Pileups de P.E.P.P.E.R. se pueden describir como matrices de 26 filas y 21 columnas, donde en las 4 primeras filas se resume la información de referencia y se marca si ha ocurrido un SNP, una Inserción o una Deleción. Mientras que de la fila 5 a la 15 se resume la información en una lectura del genoma hacia delante y de la fila 16 al final se realiza la lectura hacia atrás. La representación en Clair3 es casi idéntica, pero tratándose de matrices de 18 filas y 33 columnas. En la Figura 3.1, la parte de arriba corresponde a la lectura hacia delante y la de abajo a la lectura hacia atrás. Cabe destacar, que Clair3 utiliza ventanas más grandes que P.E.P.P.E.R., pero en la última versión de P.E.P.P.E.R. r0.8 se ha actualizado la representación de Pileups a matrices de 26 filas y 33 columnas.

En [39] se explica por qué se realiza una lectura hacia delante y otra hacia atrás en la secuenciación por pares. Las tecnologías de secuenciación realizan las lecturas de ADN en una dirección de 5′ a 3′, esto es debido a que así es como funcionan los mecanismos que se encargan de replicar el ADN de los organismos. Por eso la secuenciación de las cadenas se sintetizan en una dirección de 5′ a 3′. Por lo que se termina con ambas cadenas de ADN y para leerlas, se utilizan adaptadores no complementarios a los extremos 3′ y 5′ de las lecturas de ADN, tal y como se muestra en la Figura 3.2. [39]

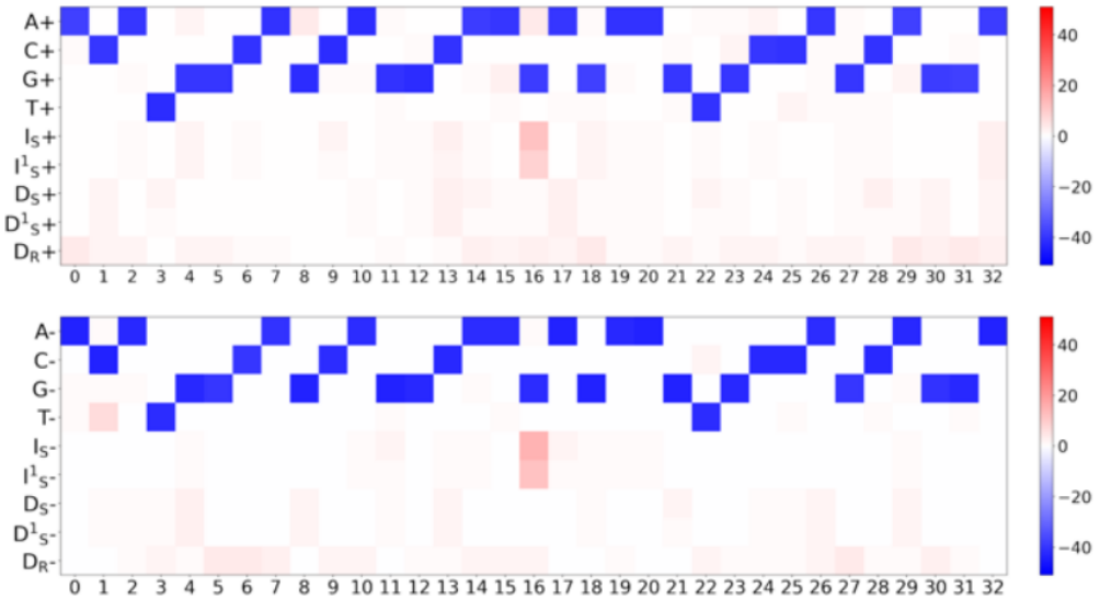

Figura 3.1: Pileup de Clair3. [34]

Como se muestra en la Figura 3.2, en la secuenciación convencional, denominada por pares, se secuencia utilizando el adaptador para un extremo. Y cuando se ha terminado, se vuelve a secuenciar utilizando el adaptador por el otro extremo. Por lo tanto, una lectura hacia delante es en la dirección 5′ a 3′ y hacia atrás es de 3′ a 5′. [39]

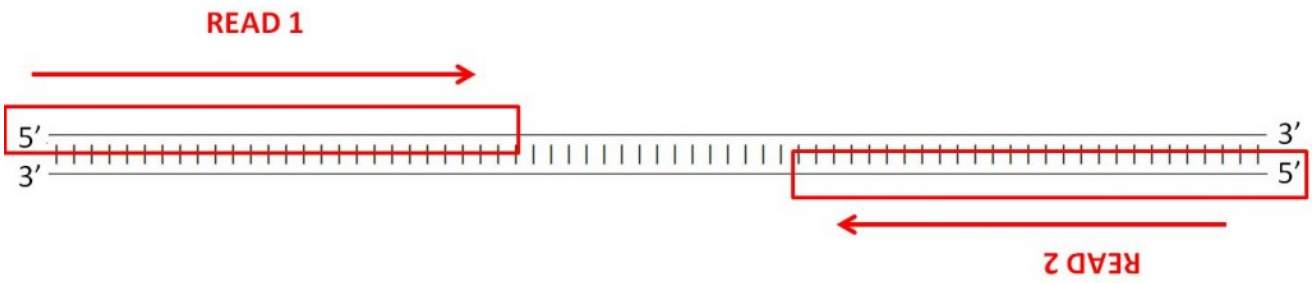

Figura 3.2: Lectura hacia delante y hacia atrás en la secuenciación por pares. [39]

Debido a que las representación que se construye en las dos tecnologías del Estado del Arte son casi idénticas, se ha decidido estudiar en profundidad el Pileup de P.E.P.P.E.R.-Margin-DeepVariant, desarrollarla y emplearla.

Como se ha expuesto anteriormente, un Pileup es una imagen o matriz que resume toda la información de una región de interés del genoma del individuo que se está estudiando. Dependiendo de cómo de rico y detallado sea este Pileup este será más discriminante para la detección de variantes.

En el caso del Pileup de P.E.P.P.E.R., se identifican las variantes candidatas y se generan las imágenes de estas. Para una variante candidata se utilizan todas sus lecturas y se calcula la frecuencia de cada Nucleótido o base para una determinada región en comparación con el Nucleótido de referencia. A partir de esta información se genera el resumen del alineamiento leído (Pileup).

En el primer bloque del Pileup, la primera fila corresponde a la secuencia o Genoma de referencia, donde cada uno de los 4 Nucleótidos tiene un valor del 1 al 4. Para las filas 2, 3 y 4 denominadas S, I y D se codifica la longitud de la variante SNP, Inserción y/o Deleción, caso de no existir variación se utiliza el valor 0.

En el segundo bloque del Pileup, se tiene la información de la lectura hacia delante (dirección 5′ a 3′), donde la primera fila cuenta cuantas lecturas tienen el mismo alelo (forma alternativa de un gen) que el de referencia. Las filas 2, 3 y 4 denominadas S, I y D cuentan cuántas lecturas tienen el mismo alelo que el de referencia basándose en la variante candidata SNP, Inserción y/o Deleción. Las filas 5, 6, 7, 8, denominadas A, C, G, T cuentan el número de veces que se expresa cada Nucleótido en la región que se está estudiando. Finalmente, las filas 9, 10 y 11, denominadas I, D y * cuentan el número de Inserciones/Deleciones en cada posición.

Para el tercer bloque del Pileup, se contiene la información de la lectura hacia atrás. Los cálculos a realizar son los mismos que en la lectura hacia delante pero en una dirección 3′ a 5′.

En resumen, a partir del archivo BAM que contiene las lecturas de la secuencia del individuo y del archivo FASTA que contiene el Genoma de referencia, se buscan variantes candidatas y se genera un Pileup de cada una de ellas siguiendo la estructura de P.E.P.P.E.R. de la Figura 2.5.

3.3 Arquitecturas Neuronales

La idea de utilizar *Deep Learning* para la tarea de detección de variantes es el enfoque que mejores resultados obtiene en la actualidad, debido a esto es interesante proponer diferentes arquitecturas que se adapten al análisis de los Pileups de las variantes candidatas.

La naturaleza del problema de detección de variantes consiste en identificar si ocurre una variación respecto al Genoma de referencia o no, y en caso de ocurrir clasificar en variante homocigota o heterocigota. Por lo que para modelizar esta tarea, se tienen tres clases:

- **Homocigoto Referencia**: las lecturas secuenciadas coinciden con la de referencia.
- **Homocigoto Alterado**: las lecturas secuenciadas proponen un Nucleótido diferente al de la secuencia de referencia.
- **Heterocigoto Alterado**: las lecturas secuenciadas proponen más de un Nucleótido diferente al de la secuencia de referencia.

En la Figura 3.3 se presenta un ejemplo de [40], donde en la imagen de la izquierda, la cadena de abajo es la secuencia de referencia y el resto son lecturas secuenciadas de un organismo. Se puede apreciar, que para las columnas o regiones de ADN 1, 3, 5, 6, 7 y 9 las lecturas proponen el mismo Nucleótido que el de referencia, por lo que estas se clasifican como Homocigoto Referencia, es decir, no hay variación. En la región 8 se puede apreciar que se proponen tres Ts y el resto son Cs que coinciden con la referencia, por lo que al ser baja la frecuencia de las Ts se considera que no hay variación. Se puede apreciar un caso de Homocigoto Alterado en la región 4, donde todas las lecturas proponen Gs y la secuencia de referencia propone A. Además, también hay un caso de Heterocigoto Alterado en la región 2, donde se proponen Ts y As y la referencia es T. Este ejemplo simplificado del problema permite explicar y entender las tres clases posibles en la detección de variantes.

