Revisión de algoritmos de última generación para procesos de análisis de datos genómicos

Ángel Canal-Alonso1, Noelia Egido1, Pedro Jiménez1, Javier Prieto[1], Juan Manuel Corchado[1]

[1]Departamento de Bioinformática y Biología Computacional, Instituto AIR,
Carbajosa de la Sagrada, España

Correo electrónico: acanal@air-institute.com

ABSTRACTO

La llegada de big data y tecnologías avanzadas de secuenciación genómica ha presentado desafíos en términos de procesamiento de datos para uso clínico. La complejidad de detectar e interpretar variantes genéticas, junto con la amplia gama de herramientas y algoritmos y la gran carga de trabajo computacional, ha hecho que el desarrollo de plataformas integrales de análisis genómico sea crucial para permitir a los médicos proporcionar rápidamente resultados genéticos a los pacientes. Este capítulo revisa y describe el proceso para analizar datos genómicos masivos utilizando tecnologías de lectura corta y larga, discutiendo el estado actual de las principales herramientas utilizadas en cada etapa y el papel de la inteligencia artificial en su desarrollo. También presenta DeepNGS (deepngs.eu), una plataforma web de análisis genómico de un extremo a otro, incluidas sus funciones y aplicaciones clave.

PALABRAS CLAVE

Genómica, medicina de precisión, sistema de gestión de flujo de trabajo, computación en la nube

1 INTRODUCCIÓN

La finalización del Proyecto Genoma Humano ha supuesto avances significativos en la secuenciación de alto rendimiento, convirtiendo esta tecnología en una herramienta de rutina utilizada en muchos laboratorios de investigación y centros genéticos para diversos fines. Se ha descubierto que nuestros genomas contienen las respuestas a muchas de las preguntas que la humanidad se ha estado planteando durante siglos. Sin embargo, este nuevo paradigma genera una gran cantidad de datos que los métodos computacionales tradicionales no pueden manejar, lo que resulta en el desarrollo de diversas herramientas y algoritmos para analizar y gestionar los datos recopilados por diferentes plataformas de secuenciación.

DeepNGS es una plataforma web que ayuda y simplifica el proceso de análisis de datos de muestras genómicas humanas para detectar e interpretar su posible patogenicidad y posible asociación con un fenotipo clínico. Con solo un clic, esta plataforma automatiza el tedioso proceso de análisis de datos y lo entrega a sistemas expertos de gestión de flujo de trabajo implementados tanto en la nube como en las instalaciones, para optimizar las cargas computacionales y la paralelización en un entorno confiable y seguro para trabajar con datos genómicos de Relevancia. El uso de nuevos algoritmos de IA permite una detección e interpretación clínicas óptimas, con métodos de extracción de patrones basados en aprendizaje automático que aprovechan la información de múltiples fuentes para crear un modelo adaptado a cada fenotipo.

2 CANAL DE ANÁLISIS DE DATOS GENÓMICOS

2.1 Evaluación de calidad

Los archivos FASTQ se utilizan para almacenar datos de secuencia de alto rendimiento y contienen datos sin procesar y puntuaciones de calidad para cada base.[1]. Los puntajes de calidad se representan como caracteres ASCII; un valor ASCII más alto indica un puntaje de calidad más alto. Los puntajes de calidad se utilizan para evaluar la precisión de las llamadas de base realizadas por el secuenciador y se pueden usar para filtrar lecturas o bases de baja calidad.

Recortar es el proceso de eliminar secuencias no deseadas o de baja calidad de los finales de las lecturas. Esto se puede hacer para mejorar la situación general.

calidad de los datos, o para eliminar secuencias de adaptadores que puedan haberse introducido durante el proceso de secuenciación. Hay muchas herramientas de software disponibles para recortar archivos FASTQ, y la herramienta y los parámetros específicos utilizados dependerán de las necesidades específicas del análisis.

FASTQC y Trimmomatic son dos herramientas comúnmente utilizadas para analizar y procesar datos de secuencia de alto rendimiento.[2], [3]. FASTQC es un software que proporciona una serie de gráficos y estadísticas para evaluar la calidad de las lecturas, incluidas medidas como la calidad promedio por base, la distribución de nucleótidos indeterminados (N) y el contenido de GC. Trimmomatic es una potente herramienta para filtrar secuencias utilizando múltiples parámetros y está optimizada para su uso con datos de la plataforma de secuenciación Illumina. Se puede utilizar para eliminar secuencias adaptadoras, que son tramos cortos de ADN que pueden introducirse durante el proceso de preparación de la biblioteca y pueden interferir con los análisis posteriores.

Hay muchas otras herramientas disponibles para analizar y procesar datos de secuencia de alto rendimiento, incluido NGSQC Toolkit[4], PRINSEQ[5], galaxia[6], PathoQC[7], Después del control de calidad[8], rápido[9], FastProNGS[10]y FQStat[11]. Estas herramientas ofrecen diferentes características y pueden ser más adecuadas para ciertos tipos de análisis o conjuntos de datos. Es importante evaluar cuidadosamente las opciones disponibles y elegir la herramienta que sea más adecuada para sus necesidades específicas.

2.2 Alineación y posprocesamiento

Mapear o alinear lecturas con un genoma de referencia es un paso importante en el análisis de datos de secuenciación de alto rendimiento. Implica colocar las lecturas en sus posiciones correctas dentro del genoma, lo que puede resultar un desafío debido a la complejidad del genoma y la presencia de regiones repetitivas o mal caracterizadas. Hay muchas herramientas diferentes disponibles para alinear lecturas con un genoma de referencia, que se pueden clasificar en términos generales según el algoritmo que utilizan.[12].

Algoritmos basados en hash, como RMAP[13], JABÓN[14], Novoalign[15]y CAMARONES[dieciséis], utilizan funciones hash para indexar rápidamente la posición de cada lectura, pero son propensas a errores. Algoritmos basados en el algoritmo de Smith-Waterman, como BFAST[17], son más precisos pero también requieren más tiempo. Algoritmos basados en la transformada de Burrows-Wheeler, como BWA[18], Corbata de moño[19]y SOAP2[20], ofrecen un equilibrio entre eficiencia, sensibilidad y especificidad, y son adecuados para lecturas breves.

Según estudios, los alineadores que mejor rendimiento ofrecen para la identificación de variantes son BWA, Bowtie, Novoalign y SOAP. BWA y Bowtie se encuentran entre los alineadores más utilizados, aunque algunos estudios han encontrado que BWA puede no ser preciso cuando la tasa de error en las lecturas es baja.[21]y que Novoalign funciona mejor cuando se usa junto con el Genome Analysis Toolkit (GATK)[22]. SOAP y sus versiones mejoradas tienen una alta precisión incluso cuando hay altas tasas de error en el mapeo, lo que los convierte en una buena opción para identificar polimorfismos de un solo nucleótido (SNP).[21]. Es importante evaluar cuidadosamente las opciones disponibles y elegir el alineador más apropiado para sus necesidades específicas.

A menudo se aplican pasos de preprocesamiento para preparar el archivo SAM para su posterior análisis. Estos pasos pueden incluir ordenar e indexar las lecturas mapeadas según su posición genómica utilizando

SAMtools.[23]y convertir el archivo SAM a una versión binaria (BAM) para una mejor gestión. Además, las estadísticas de alineación se pueden resumir utilizando Qualimap.[24]o más profundidad[25]y las secuencias duplicadas se pueden marcar o eliminar usando Picard[26]. Otros pasos, como realinear las lecturas en torno a inserciones y eliminaciones mediante la recalibración del nivel de calidad base GATK, también pueden mejorar la precisión y confiabilidad de la identificación de variantes.[27]. En resumen, el preprocesamiento de los datos de secuenciación es crucial para mejorar la precisión de la identificación de la variante final eliminando el sesgo y ajustando la calidad de las variantes identificadas.