En este apartado de la Metodología se presentan las diferentes arquitecturas de Redes Neuronales Artificiales propuestas en este estudio para resolver la tarea de detección de variantes. Todos los modelos neuronales han sido desarrollados utilizando la librería de Python v.3.9 TensorFlow v.2.7.0 con CudaToolkit v.11.6.0 y CudNN v.8.2.1.32.

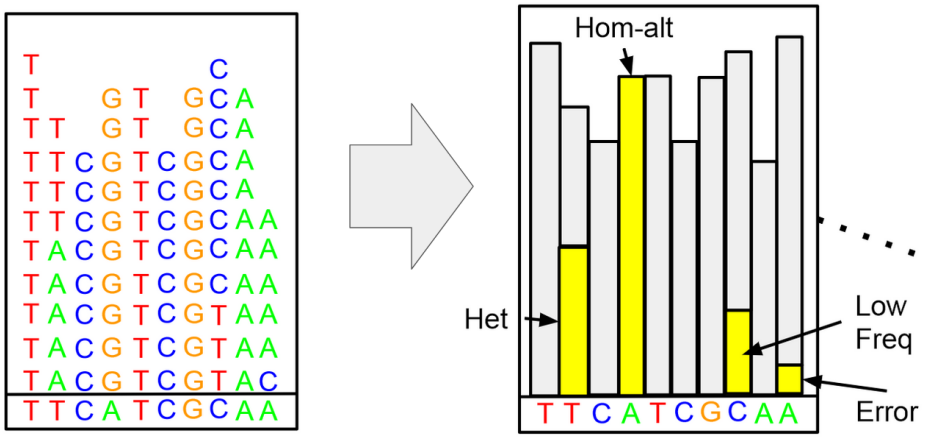

Figura 3.3: Ejemplo simple de estudio de variantes. [40]

La primera arquitectura a estudiar es el modelo por defecto presentado en el artículo de PEPPER-Margin-DeepVariant [32] (Figura 3.4), que es una Red Neuronal Recurrente (RNN) compuesta por dos capas de neuronas *Gated Recurrent Unit* (GRU) propuestas en [41]. El uso de neuronas recurrentes es debido a que los Pileups que se utilizan de entrada a la red son de dos dimensiones, por lo que la red procesa los datos columna a columna.

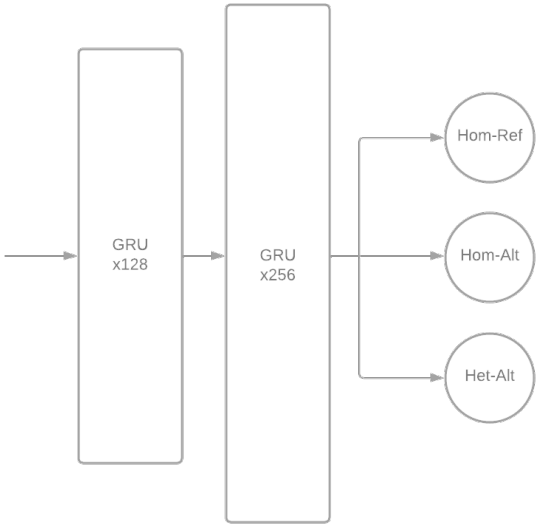

Figura 3.4: RNN de P.E.P.P.E.R. por defecto.

La segunda propuesta a estudiar, es una arquitectura de Perceptrón Multicapa (MLP) sencilla (Figura 3.5), donde se realiza un *Flatten* de los datos bidimensionales para convertirlos en un vector de dimensiones igual a: número de filas por número de columnas. Seguido al *Flatten* se aplican tres capas ocultas para la posterior clasificación. Aplicar una arquitectura simple como el MLP de la Figura 3.5 resulta interesante para analizar las prestaciones que aporta, ya que este es más sencillo computacionalmente que una RNN.

Cabe destacar que las capas ocultas del MLP propuesto son bloques con *Batch Normalization*, *Gaussian Noise*, función de activación y *Dropout*, tal y como se muestra en la Figura 3.6.

La tercera propuesta es una Red Neuronal Recurrente compleja que incorpora dos capas bidireccionales de neuronas *Long Short-Term Memory* (LSTM) presentadas en [42].

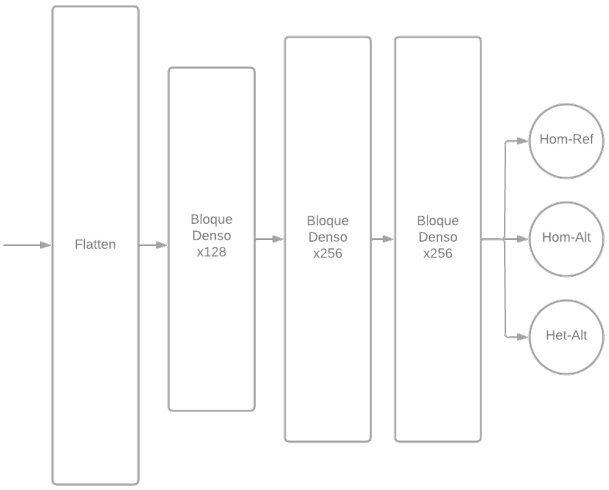

Figura 3.5: Perceptrón Multicapa (MLP).

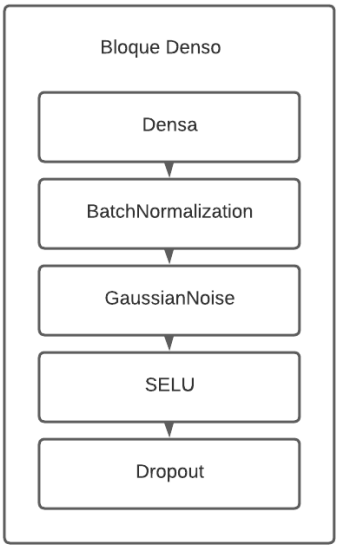

Figura 3.6: Bloque denso de la MLP.

A esta red se la denomina Bi-LSTM y ha sido diseñada pensando en la naturaleza de los datos Pileup (Figura 3.7). Es una arquitectura más compleja que la Red Recurrente de la Figura 3.4, ya que es más profunda y utiliza celdas LSTM, que son más ampliamente utilizadas en la actualidad.

Debido a que el enfoque de Google original es transformar el problema de detección de variantes a un problema de *Computer Vision* es interesante aplicar Redes Neuronales Convolucionales (CNN) con mecanismos de Self-Attention inspirados en el Transformer de [43] (Figura 3.8). Ya que las CNNs son el estándar en las tareas de imagen y vídeo. Para aplicar CNNs la imagen Pileup tiene que redimensionarse de una matriz a un Tensor con 1 canal.

Finalmente, la última arquitectura propuesta para esta tarea está inspirada en el campo del *Computer Vision* [44]. Este enfoque es clave para las tareas de categorización de

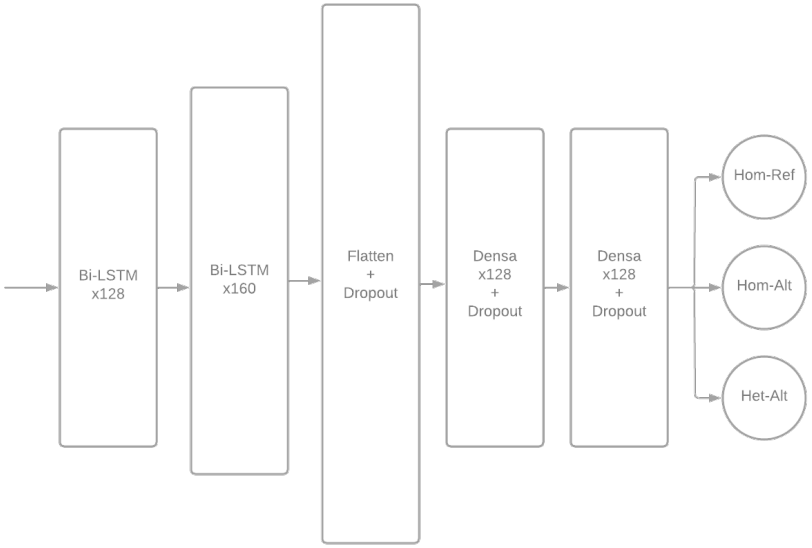

Figura 3.7: Bidirectional LSTM (Bi-LSTM).