2.3 llamada variante

Las herramientas de llamada de variantes SNP (polimorfismo de un solo nucleótido) e indel (inserción/deleción) se utilizan para identificar variaciones en las secuencias de ADN a nivel de un solo nucleótido. Estas herramientas se pueden dividir en dos categorías: métodos heurísticos y métodos probabilísticos. Métodos heurísticos, como VarScan2[28], utilizan múltiples fuentes de información sobre la calidad de los datos para asignar variantes y también pueden utilizar pruebas estadísticas, como la prueba de Fisher, para comparar las variantes con distribuciones teóricas.[12]. Los métodos probabilísticos, como SAMtools y GATK, se basan en enfoques bayesianos que optimizan la probabilidad de identificar genotipos. GATK es una herramienta de llamada de variantes probabilísticas ampliamente utilizada que se caracteriza por su confiabilidad y precisión, y también tiene un diseño modular que permite la detección de diferentes tipos de variantes y tiene funciones para filtrar y recalibrar resultados. Otras herramientas, como FreeBayes, pueden ser más precisas a la hora de detectar variantes de alta calidad, pero pueden detectar menos variantes en general.[29].

Los llamadores de la línea germinal se especializan en detectar variaciones que están presentes en cada célula de un individuo y se transmiten a su descendencia, mientras que los llamadores somáticos se utilizan para identificar variaciones que ocurren en un tejido o tipo de célula específico y no se transmiten a la descendencia. La llamada de variantes somáticas normalmente implica comparar los resultados de la secuenciación de dos muestras del mismo paciente, una muestra de tejido tumoral y una muestra de tejido normal, para distinguir las variantes somáticas de las variantes de la línea germinal. Existen varios algoritmos utilizados por diferentes herramientas de llamada de variantes somáticas, incluidos algoritmos heurísticos, análisis de genotipo conjunto, análisis de frecuencia alélica y estrategias basadas en haplotipos. Algoritmos heurísticos, como VarScan2[28], utilizan ciertos umbrales para detectar variantes y luego aplican pruebas estadísticas, como la prueba exacta de Fisher, para filtrar los resultados y obtener solo variantes somáticas. Análisis de genotipo conjunto, utilizado por herramientas como SomaticSniper[30]y SAMtools, asume diploidía en ambas muestras e intenta inferir los genotipos conjuntos utilizando el teorema de Bayes. Sin embargo, esta suposición puede no ser válida para muestras de tumores con subclones altamente heterogéneos. Análisis de frecuencia alélica, empleado por herramientas como Strelka[31]y MuTect[32], intenta alejarse del enfoque clásico de genotipos conjuntos considerando la frecuencia alélica de las variantes. Estrategias basadas en haplotipos, utilizadas por herramientas como Platypus[33], FreeBayes y MuTect2[21], implican ensamblar lecturas localmente en regiones específicas y generar genotipos basados en haplotipos.

Las herramientas de llamada de variantes estructurales se utilizan para identificar variaciones a mayor escala, como eliminaciones, inserciones, inversiones o translocaciones, que no pueden ser detectadas por SNP y las herramientas de llamada de variantes indel. Estas variaciones pueden tener impactos significativos en la función genética y pueden estar involucradas en el desarrollo de enfermedades. Algunas herramientas para llamar a variantes estructurales incluyen BreakDancer[34], CNVnador[35]y DELLY[36].

2.4 Anotación variante

La anotación de variantes es el proceso de asignar significado biológico a los resultados obtenidos de la llamada de variantes. Implica buscar información sobre variantes en varias bases de datos y recursos en línea, como dbSNP.[37]o el proyecto 1000 Genomas[38]y utilizando métricas como Condel[39], polifenol[40], o TAMIZAR[41]para evaluar el impacto clínico potencial de una variante. Estas métricas proporcionan una puntuación de predicción basada en la anotación de la variante y clasifican la variante según su posible impacto

clínico. Las variantes se pueden clasificar como patógenas, neutras, posiblemente benignas o variantes de significado incierto (VUS), según el nivel de confianza en su importancia clínica.[42]. Las variantes también se pueden definir según el efecto que tienen en la cadena de proteínas, como ser sinónimos o no sinónimos, o causar una mutación por cambio de marco. Las variantes no sinónimas dan como resultado un cambio en la secuencia de la proteína, lo que puede tener consecuencias funcionales, mientras que las variantes sinónimas no dan como resultado un cambio neto en la secuencia de la proteína debido a la degeneración del código genético. Las mutaciones de cambio de marco, que son causadas por la ganancia o pérdida de nucleótidos, pueden alterar la lectura normal de la secuencia de ADN y dar como resultado una secuencia de proteínas completamente diferente.

Hay varias herramientas disponibles para la anotación funcional, incluida ANNOVAR[43], NGS-SNP[44], snpEff[45]y PEV[46]. Estas herramientas se pueden utilizar en una interfaz de línea de comandos o mediante una interfaz gráfica, y ANNOVAR y VEP se encuentran entre las opciones más utilizadas y completas. Además de la anotación funcional, también es importante considerar otros factores como la frecuencia de una variante en la población, su presencia en genes conocidos asociados a enfermedades y la presencia de elementos funcionales relevantes en la región genómica de la variante. Recursos como el Proyecto Atlas del Genoma del Cáncer – TCGA -[47], CÓSMICO[48]y ClinVar[49]se puede utilizar para priorizar o filtrar variantes y evitar la necesidad de utilizar múltiples llamadores de variantes.

La anotación funcional puede ser un proceso complejo porque la predicción funcional de las variantes detectadas no siempre es sencilla. Si bien anteriormente se pensaba que tanto las variantes puntuales como las estructurales daban como resultado una proteína nociva que causaba un cambio en la secuencia de aminoácidos, estudios recientes han demostrado que este no es siempre el caso.[50]. Por lo tanto, es importante considerar una variedad de factores al evaluar la importancia clínica de una variante y determinar su impacto potencial en la función genética.

3 CANAL DE ANÁLISIS DE DATOS DE SECUENCIACIÓN DE NANOPOROS

La tecnología de secuenciación de lectura larga, como la secuenciación Single Molecule Real Time (SMART) y la tecnología Oxford Nanopore, ha ampliado las capacidades de la secuenciación masiva en entornos clínicos y tiene el potencial de convertirse en una práctica de rutina.[51]. Estas tecnologías se basan en secuenciación de lecturas largas y no requieren la amplificación de fragmentos mediante PCR, lo que las hace útiles para la detección de variantes estructurales y ensamblaje del genoma.[52]. Sin embargo, aún no han alcanzado la precisión de las tecnologías de secuenciación de lectura corta como Illumina. El análisis de datos de lectura larga requiere una canalización con varios pasos de procesamiento, que incluyen llamada de base, control de calidad, corrección de errores, detección de modificaciones, ensamblaje del genoma, análisis del transcriptoma, llamada de variantes y haplotipado/fase. En entornos clínicos, la secuenciación de Nanopore es la tecnología de lectura larga más utilizada y es particularmente útil para la detección de diferentes tipos de variantes de interés clínico.

Basecalling es el proceso de convertir mediciones de cambios de corriente sin procesar de cadenas de ADN o ARN que pasan a través de nanoporos en datos de secuencia. Es un paso crítico en el proceso de análisis de datos de nanoporos. Existen herramientas comerciales internas desarrolladas por Oxford Nanopore Technologies (ONT) y opciones de software de código abierto disponibles para llamadas base, utilizando diferentes métodos algorítmicos. El paquete de software más utilizado es Guppy, que forma parte del conjunto de herramientas proporcionadas por la ONT, que incluye algoritmos como Scrappie y Bonito y se utiliza junto con el servicio MinKNOW para gestionar el proceso de secuenciación y obtener lecturas secuenciadas.[53]. Otras herramientas populares de llamadas base de terceros incluyen Causalcall (que utiliza una red convolucional temporal)[54], DeepNano (utilizando una red neuronal recurrente)[55]y fast-bonito (una reimplementación de Bonito utilizando una técnica de búsqueda de arquitectura neuronal para acelerar la ejecución)[56].