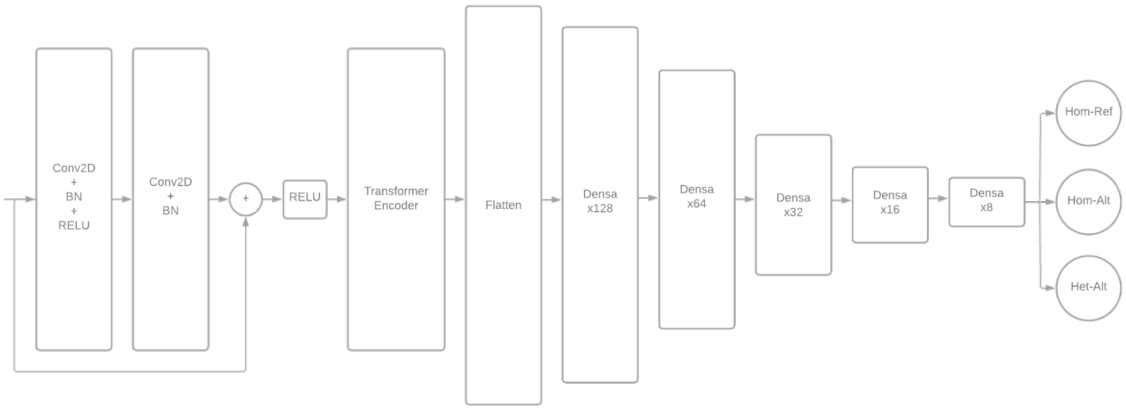

Figura 3.8: Red Neuronal Convolucional con Transformer (CNN+Transformer).

grano fino, donde las clases tienen diferencias sutiles. Se compone de dos extractores de características cuyas salidas se unen multiplicándose por el *outer product* o producto exterior (Figura 3.9).

En este trabajo se proponen diferentes arquitecturas neuronales para abarcar el mismo problema, desde topologías sencillas como el MLP (Figura 3.5) y RNN GRU (Figura 3.4) hasta topologías modernas y recientes como RNN bidireccionales (Figura 3.7) y CNN con mecanismos de Self-Attention (Figura 3.9).

3.4 Diseño Experimental

Utilizando las variantes etiquetadas en el archivo VCF de GIAB [37] y el archivo BAM de lecturas secuenciadas del organismo, se buscan variantes candidatas y se generan sus correspondientes Pileups. De esta forma se genera el conjunto de datos HG001 (o NA12878).

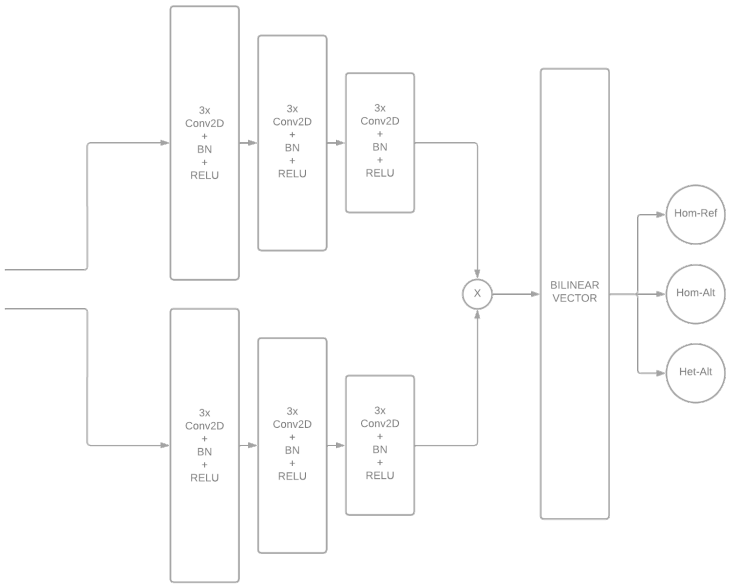

Figura 3.9: Red Neuronal Convolucional Bi-linear (Bi-linear CNN).

El diseño experimental que se realiza en este trabajo está basado en el que utiliza Google para DeepVariant [14], que consiste en dividir en entrenamiento (o *train*) y prueba (o *test*) el conjunto de datos en función de los Cromosomas.

Un Cromosoma es un paquete ordenado de ADN que se encuentra en el núcleo de la célula, este contiene la mayor parte de la información genética de un ser vivo. En la Figura 3.10 se puede observar la estructura de un Cromosoma y su composición. Los humanos tienen 23 pares de Cromosomas, donde 22 pares son autosómicos y el vigésimo tercer par es de Cromosomas sexuales X e Y.

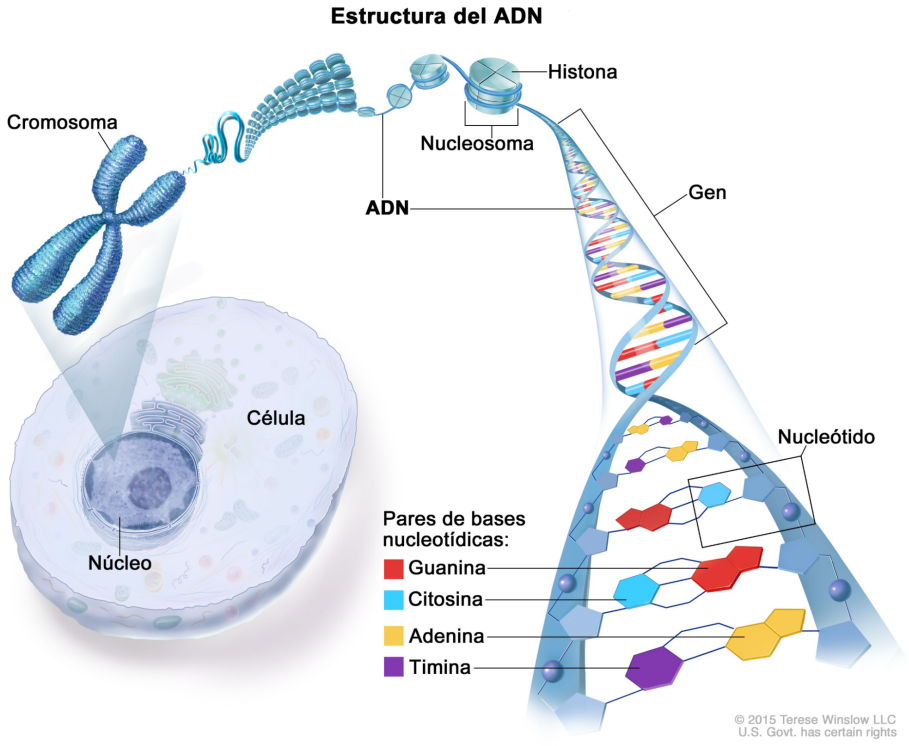

Figura 3.10: Representación del ADN, desde los Nucleotidos a la célula. [45]

La división que se realiza de *train* y *test* se presenta en la Figura 3.11. Donde el *train* contiene desde los Cromosomas número 1 al 19 y el *test* contiene del 20 al 22. Además, los Cromosomas sexuales se descartan del conjunto de datos y no se utilizan.

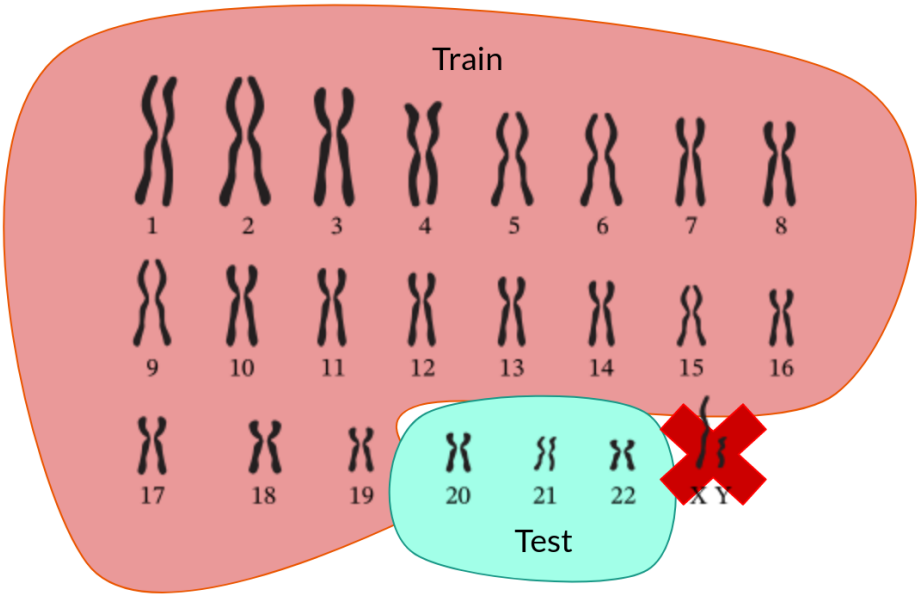

Figura 3.11: Diseño experimental de train y test.

Aplicando el diseño experimental de la Figura 3.11 al conjunto de datos que se proponen en este trabajo, se obtiene la Tabla 3.1. En dicha tabla se presentan el número de variantes candidatas para *train* y *test*. Se puede apreciar que el problema está desbalanceado, ya que la mayoría de *samples* corresponde a la clase Homocigoto Referencia, es decir, la clase que representa que no hay variación.

Longitud del Train			Longitud del Test		
15.531.138			2.463.331		
Hom-Ref	Het-Alt	Hom-Alt	Hom-Ref	Het-Alt	Hom-Alt
12.313.138	1.961.998	1.255.324	2.296.928	106.458	59.945

Tabla 3.1: Tamaño del conjunto de datos.

Para los problemas de clasificación existen tres principales medidas para estudiar los resultados del modelo: *Precision*, *Recall* y *F1-Score*. Si para un problema desbalanceado como el que se trata en este trabajo se utiliza la *Precision*, entonces se prioriza minimizar los Falsos Positivos (FP). En cambio, si se usa el *Recall* se prioriza minimizar los Falsos Negativos (FN). Pero en lugar de elegir entre *Precision* o *Recall*, es interesante combinarlas para obtener el *F1-Score* y de esta forma resumir el rendimiento del modelo.

Con el número de Verdaderos Positivos (TP) y Falsos Positivos (FP) se calcula la *Precision*:

$$precision = \frac{TP}{TP + FP}$$

El *Recall* se obtiene mediante los Falsos Negativos (FN):

$$precision = \frac{TP}{TP + FN}$$

A partir de la *Precision* y *Recall* se calcula el *F1-Score*:

$$F1Score = \frac{2 \times precision \times recall}{precision + recall}$$

En resumen, las distintas arquitecturas neuronales son entrenadas utilizando el diseño experimental propuesto (Figura 3.11) y sus predicciones se analizan y comparan con las del archivo VCF que contiene las verdaderas variantes. Finalmente, la comparativa entre modelos (*benchmarking*) y sus clasificaciones se realiza utilizando la herramienta Haplotype VCF comparison tools (v3.12.1) [46]. Esta herramienta calcula la *Precision*, *Recall*, *F1-Score* de los SNPs, Inserciones, Deleciones e Indels.

CAPÍTULO 4
Experimentos y Resultados

En este capítulo se realizan los experimentos siguiendo la Metodología descrita anteriormente. Se entrena cada una de las redes neuronales propuestas en las Figuras 3.4, 3.5, 3.7, 3.8 y 3.9 y se clasifican las variantes candidatas para su posterior comparativa con el VCF que contiene la variantes etiquetadas. Finalmente se presentan las comparativas de los modelos neuronales entrenados con los modelos del Estado del Arte.

4.1 GRU de P.E.P.P.E.R.

En el primer experimento se ha entrenado la arquitectura neuronal presentada en la Figura 3.4, que es la propuesta en el artículo [32] para P.E.P.P.E.R.. Los resultados obtenidos se presentan en la Tabla 4.1 donde se puede apreciar que para la detección de SNPs se alcanza un *F1-Score* del 98 %, mientras que para la identificación de Indels, Inserciones y Deleciones las métricas no alcanzan el 53 %. Pero, en general, como se dan más SNPs que otro tipo de variaciones, el *F1-Score* es de 92 %. Además se puede apreciar que para los SNPs el *Recall* es más alto que la *Precision*, mientras que para los Indels, Inserciones y Deleciones la *Precision* es más alta que el *Recall*. Cabe mencionar que los valores de *Precision* y *Recall* están bastante igualados.

GRU P,E,P,P,E,R,	Precision	Recall	F1-Score
Overall	**0,9171**	**0,9234**	**0,9203**
SNP	0,9743	0,9926	0,9834
Indel	0,5411	0,5058	0,5228
Ins	0,5384	0,4995	0,5183
Del	0,5436	0,5119	0,5273

Tabla 4.1: Resultados arquitectura GRU P.E.P.P.E.R..

4.2 Perceptrón Multicapa

Para la arquitectura propuesta en la Figura 3.5, el MLP en función de los resultados de la Tabla 4.2, obtiene un 98 % para la detección de SNPs, mientras que para la detección de Indels, Inserciones y Deleciones se queda sobre el 56 %. En general el *F1-Score* alcanzado es del 93 %. En este caso se tiene para las *Precision* de los Indels, Inserciones y Deleciones los valores son 10 puntos más altos que en los de *Recall*.

MLP	Precision	Recall	F1-Score
Overall	**0,9401**	**0,9208**	**0,9304**
SNP	0,9828	0,9898	0,9863
Indel	0,6299	0,5141	0,5661
Ins	0,6342	0,5159	0,5690
Del	0,6257	0,5123	0,5634

Tabla 4.2: Resultados arquitectura Perceptrón Multicapa.

4.3 Bidirectional LSTM

El tercer experimento es el entrenamiento y clasificación de la arquitectura *Bidirectional* LSTM de la Figura 3.7. Los resultados de esta red neuronal se muestran en la Tabla 4.3. Con este modelo se alcanza un *F1-Score* del 99 % para SNPs y sobre el 60 % para Indels, Inserciones y Deleciones. Por lo que en general se tiene casi un 94 % de *F1-Score*. La precisión obtenida de Deleciones es casi 20 puntos más alta que el *Recall*, debido a eso el *F1-Score* obtiene un 60 %.

BI-LSTM	Precision	Recall	F1-Score
Overall	**0,9586**	**0,9219**	**0,9399**
SNP	0,9915	0,9896	0,9906
Indel	0,6993	0,5232	0,5986
Ins	0,6798	0,5221	0,5906
Del	0,7197	0,5243	0,6067

Tabla 4.3: Resultados arquitectura *Bidirectional* LSTM.

4.4 CNN + Transformer

La Red Neuronal Convolucional con el Transformer presentada en la Figura 3.8 obtiene un *F1-Score* del 98 % para SNPs y sobre el 47 % para Indels, Inserciones y Deleciones. Por lo que en general obtiene un 90 % de *F1-Score*. Los resultados se exponen en la Tabla 4.4. Para esta arquitectura se tiene que los valores obtenidos de *Recall* son más altos que las *Precision*.

CNN+TRANS	Precision	Recall	F1-Score
Overall	**0,8881**	**0,9179**	**0,9028**
SNP	0,9739	0,9867	0,9803
Indel	0,4442	0,5127	0,4760
Ins	0,4425	0,5280	0,4815
Del	0,4460	0,4975	0,4704

Tabla 4.4: Resultados arquitectura CNN + Transformer.

4.5 Bi-linear CNN

La arquitectura de la Figura 3.9 inspirada en los problemas de *Computer Vision* obtiene los resultados de la Tabla 4.5. El *F1-Score* para SNPs es del 98 % mientras que para Indels,

Inserciones y Deleciones está sobre el 53 %. En general el *F1-Score* es del 92 %. Se tiene que las *Precisions* y *Recalls* están bastante igualadas.

BIL-CNN	Precision	Recall	F1-Score
Overall	**0,9209**	**0,9230**	**0,9219**
SNP	0,9887	0,9892	0,9890
Indel	0,5261	0,5329	0,5295
Ins	0,5756	0,5218	0,5474
Del	0,4863	0,5439	0,5135

Tabla 4.5: Resultados arquitectura *Bi-linear* CNN.

4.6 Benchmarking con el Estado del Arte

En este apartado se realiza el *benchmarking* o evaluación comparativa entre los modelos entrenados en este trabajo y el Estado del Arte para la detección de variantes. La comparativa se realiza desde 5 perspectivas de estudio: SNPs, Inserciones, Deleciones, Indels y resultados en general. En las tablas de resultados se marcan con negrita los mejores resultados para cada métrica y se subrayan aquellas propuestas neuronales de este trabajo que mejor rendimiento tienen. Además, en las tablas comparativas se añade la columna C.I. correspondiente al Intervalo de Confianza, que se calcula con la siguiente expresión:

$$C.I. = 1,96 * \sqrt{\frac{precision(1 - precision)}{n}}$$

La Tabla 4.6 recoge los resultados de *Precision*, *Recall* y *F1-Score* en la detección de SNP de las 5 arquitecturas entrenadas y las dos principales tecnologías del Estado del Arte para las *Third Generation Sequencing* (TGS). En cuanto a la *Precision*, la CNN+Transformer y la GRU P.E.P.P.E.R. obtienen un 97,4 %, siendo los valores más bajos. Pero la propuesta de The University of Hong Kong Clair3 obtiene el tercer valor más bajo de *Precision* para SNPs. Mientras que la *Precision* más alta es la obtenida por PEPPER-Margin-DeepVariant con un 99,5 %. Cabe destacar que la arquitectura *Bidirectional* LSTM no se aleja mucho del resultado de PEPPER-Margin-DeepVariant, ya que alcanza un 99,1 %.

Sobre los resultados de *Recall* obtenidos se tiene que el más bajo es el de la arquitectura CNN+Transformer, con un 98,6 %, mientras que la MLP, Bi-LSTM y Bi-linear CNN obtienen 98,9 %. El valor más alto de *Recall* lo consigue PEPPER-Margin-DeepVariant y Clair3 con 99,78 % y 99,72 %. La arquitectura propuesta en este trabajo que mayor *Recall* alcanza es la GRU P.E.P.P.E.R. con 99,2 %.

Debido a la naturaleza desbalanceada del problema, la columna correspondiente al *F1-Score* es más importante. PEPPER-Margin-DeepVariant logra un 99,6 % de *F1-Score* y Clair3 un 98,6 %. La arquitectura propuesta que mayor valor alcanza es la Bi-LSTM con un 99,06 % de *F1-Score* superando a la tecnología del Estado del Arte Clair3.

Para las Inserciones, los resultados se presentan en la Tabla 4.7. En cuanto a *Precision*, el valor más bajo lo obtiene la arquitectura CNN+Transformer con un 44,2 %. Por otro lado, el valor más alto lo consigue Clair3 con un 81,1 %, seguido de PEPPER-Margin-DeepVariant con un 78,5 %. Y el tercer puesto lo tiene la Bi-LSTM alcanzando 67,9 % de *Precision*.