El control de calidad y el preprocesamiento son pasos importantes en el análisis de datos de lectura larga, similar al análisis de datos de lectura corta. Estos pasos son necesarios para garantizar la precisión de los análisis posteriores y evaluar la calidad de los fragmentos generados. Hay varias herramientas disponibles para este propósito, como LongQC[57], Poretools[58], por[59], Nano OK[60], poro HPG[61], nanopaquete[62]y Filtlong[63]. Estas herramientas se pueden utilizar para evaluar la calidad y el estado general de los datos

generados a través de métricas, visualizaciones y análisis estadísticos, así como para realizar recortes para mejorar la calidad de una gran parte de los fragmentos.

El proceso de generación de lectura basado en nanoporos tiene una tasa de error inherente de alrededor del 15%, lo que puede provocar una cantidad significativa de bases secuenciadas incorrectamente. Esta tasa de error puede ser una limitación para la detección de cambios puntuales o ensamblaje del genoma de novo. Para solucionar este problema, se pueden utilizar métodos de corrección de errores. Estos métodos se pueden dividir en dos categorías: métodos de autocorrección, que utilizan gráficos para generar secuencias de consenso (como Canu[64]o LoRMA[sesenta y cinco]), y métodos híbridos, que utilizan lecturas cortas para corregir lecturas largas a través de alineaciones (como Nanocorr[66], Ratatosk[67]o FMLRC[68]). Se ha demostrado que los métodos híbridos reducen la tasa de error al 1-4%, que es similar a la tasa de error de lecturas cortas.[69].

Las lecturas largas tienen la ventaja de cubrir regiones genómicas más grandes, lo que hace que su posicionamiento sea más exclusivo en comparación con las lecturas cortas, que a menudo tienen múltiples alineamientos o mapeos secundarios posibles, especialmente en regiones repetitivas. Esto ha llevado al desarrollo de varias herramientas de alineación diseñadas específicamente para lecturas largas, como LAST[70], Mapa gráfico[71], minimapa2[72], NGLMR[73]y GraphMap2[74]. Entre estos, se ha demostrado que minimap2 con su algoritmo de alineación de cadena de semillas funciona bien en términos de precisión y rendimiento en conjuntos de llamadas, y también se puede utilizar para mapear lecturas cortas.[75]. Otra herramienta notable es GraphMap2, una reimplementación de GraphMap que puede mapear en modo de empalme para detectar con precisión los extremos de los exones.

El objetivo de un proceso de análisis de datos de secuenciación de nanoporos es detectar variantes de interés clínico, incluidas variantes de un solo nucleótido (SNV) y variantes estructurales (SV). Una de las principales aplicaciones de esta tecnología es la detección de grandes cambios genómicos estructurales, gracias a la capacidad de lecturas largas para abarcar estos cambios y detectar con precisión puntos de interrupción. Para ello, herramientas como NanoSV[76], Resoplidos[73], Quisquilloso[77], Nanovar[78], y se puede utilizar disgu[79]. Entre estos, se ha demostrado que Sniffles y Dysgu tienen el mejor rendimiento general.[75]. Para la detección de SNV, puede ser necesario un paso previo de corrección de errores y algoritmos de llamada de variantes conscientes de haplotipos. Se ha demostrado que el canal PEPPER-Margin-DeepVariant, que utiliza la herramienta DeepVariant de Google, detecta con precisión los SNV.[80].

4 APRENDIZAJE AUTOMÁTICO

Los algoritmos de aprendizaje automático se pueden utilizar en procesos de análisis de variantes genómicas para identificar variantes asociadas con enfermedades[81]. Estos algoritmos se dividen en dos categorías principales: aprendizaje supervisado y aprendizaje no supervisado. Los algoritmos de aprendizaje supervisado, como los métodos de aprendizaje profundo y las redes neuronales, se utilizan para predecir el valor de salida de una variable de entrada en función de un conjunto de datos de entrenamiento que ya ha sido etiquetado. Estos algoritmos se utilizan a menudo en llamadas y anotaciones de variantes porque ofrecen un mejor rendimiento que otros métodos, como clasificadores o técnicas de agrupación. Los algoritmos de aprendizaje no supervisados, como el aprendizaje de reglas de asociación o agrupamiento, se utilizan para descubrir patrones ocultos en un conjunto de datos sin etiquetas previas.