Los valores de *Recall* obtenidos son más bajos que los de *Precision*. Por ejemplo, los más altos son: primero por parte de PEPPER-Margin-DeepVariant con un 60 %, Clair3

Arquitectura	Precision	Recall	F1-Score	C.I.
GRU PEPPER	0,9743	0,9926	0,9834	0,00080
MLP	0,9828	0,9898	0,9863	0,00072
BI-LSTM	0,9915	0,9896	0,9906	0,00051
CNN+TRANS	0,9739	0,9867	0,9803	0,00088
BIL-CNN	0,9887	0,9892	0,9890	0,00058
PEPPER-MARGIN-DEEPVARIANT	**0,9951**	**0,9978**	**0,9964**	0,00029
CLAIR3	0,9750	0,9972	0,9860	0,00079

Tabla 4.6: *Benchmarking* en SNPs.

con un 59,1 % y BCNN+Transformer con un 52,8 %. Mientras que el valor más bajo lo alcanza la arquitectura GRU P.E.P.P.E.R. con un 49,9 %.

Debido a la caída de los valores de *Recall*, el *F1-Score* obtiene 68 %, 68,4 % y 59 % para PEPPER-Margin-DeepVariant, Clair3 y Bi-LSTM, siendo estos tres los más altos. Y el más bajo es el modelo CNN+Transformer con un 48,8 %.

Arquitectura	Precision	Recall	F1-Score	C.I.
GRU PEPPER	0,5384	0,4995	0,5183	0,00750
MLP	0,6342	0,5159	0,5690	0,00816
BI-LSTM	0,6798	0,5221	0,5906	0,00807
CNN+TRANS	0,4425	0,5280	0,4815	0,00743
BIL-CNN	0,5756	0,5218	0,5474	0,00811
PEPPER-MARGIN-DEEPVARIANT	0,7854	**0,6001**	0,6803	0,00549
CLAIR3	**0,8111**	0,5913	**0,6840**	0,00728

Tabla 4.7: *Benchmarking* en Inserciones.

En cuanto a la Deleciones, los resultados se muestran en la Tabla 4.8. La *Precision* más alta la consigue PEPPER-Margin-DeepVariant con un 90,3 %, mientras que la segunda posición es para la arquitectura Bi-LSTM con un 71,9 % y el tercer valor más alto es de Clair3 con un 70 %. En cuanto al valor más bajo, lo obtiene la arquitectura CNN+Transformer, con un 44,6 %.

Los valores de *Recall* son bastante más bajos que los de *Precision*, ya que por primera vez, el valor más alto no es de PEPPER-Margin-DeepVariant, sino de Clair3 con un 58,7, seguido de la arquitectura *Bi-linear* CNN que alcanza un 54,3 % y en tercera posición PEPPER-Margin-DeepVariant con un 52,9 %. El valor más bajo es de la red CNN+Transformer con un 49,7 %. Cabe destacar que el resultado de la Bi-LSTM se mantiene en concordancia con el de PEPPER-Margin-DeepVariant, consiguiendo un 52,4 %.

Sobre los valores de *F1-Score*, se tiene que PEPPER-Margin-DeepVariant alcanza el valor más alto con un 66,7 %, el segundo valor es Clair3 con un 63,9 % y el tercero es la arquitectura Bi-LSTM con un 60,6 %. El valor más bajo es el de la CNN+Transformer con un 47 %.

La cuarta comparativa corresponde a la Tabla 4.9 y al estudio de Indels. Sobre *Precision*, PEPPER-Margin-DeepVariant vuelve al desmarcarse del resto obteniendo un 83,7 %, seguido de Clair3 con un 75,1 % y Bi-LSTM con 69,9 %. Los valores más bajos son 44,4 % y 52,6 % de la CNN+Transformer y *Bi-linear* CNN.

Sobre el *Recall*, se tiene que Clair3 vuelve a superar a PEPPER-Margin-DeepVariant con un 58,9 %, PEPPER-Margin-DeepVariant consigue un 56,4 % y el tercer mejor valor lo alcanza la *Bi-linear* CNN con un 53,2 %. El valor más bajo es 50,5 % por parte de la aquitectura GRU P.E.P.P.E.R..

Arquitectura	Precision	Recall	F1-Score	C.I.
GRU PEPPER	0,5436	0,5119	0,5273	0,00741
MLP	0,6257	0,5123	0,5634	0,00814
BI-LSTM	0,7197	0,5243	0,6067	0,00787
CNN+TRANS	0,4460	0,4975	0,4704	0,00752
BIL-CNN	0,4863	0,5439	0,5135	0,00766
PEPPER-MARGIN-DEEPVARIANT	**0,9034**	0,5294	**0,6676**	0,00410
CLAIR3	0,7005	**0,5875**	0,6391	0,00728

Tabla 4.8: *Benchmarking* en Deleciones.

En cuanto a *F1-Score*, PEPPER-Margin-DeepVariant consigue la primera posición con un 67,4 %, seguido de Clair3 con un 66 % y de Bi-LSTM con 59,8 %. Las arquitecturas que peores métricas consiguen son GRU y *Bi-linear* CNN con 52,2 % y 52,9 %.

Arquitectura	Precision	Recall	F1-Score	C.I.
GRU PEPPER	0,5411	0,5058	0,5228	0,00527
MLP	0,6299	0,5141	0,5661	0,00576
BI-LSTM	0,6993	0,5232	0,5986	0,00564
CNN+TRANS	0,4442	0,5127	0,4760	0,00528
BIL-CNN	0,5261	0,5329	0,5295	0,00559
PEPPER-MARGIN-DEEPVARIANT	**0,8372**	0,5644	**0,6742**	0,00356
CLAIR3	0,7514	**0,5894**	0,6606	0,00500

Tabla 4.9: *Benchmarking* en Indels.

Finalmente, en la Tabla 4.10 se presentan los resultados de *Precision*, *Recall*, *F1-Score* en general. En cuanto a *Precision*, el valor más bajo lo obtiene la CNN+Transformer con un 88,8 %. Mientras que los valores más altos son de PEPPER-Margin-DeepVariant, Bi-LSTM y Clair3 con 97,9 %, 95,8 % y 94,9 %.

En cuanto al Recall, Clair3 alcanza la primera posición con un 93,9 %, seguido de PEPPER-Margin-DeepVariant con un 93,5 %. Mientras que la tercera posición es de la arquitectura GRU con un 92,34 %. Cabe destacar que a excepción de la CNN+Transformer, todos los modelos neuronales propuestos obtienen resultados alrededor del 92 %.

La medida más significativa para esta tarea es el *F1-Score*. Y el valor más alto es del 95,6 % por parte de PEPPER-Margin-DeepVariant, el segundo es de Clair3 con un 94,4 % y el tercero de la arquitectura Bi-LSTM con un 93,9 %. Por otro lado, el peor es de la CNN+Transformer con un 90,2 %.

Arquitectura	Precision	Recall	F1-Score	C.I.
GRU PEPPER	0,9171	0,9234	0,9203	0,00126
MLP	0,9401	0,9208	0,9304	0,00120
BI-LSTM	0,9586	0,9219	0,9399	0,00101
CNN+TRANS	0,8881	0,9179	0,9028	0,00155
BIL-CNN	0,9209	0,9230	0,9219	0,00135
PEPPER-MARGIN-DEEPVARIANT	**0,9791**	0,9358	**0,9569**	0,00054
CLAIR3	0,9498	**0,9393**	0,9445	0,00102

Tabla 4.10: *Benchmarking* en general.

CAPÍTULO 5
Conclusiones y Trabajos Futuros

En este capítulo se exponen si los objetivos se han cumplido y se presenta la discusión de los resultados obtenidos, sus correspondientes conclusiones y finalmente se analizan y exponen las futuras líneas de investigación que abre este Trabajo Final de Máster.

El primer objetivo comprendía el estudio del Estado del Arte. Se ha profundizado en los dos enfoques tradicionales, el **Bayesiano** y el **basado en el Ensamblaje**. Además, se han presentado las tecnologías que emplean *Deep Learning* para abarcar la tarea de detección de variantes y sus flujos de trabajo. Se ha comprobado que las técnicas neuronales son capaces de obtener mejores resultados que los métodos tradicionales como GATK [10] y tienen mejor rendimiento en regiones complejas del ADN.

El tratamiento y manipulación de datos genómicos era el segundo objetivo. Se han presentado los formatos de archivos más extendidos para cada conjunto de datos:

- **BAM**: almacena las lecturas secuenciadas alineadas de un organismo con un Genoma de referencia.
- **FASTA**: formato para almacenar lecturas secuenciadas de ADN, en este trabajo se ha utilizado para almacenar el Genoma de referencia.
- **VCF**: formato más extendido para etiquetar las variantes de un organismo.

A parte de conocer los formatos y distintas formas de almacenar la información, se ha estudiado el preprocesado y tratamiento necesario para poder trabajar con datos genómicos. Como por ejemplo, utilizar algoritmos de alineamiento o alineadores, como Minimap2 [20] que mediante las lecturas de un organismo y un Genoma de referencia, permiten obtener un archivo BAM.

El tercer objetivo trataba de diseñar diferentes propuestas de arquitecturas neuronales para abarcar el problema de la detección de variantes. Por lo que se han propuesto arquitecturas sencillas como un Perceptrón Multicapa o una GRU simple, hasta arquitecturas avanzadas recurrentes (*Bidirectional* LSTM) o modelos de *Computer Vision* como CNN+Transformer y *Bi-linear* CNN.

El cuarto objetivo trata de realizar una evaluación comparativa entre los diferentes modelos propuestos y los del Estado del Arte con el fin de abrir una discusión y exponer las líneas de investigación que abre este trabajo para futuras investigaciones.

La evaluación comparativa desde el punto de vista general (Tabla 4.10) da lugar a la gráfica que se presenta en la Figura 5.1. Los primeros cinco modelos son los propuestos y entrenados en este trabajo, mientras que los dos últimos son PEPPER-Margin-DeepVariant y Clair3, que son las tecnologías el Estado del Arte que mejores resultados obtienen para *Third Generation Sequencing* (TGS).

Se puede apreciar en la Figura 5.1, que el modelo neuronal CNN+Transformer, es el que peores resultados alcanza, a pesar de ser una arquitectura especializada para el reconocimiento de imágenes con mecanismos de Self-Attention. Por otro lado, se puede comentar lo mismo de la arquitectura *Bi-linear* CNN. Aunque se trata de una arquitectura especializada y compleja, inspirada en el campo del *Computer Vision* para problemas donde las clases tienen diferencias sutiles, no ha logrado aportar resultados interesantes. Aunque cabe destacar que para Deleciones e Indels (Tablas 4.8 y 4.9), esta obtiene el mayor *Recall* de las topologías propuestas.

Otro punto interesante de los resultados desde el punto de vista general, es que la propuesta de [32] de la arquitectura GRU se queda atrás en comparación con el Perceptrón Multicapa y la *Bidirectional* LSTM. A pesar de la sencillez de la MLP, esta ha logrado el segundo puesto en las arquitecturas propuestas, con un 93,04 % de *F1-Score* en los resultados generales.

Lo más destacable es que la arquitectura neuronal de este trabajo que mejor rendimiento tiene, alcanzando casi a Clair3, es la *Bidirectional* LSTM. Con un *F1-Score* de 93,99 %, esta arquitectura proporciona una muy buena flexibilidad y adaptación al problema. Esta obtiene una *Precision* del 95,86 % pasando a Clair3 y únicamente superada por PEPPER-Margin-DeepVariant con un 97,91 %.

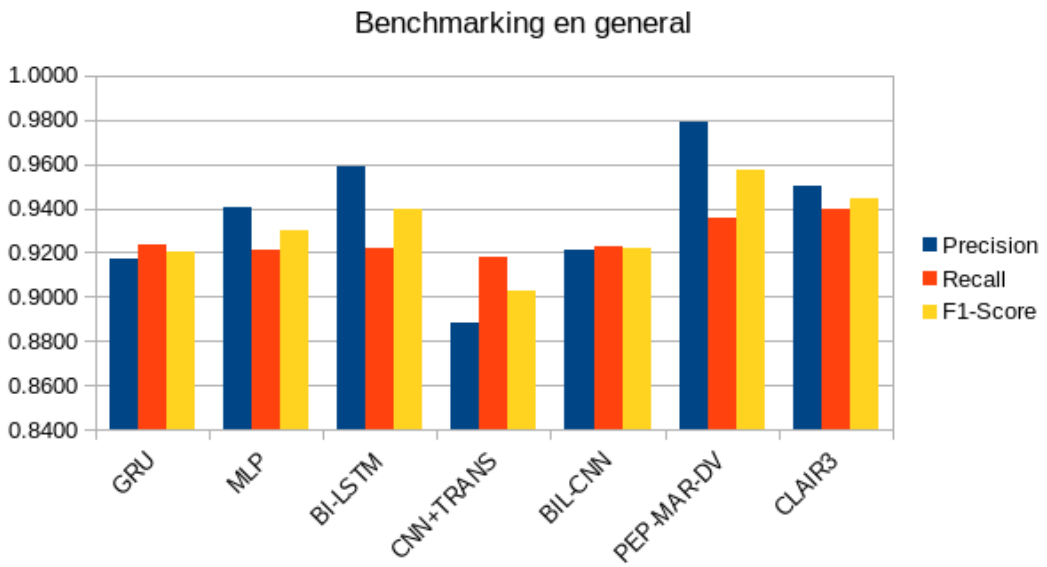

Figura 5.1: Gráfica de *benchmarking* en general.

En la Figura 5.2, se presenta un gráfico del desempeño en general de los modelos propuestos en este trabajo, donde es más que evidente la superioridad de la *Bidirectional* LSTM frente al resto. Además, el comportamiento de la GRU y la *Bi-linear* CNN son muy parecidos.

Debido al estudio de diferentes arquitecturas neuronales expuesto en este trabajo, se puede afirmar que las *Bidirectional* LSTM o Redes Neuronales Recurrentes Bidireccionales son capaces de entender y proporcionar un rendimiento más alto que con el uso de neuronas densas normales (MLP) o Redes Neuronales Convolucionales. Ya que, como se ha comprobado en los experimentos, la Bi-LSTM es capaz de alcanzar y superar a Clair3 en determinadas situaciones, como en la detección de SNPs. Además, se debe tener en cuenta que Clair3 y PEPPER-Margin-DeepVariant han sido entrenadas con diferentes conjuntos de datos y la Bi-LSTM sólo con uno (HG001).

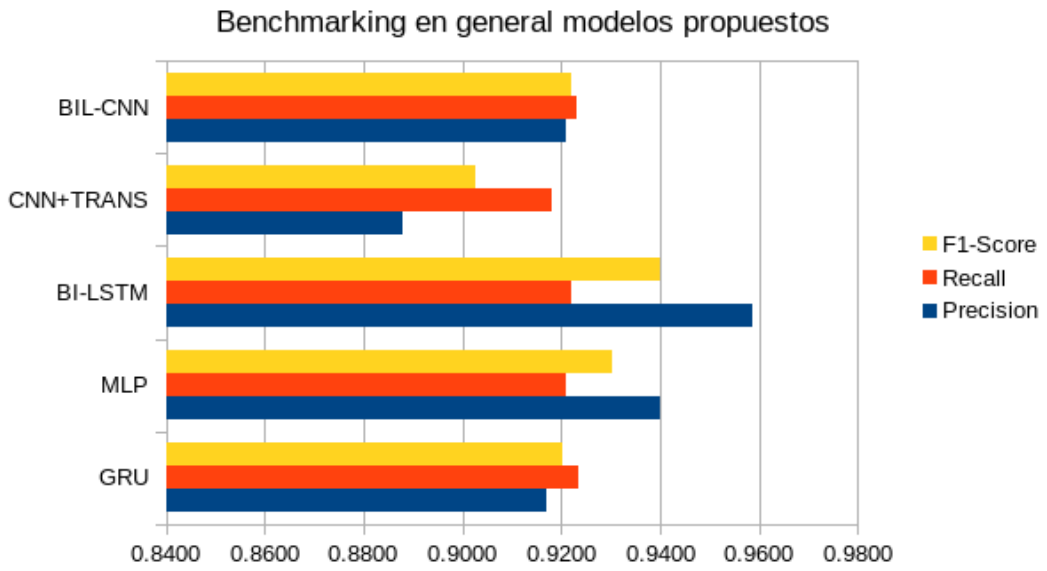

Figura 5.2: Gráfica de benchmarking en general de los modelos propuestos en este trabajo.

Otro punto a destacar de los resultados son las bajas tasas de *F1-Score* para las Inserciones, Deleciones e Indels. Este fenómeno ocurre tanto para las tecnologías del Estado del Arte como para las entrenadas en este trabajo. Debido a que la naturaleza del problema es desbalanceada y en el ADN de los individuos se producen más SNPs que Inserciones, Deleciones o Indels las tasas de detección de SNPs son mayores, ya que este caso tiene muchas muestras.

En la Figura 5.3 se muestran los resultados de la Tabla 4.9 correspondiente al *benchmarking* de Indels, donde se destaca que en los modelos propuestos los F1-Scores obtienen valores bajos. Esto es lógico, ya que en este trabajo se ha utilizado únicamente la muestra de un sólo organismo, por lo que si se usan más muestras se tendría una mejor representación de las variaciones relacionadas con Inserciones, Deleciones e Indels y se podrían alcanzar los porcentajes de PEPPER-Margin-DeepVariant o Clair3.

Pero los valores bajos en Inserciones, Deleciones e Indels no es únicamente debido a la falta de datos, sino también a dos principales factores. El primero es que en el ADN existen regiones difíciles de modelizar como ya se ha tratado en capítulos anteriores. Y el segundo factor es que la información que se está tratando para detectar las variaciones es un resumen o Pileup de lo que realmente es el problema. De ahí puede darse, que la codificación que se realiza de las regiones candidatas no sea lo suficientemente representativa en regiones difíciles del ADN. Por lo que aunque se diese el caso de tener infinitas muestras de entrenamiento, no se podrían alcanzar tasas altas de acierto, ya que el problema puede no estar bien representado en ciertas regiones.

Con este trabajo se demuestra también cómo el *Deep Learning* es capaz de abarcar gran variedad de problemas y mejorar las tecnologías tradicionales de diferentes áreas. Ya que este subcampo del Machine Learning se aplicó por primera vez de la mano de Google a las *Next Generation Sequencing* (NGS) y consiguieron superar a los métodos tradicionales. Debido a esto, desde el nacimiento de las *Third Generation Sequencing* (TGS), el uso de la tecnología *Deep Learning* ha sido el estándar para la detección de variantes.