Se han desarrollado varias herramientas de aprendizaje automático para la detección de variantes, incluidas Scotch y Metal.[82], que obtienen nuevos indeles no detectados previamente, y DeepVariant[83], que es una herramienta versátil basada en una red neuronal convolucional profunda que puede generalizarse a través de diferentes construcciones genómicas, plataformas y diseños experimentales. También se han realizado esfuerzos para desarrollar herramientas de aprendizaje automático para la detección de variantes somáticas en el cáncer, como NeuSomatic.[84], el primer algoritmo basado en una red neuronal convolucional para la detección precisa de variantes somáticas, y DeepSVR[85], un modelo de aprendizaje profundo que incluye tres algoritmos diferentes (regresión logística, bosque aleatorio y aprendizaje profundo) que se utilizan para refinar un conjunto de variantes ya detectadas y obtener un buen conjunto de variantes somáticas verdaderas.

Además de estas herramientas generales de aprendizaje automático, también existen herramientas específicas para plataformas de secuenciación de lectura larga como Oxford Nanopore y Pacific Biosciences, como DeepNano, una red neuronal recurrente para lecturas del secuenciador MinION, y Clairvoyante.[86], una

red neuronal profunda convolucional diseñada específicamente para la plataforma de secuenciación SMRT, aunque también es válida para otras plataformas.

También existen algoritmos que se centran en mejorar el proceso de anotación funcional de variantes detectadas, que son clave para el análisis de variantes somáticas desconocidas relacionadas con el cáncer o para mutaciones en regiones no codificantes. Ejemplos de estos algoritmos incluyen DeepSEA[87], un algoritmo basado en aprendizaje profundo que identifica características funcionales de variantes no codificantes a través de información de secuencias reguladoras y perfiles de cromatina, BadMut[88], un metaestimador que utiliza algoritmos de aprendizaje profundo para integrar varias puntuaciones de predicción deletéreas y predecir el potencial patógeno de una variante, y DeepGene[89], software que es un clasificador basado en aprendizaje profundo que predice el posible efecto funcional de las mutaciones sin sentido en la estructura y función de las proteínas.

5 SISTEMAS DE GESTIÓN DEL FLUJO DE TRABAJO.

El análisis de datos de secuenciación de próxima generación (NGS) implica numerosas etapas que a menudo se integran en un proceso. Estos canales y sistemas de gestión de flujo de trabajo se han desarrollado para facilitar el análisis de datos NGS para investigadores con poca experiencia en TI y para estandarizar y aumentar la reproducibilidad en biología computacional.

Estas herramientas suelen incluir una interfaz gráfica y ofrecen pasos y procesos predefinidos, pero pueden carecer de flexibilidad para modificar o reemplazar ciertos módulos. Los sistemas de gestión de flujos de trabajo, por otro lado, ofrecen mayor apertura y flexibilidad, funciones avanzadas y la capacidad de visualizar procesos en tiempo real y trabajar en la nube. Ejemplos de sistemas de gestión de flujo de trabajo incluyen Galaxy, Taverna[90]y serpiente[91]. Estas plataformas permiten a los investigadores personalizar sus procesos de análisis y elegir las herramientas adecuadas para cada paso del proceso.

Nextflow es un sistema de gestión de flujo de trabajo que permite la automatización y paralelización de canales de análisis de datos.[92]. Algunas de sus características principales incluyen:

- Soporte para diversos lenguajes de programación y entornos de ejecución, como Bash, Python, R, etc.
- Capacidad de adaptación a diferentes sistemas de almacenamiento y plataformas informáticas, como clusters informáticos, plataformas en la nube y sistemas de contenedores.
- Gestión eficiente de dependencias y recursos, para evitar tareas redundantes y optimizar el uso de los recursos disponibles.
- Monitoreo y seguimiento en tiempo real de la ejecución del flujo de trabajo mediante la visualización de gráficos y estadísticas.
- Integración con repositorios de código y sistemas de control de versiones, como Git y GitHub, para facilitar el trabajo colaborativo y la reproducibilidad de los análisis.

Además de estas herramientas generales, también hay herramientas y procesos más especializados disponibles para aplicaciones NGS específicas, como ChIP-seq y RNA-seq, y para propósitos específicos, como control de calidad y alineación. Estos canales y herramientas especializados se pueden utilizar junto con marcos de canales generales para abordar las necesidades específicas de un proyecto determinado.