Otra ventaja del uso de tecnologías neuronales es la eliminación de la necesidad del conocimiento experto, ya que estos algoritmos consiguen mediante el reconocimiento de patrones adaptarse a cualquier tipo de tarea. Este punto es destacable, ya que la mayo-

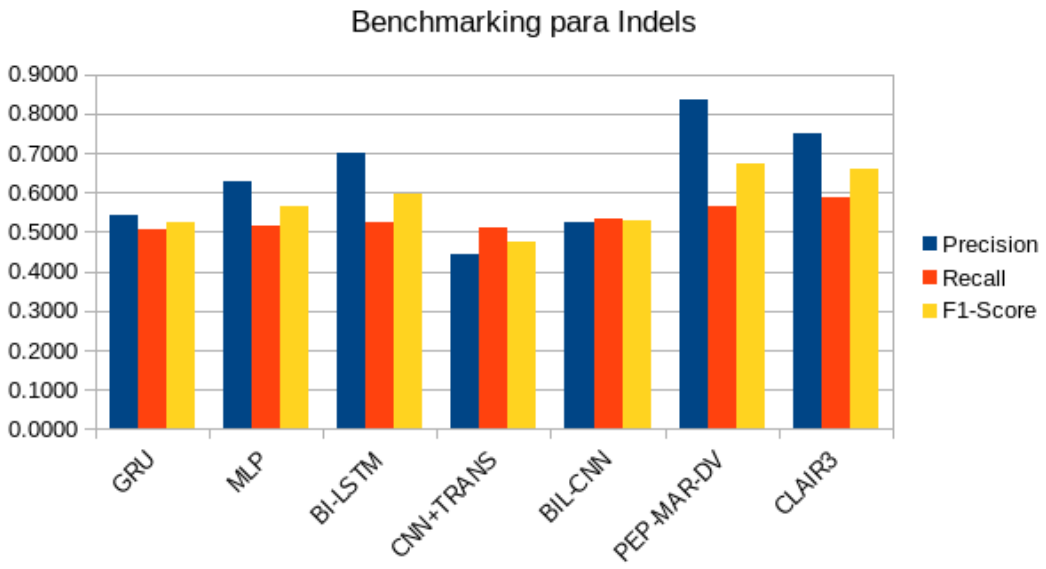

Figura 5.3: Gráfica de benchmarking para Indels.

ría de tecnologías tradicionales para la detección de variantes como GATK [10] utilizan conocimiento experto y parametrizaciones específicas derivadas de este conocimiento.

Debido al amplio abanico de tecnologías de secuenciación y las particularidades, características y tasas de error propias de cada una, el enfoque tradicional era dependiente de la tecnología. Pero con la llegada de DeepVariant [14], PEPPER-Margin-DeepVariant [32] y Clair3 [34] y convertir el problema de detección de variantes en uno de reconocimiento de imágenes, el paradigma cambia. El enfoque deja de depender de la tecnología y las Redes Neuronales Artificiales son capaces de tratar las variantes candidatas con independencia de la tecnología de secuenciación, ya que a estas se les pasa una imagen o Pileup.

Pero esto no es cierto del todo, ya que el flujo de trabajo del Estado del Arte en la actualidad se divide en dos fases independientes: Preproceso o Extracción de Características y Clasificación. Por lo que en la fase de Extracción de Características es donde se encuentran o se preprocesan las secuencias de forma diferente dependiendo de la tecnología. Un ejemplo es como DeepVariant está desarrollada para NGS, pero no para TGS. La versión de lecturas largas o TGS de DeepVariant es PEPPER-Margin-DeepVariant. Por lo que se consigue un modelo neuronal o clasificador independiente de las tecnologías de secuenciación, siempre y cuando el preproceso de las lecturas siga la misma lógica para NGS o TGS.

Un punto interesante a tratar es desde el punto de vista de las TGS. En la Tabla 1.2 se realiza una comparativa entre TGS y NGS, donde la principal característica de las *Third Generation Sequencing* son sus lecturas largas y el alto error que estas conllevan. Por lo que una mejora en las tecnologías de secuenciación, que permitan disminuir el error de secuenciación puede dar lugar a mejores tasas a la hora de detectar variantes en regiones difíciles o conflictivas del ADN.

Debido al interés e importancia que este problema conlleva para el diagnóstico, predicción y prevención de enfermedades, así como el desarrollo de tratamientos personalizados y efectivos. Es interesante exponer las líneas de investigación que este Trabajo Final de Máster abre para la continuación en la Tesis Doctoral.

Como se ha mencionado, el esquema principal del Estado del Arte se basa en dos fases independientes: Extracción de Características y Clasificador. Este enfoque abre dos perspectivas interesantes de estudio. La primera sería el análisis y diseño de arquitecturas capaces de realizar las dos fases utilizando una única Red Neuronal Artificial que se encargue de la Extracción de Características y la Clasificación. La segunda perspectiva sería cambiar el enfoque de transformar el problema al área del reconocimiento de imágenes, ya que este permite buenos resultados, pero no es capaz de representar las regiones conflictivas del ADN. El cambio sería que en lugar de utilizar imágenes, usar las secuencias de ADN directamente, que corresponde con una forma más natural para esta tarea. Por lo que estas dos perspectivas abren un camino al diseño y evaluación de modelos End-to-End completos para la detección de variantes.

Además de proponer diferentes formas para abarcar el problema, se abren nuevas fronteras cada día más relacionadas con la Ética y Explicabilidad de los modelos de Inteligencia Artificial. Debido a que el área en el que se enmarca este trabajo es la Genómica y la Medicina de Precisión, es extremadamente importante el análisis del comportamiento de las tecnologías para el diagnóstico de enfermedades. Ya que si a un paciente se le diagnostica que desarrollará un cáncer y no lo acaba teniendo, es mucho mejor que si no se le diagnostica y lo acaba teniendo. Por lo que el estudio de las tecnologías desde una perspectiva XAI (*Explainable Artificial Intelligence*) es otro de los caminos que se abren para posteriores estudios.

Glosario

A Adenina. 1, 3

ADN Ácido Desoxirribonucleico. 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 14, 18, 20, 24, 33, 35, 36, 37

ARN Ácido ribonucleico cuya función es transferir información del genoma a las proteínas. 7

Bases componentes de ADN, existen 4 tipos: Adenina (A), Guanina (G), Citosina (C) y Timina (T). Es sinónimo de nucleótido. 1, 2, 3, 4, 5, 7, 9

C Citosina.. 1, 3

Clair3 tecnología del Estado del Arte para la detección de variantes TGS propuesta por The University of Hong Kong. 13, 14, 15, 18, 29, 30, 31, 33, 34, 35, 36

CNN Convolutional Neural Network. 14, 22, 23, 29, 30, 31, 33, 34

Cromosoma paquete ordenado de ADN que se encuentra en el núcleo de la célula. 24, 25

DeepVariant tecnología del Estado del Arte para la detección de variantes NGS propuesta por Google. 4, 10, 11, 12, 14, 15, 24, 36

Deleción tipo de variación, que implica la pérdida de uno o más nucleótidos de un segmento de ADN. 18, 19, 20

F1-Score medida de precisión. 25, 26, 27, 28, 29, 30, 31, 34, 35

G Guanina. 1, 3

Genoma de referencia secuencia estándar completa del ADN de un organismo. 3, 4, 5, 7, 9, 11, 12, 15, 17, 18, 19, 20, 33

Genómica campo de la biología que se encarga de estudiar la secuencia de ADN completa de los organismos. 1, 7, 37

GIAB Genome In A Bottle. 7, 17, 23

GRU Gated Recurrent Unit. 21, 23, 29, 30, 31, 33, 34

Heterocigoto Alterado las lecturas secuenciadas proponen más de un nucleótido diferente al de la secuencia de referencia. 20

Homocigoto Alterado las lecturas secuenciadas proponen un nucleótido diferente al de la secuencia de referencia. 20

Homocigoto Referencia las lecturas secuenciadas coinciden con la de referencia.. 20, 25

Indel contracción de las palabras Inserción o Deleción, se utiliza para referirse a inserciones y/o deleciones de ADN. 3, 7, 9, 10, 12, 15, 17, 26, 27, 28, 29, 30, 35

Inserción tipo de variación, que implica la inserción de uno o más nucleótidos en un segmento de ADN. 18, 19, 20

LSTM celda neuronal recurrente, conocida como Long Short-Term Memory. 21, 22, 28, 29, 30, 31, 33, 34

Medicina de Precisión es un nuevo enfoque de la medicina que se centra en el individuo, sus genes, su estilo de vida y su entorno. 1, 7, 37

NGS Next Generation Sequencing.. 3, 4, 6, 10, 12, 14, 15, 35, 36

Nucleótido componente de ADN, existen 4 tipos: Adenina (A), Guanina (G), Citosina (C) y Timina (T). Es sinónimo de base. 3, 7, 10, 15, 19, 20

ONT Oxford Nanopore Technologies. 6, 7, 15, 17, 18

PEPPER-Margin-DeepVariant tecnología del Estado del Arte para la detección de variantes TGS propuesta por Google y diferentes centros de investigación genómica. 12, 13, 15, 21, 29, 30, 31, 33, 34, 35, 36

Pileup imagen resumen de una variantes candidata. 10, 11, 12, 18, 19, 20, 22, 35, 36

Precision medida de precisión. 25, 26, 27, 28, 29, 30, 31, 34

Recall medida de precisión. 25, 26, 27, 28, 29, 30, 31, 34

RNN Recurrent Neural Network. 14, 21, 23

Self-Attention mecanismo de atención que es capaz de relacionar diferentes posiciones de una misma secuencia con el fin de representar dicha secuencia. 22, 23, 34

SNP Single Nucleotide Polymorphism. 3, 7, 12, 17, 18, 19, 20, 26, 27, 28, 29, 34, 35

SNV Single Nucleotide Variation. 10, 15

T Timina. 1, 3

TGS Third Generation Sequencing. 3, 6, 12, 13, 14, 15, 18, 29, 33, 35, 36

Transformer arquitectura de Deep Learning que incorpora el mecanismo de Self-Attention. 22, 28, 29, 30, 31, 33, 34

Bibliografía

[1] National Human Genome Research Institute. Genómica. Consultado en https://www.genome.gov/es/genetics-glossary/Genomica.

[2] Noelia Izquierdo. *El futuro de la genómica, a la espera de un plan nacional operativo*. Revista Española de Economía de la Salud, 2022. Consultado en https://economiadelasalud.com/topics/difusion/futuro-genomica-plan-nacional-operativo-estrategia-gobierno/.

[3] Hodson, R. Precision medicine. *Nature 537*, S49 (2016). https://doi.org/10.1038/537S49a.

[4] MedlinePlus. ¿Qué es el ADN?. Consultado en https://medlineplus.gov/spanish/genetica/entender/basica/adn/.

[5] YourGenome. What is the 'Central Dogma'?. 2021. Consultado en https://www.yourgenome.org/facts/what-is-the-central-dogma.

[6] Bamshad, M., Ng, S., Bigham, A. et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet 12*, 745–755 (2011). https://doi.org/10.1038/nrg3031.

[7] Jiao, Y., Zhao, H., Ren, L. et al. Genome-wide genetic changes during modern breeding of maize. *Nat Genet 44*, 812–815 (2012). https://doi.org/10.1038/ng.2312.

[8] A. Ramachandran, H. Li, E. Klee, S. S. Lumetta and D. Chen. Deep Learning for Better Variant Calling for Cancer Diagnosis and Treatment. *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, (2018), pp. 16-21, https://doi.org/10.1109/ASPDAC.2018.8297276.

[9] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics, Volume 25*, (2009), pp. 16-21, Issue 16, Pages 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.

[10] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M. A.. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research vol. 20*, (2010), 1297-303, https://doi.org/10.1101/gr.107524.110.

[11] Heng Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, (2014), 30(20):2843-51. https://doi.org/10.1093/bioinformatics/btu356.

[12] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapess. *Science*, (2013), 29;339(6127):1546-58. https://doi.org/10.1126/science.1235122.

[13] Hodgson, S. Mechanisms of inherited cancer susceptibility. *J Zhejiang Univ. Sci.*, (2008), B 9, 1–4. https://doi.org/10.1631/jzus.B073001.

[14] Poplin, R., Chang, PC., Alexander, D. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.*, (2018), 36, 983–987. https://doi.org/10.1038/nbt.4235.

[15] Lin, YL., Chang, PC., Hsu, C. et al. Comparison of GATK and DeepVariant by trio sequencing. *Sci. Rep.*, (2022), 12, 1809. https://doi.org/10.1038/s41598-022-05833-4.

[16] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, (2018), Volume 34, Issue 18, Pages 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

[17] Human Genome Overview - Genome Reference Consortium. NCBI. Consultado en https://www.ncbi.nlm.nih.gov/grc/human.

[18] DePristo, M., Banks, E., Poplin, R. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, (2011), 43, 491–498. https://doi.org/10.1038/ng.806.

[19] Mobley, Immy. DNA Sequencing: How to Choose the Right Technology. *Front Line Genomics*, (2021). Consultado en https://frontlinegenomics.com/dna-sequencing-how-to-choose-the-right-technology/.

[20] Li, Z., Wang, Y., Wang, F. A study on fast calling variants from next-generation sequencing data using decision tree. *BMC Bioinformatics*, (2018), 19, 145. https://doi.org/10.1186/s12859-018-2147-9.

[21] Cornelis A. Albers, Gerton Lunter, Daniel G. MacArthur, et al. Dindel: Accurate indel calls from short-read data. *Genome Research*, (2014), vol. 46, no. 8, pp. 912-8. http://www.genome.org/cgi/doi/10.1101/gr.112326.110.

[22] Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. *Genome Research*, (2009), 19(6):1117-1123. https://doi.org/10.1101%2Fgr.089532.108.

[23] Burland, T.G. DNASTAR's Lasergene Sequence Analysis Software. *Bioinformatics Methods and Protocols. Methods in Molecular Biology*, (2000), vol 132. Humana Press, Totowa, NJ. https://doi.org/10.1385/1-59259-192-2:71.

[24] Silva GG, Dutilh BE, Matthews TD, et al. Combining de novo and reference-guided assembly with scaffold_builder. *Source Code Biol Med*, (2013), 8(1):23. https://doi.org/10.1186%2F1751-0473-8-23.

[25] Feng Zeng, Rui Jiang, Ting Chen. PyroHMMvar: a sensitive and accurate method to call short indels and SNPs for Ion Torrent and 454 data. *Bioinformatics*, (2013), Volume 29, Issue 22, Pages 2859–2868. https://doi.org/10.1093/bioinformatics/btt512.

[26] Rimmer, A., Phan, H., Mathieson, I. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.*, (2014), 46, 912–918. https://doi.org/10.1038/ng.3036.

[27] Szegedy, Crhistian, et al. Rethinking the Inception Architecture for Computer Vision. *ArXiv*, (2015). https://doi.org/10.48550/arXiv.1512.00567.

[28] Nattestad, Maria, et al. Looking Through DeepVariant's Eyes. DeepVariant Blog, 2020. Consultado en https://google.github.io/deepvariant/posts/2020-02-20-looking-through-deepvariants-eyes/.

[29] Szegedy, Christian, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *ArXiv*, (2016). https://doi.org/10.48550/arXiv.1602.07261.

[30] google/deepvariant: DeepVariant is an analysis pipeline that uses a deep neural network to call genetic variants from next-generation DNA sequencing data. GitHub. Consultado en https://github.com/google/deepvariant.

[31] Ramachandran, A., Lumetta, S.S., Klee, E.W. et al. HELLO: improved neural network architectures and methodologies for small variant calling. *BMC Bioinformatics*, (2021), 22, 404. https://doi.org/10.1186/s12859-021-04311-4.

[32] Shafin, K., Pesout, T., Chang, PC. et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods*, (2021), 18, 1322–1332. https://doi.org/10.1038/s41592-021-01299-w.

[33] kishwarshafin/pepper: PEPPER-Margin-DeepVariant. GitHub. Consultado en https://github.com/kishwarshafin/pepper.

[34] Zheng, Zhenxian, et al. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *BioRXiv*, (2021). https://doi.org/10.1101/2021.12.29.474431.

[35] Luo, Ruibang, et al. Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nature Machine Intelligence*, (2020), vol. 2, pp. 220-227. https://doi.org/10.1038/s42256-020-0167-4.

[36] Luo, Ruibang, et al. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nature Communications*, (2019), vol. 10 no. 998. https://doi.org/10.1038/s41467-019-09025-z.

[37] Genome in a Bottle. NIST. Consultado en https://www.nist.gov/programs-projects/genome-bottle.

[38] Danecek, Petr, et al. Twelve years of SAMtools and BCFtools. *Giga Science*, (2021), vol. 10, no. 2. https://doi.org/10.1093/gigascience/giab008.

[39] Minikel, Eric. Forward and reverse reads in paired-end sequencing. CureFFI.org, 2012. Consultado en https://www.cureffi.org/2012/12/19/forward-and-reverse-reads-in-paired-end-sequencing/.

[40] Lawrence, Michale. Introduction to Variant Calling. Bioconductor.org, 2014. Consultado en https://bioconductor.org/help/course-materials/2014/CSAMA2014/3_Wednesday/lectures/VariantCallingLecture.pdf.

[41] Cho, Kyunghyun, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *ArXiv*, (2014). https://doi.org/10.48550/arXiv.1406.1078.

[42] Hochreiter, Sepp, and Jürgen Schmidhuber. Long Short-Term Memory. *MIT Press Direct*, (1997), vol. 9, no. 8. https://doi.org/10.1162/neco.1997.9.8.1735.

[43] Vaswani, Ashish, et al. Attention Is All You Need. *ArXiv*, (2017). https://doi.org/10.48550/arXiv.1706.03762.

[44] Lin, Tasung-Yu, et al. Bilinear CNN Models for Fine-Grained Visual Recognition. *2015 IEEE International Conference on Computer Vision (ICCV)*, (2015), pp. 1449-1457. https://doi.org/10.1109/ICCV.2015.170.

[45] Definición de ADN - Diccionario de cáncer del NCI - National Cancer Institute. Consultado en https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/adn.

[46] Illumina/hap.py: Haplotype VCF comparison tools. GitHub. Consultado en https://github.com/Illumina/hap.py.

[47] Prediger, Ellen. Allele vs Genotype vs Haplotype and More - Genotyping Terms | IDT. Integrated DNA Technologies, 2019. Consultado en https://eu.idtdna.com/pages/education/decoded/article/genotyping-terms-to-know.

[48] Wu, Leihong, et al. Direct comparison of performance of single nucleotide variant calling in human genome with alignment-based and assembly-based approaches. *Scientific Reports*, (2017), vol. 7, no. 10963. https://doi.org/10.1038/s41598-017-10826-9.

[49] Khan AR, Pervez MT, Babar ME, Naveed N, Shoaib M. A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. *Evolutionary Bioinformatics*, (2018), 10963. https://doi.org/10.1177/1176934318758650.

[50] Garrison, Erik, and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *ArXiv*, (2012). https://doi.org/10.48550/arXiv.1207.3907.