El desarrollo de estas herramientas ha contribuido a la evolución y optimización de las plataformas existentes y al desarrollo de nuevos algoritmos que pueden abordar problemas complejos en el análisis de datos NGS. Sin embargo, elegir la herramienta adecuada para cada paso puede ser un proceso complejo y que requiere mucho tiempo, especialmente con la gran cantidad de herramientas y algoritmos sofisticados disponibles. Es importante que los investigadores consideren cuidadosamente sus necesidades específicas y elijan la herramienta o canal adecuado para garantizar el éxito de su análisis.

6 PROFUNDONGS

DeepNGS (deepngs.eu) es una plataforma para el análisis e interpretación rápidos y automatizados de muestras de ADN secuenciadas humanas a partir de experimentos de secuenciación masiva. Puede usarse en entornos de computación en la nube, como Amazon Web Services, o instalarse localmente en las

instalaciones del usuario con opciones de configuración personalizadas. Tiene un flujo de trabajo optimizado y utiliza scripts avanzados para una optimización continua, así como técnicas de visualización y gráficos interactivos para ayudar a los usuarios a interpretar y presentar los resultados. Pretende obtener un conjunto de variantes genéticas que sean relevantes para el diagnóstico clínico de los pacientes e incorpora algoritmos de aprendizaje automático para optimizar el proceso de análisis y mejorar la precisión. La plataforma utiliza varios métodos computacionales y criterios ACMG para predecir el posible efecto nocivo de variantes genéticas específicas sobre las proteínas.

La plataforma utiliza AWS Batch, un servicio que le permite ejecutar cargas de trabajo informáticas por lotes en la nube de AWS. Administra los recursos informáticos subyacentes, incluidas las instancias de Amazon Elastic Compute Cloud (EC2) y los contenedores Docker, y los escala automáticamente según la carga de trabajo. Esto facilita la ejecución de canalizaciones de Nextflow a escala, utilizando una variedad de herramientas y bibliotecas para el procesamiento de datos NGS.

La arquitectura de la nube se describe a continuación (Figura 1):

- Almacenamiento de datos: los datos NGS sin procesar y el genoma de referencia se almacenan en un depósito de Amazon Simple Storage Service (S3). El depósito de S3 sirve como una ubicación central para almacenar y acceder a los datos y se puede acceder a él desde los recursos informáticos utilizados para ejecutar la canalización.
- Definición de canalización: laLa canalización de Nextflow se define en un script de Nextflow y se almacena en un sistema de control de versiones como Git. El script especifica las tareas que componen la canalización, así como las dependencias entre ellas y los datos de entrada y salida de cada tarea.
- Ejecución de canalización: la canalización se ejecuta mediante AWS Batch, que es un servicio que le permite ejecutar cargas de trabajo informáticas por lotes en la nube de AWS. Para ejecutar la canalización, la plataforma envía un trabajo a AWS Batch, que lanza los recursos informáticos necesarios y ejecuta las tareas en elGuión de flujo siguiente.
- Tareas de canalización: las tareas en elLa canalización de Nextflow se ejecuta en los recursos informáticos administrados por AWS Batch. Estas tareas pueden incluir pasos como alineación de lectura, llamada de variantes y anotaciones. Cada tarea se implementa como un proceso en el script de Nextflow y se puede escribir en una variedad de lenguajes, como Bash, Python o R.
- Procesamiento de datos: las tareas de canalización pueden utilizar una variedad de herramientas y bibliotecas para el procesamiento de datos NGS, como BWA,SAMtools y GATK. Estas herramientas y bibliotecas se instalan en los recursos informáticos mediante contenedores acoplables y Amazon ECS.
- Almacenamiento de salida: los resultados de las tareas de canalización, como archivos de alineación y llamadas de variantes, se almacenan en otro depósito de Amazon S3. Las tareas de canalización pueden acceder a este depósito según sea necesario, y elLos resultados finales se pueden descargar o analizar utilizando otras herramientas.

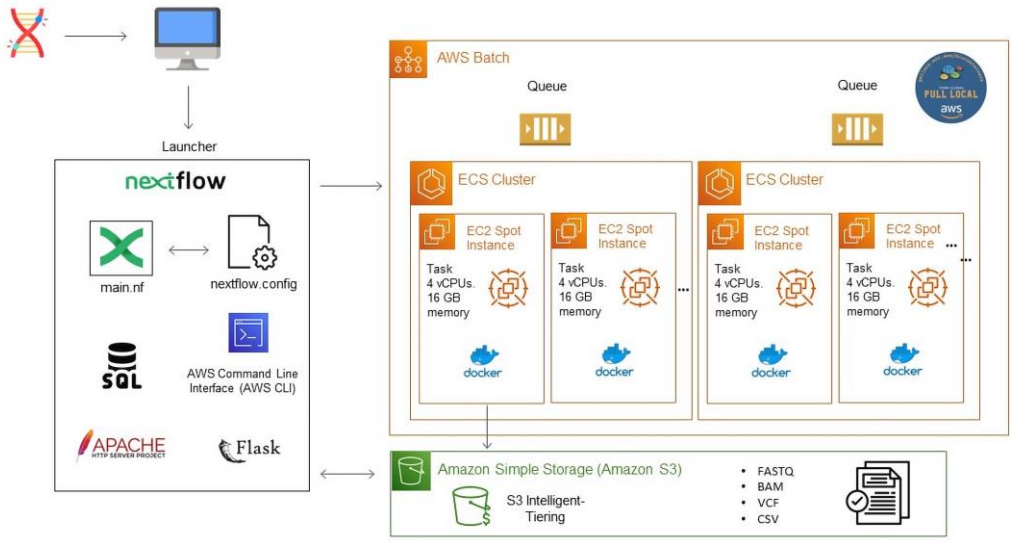

Figura 1: Arquitectura de la plataforma en la nube DeepNGS.

7 CONCLUSIONES

Hay muchas herramientas y plataformas disponibles para analizar datos de proyectos de secuenciación masiva, pero esto puede dificultar el logro de la estandarización y la reproducibilidad en el campo de la genómica computacional. Para abordar este problema, se podrían desarrollar directrices y canales para diferentes aplicaciones para mejorar la transparencia y la reproducibilidad. DeepNGS es una plataforma optimizada para que la utilicen médicos y proporciona anotaciones biológicas e informes gráficos intuitivos para ayudar en la interpretación de los resultados. También incorpora algoritmos de aprendizaje automático para mejorar los resultados del análisis. En el futuro, está previsto lanzar una versión premium de DeepNGS, que incluirá funciones adicionales como acceso a bases de datos específicas de cáncer, la capacidad de analizar muestras genómicas complejas y la capacidad de analizar casos familiares. También está prevista una versión de la plataforma diseñada para servidores locales on premise, con soporte técnico continuo y un sistema de gestión del flujo de trabajo. Esta versión será personalizable a nivel de instalación según los requerimientos computacionales de cada usuario y las características de sus servidores. En general, el objetivo de DeepNGS es proporcionar una plataforma sólida y optimizada para analizar e interpretar datos genómicos, con el objetivo de mejorar la precisión y la exhaustividad de los resultados.

AGRADECIMIENTOS

El presente estudio ha sido financiado por el proyecto AIR Genomics (con número de expediente CCTT3/20/SA/0003), mediante la convocatoria 2020 PROYECTOS I+D ORIENTADOS A LA EXCELENCIA Y MEJORA COMPETITIVA DE LOS CCTT por el Instituto de Competitividad Empresarial de Castilla y León y fondos FEDER.

REFERENCIAS

[1] "Explicación de los archivos FASTQ". https://emea.support.illumina.com/bulletins/2016/04/fastq-files-explained.html (consultado el 19 de septiembre de 2022).
[2] 'Babraham Bioinformatics - FastQC Una herramienta de control de calidad para datos de secuencia de alto rendimiento'. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (consultado el 14 de abril de 2021).
[3] AM Bolger, M. Lohse y B.Usadel, 'Trimmomatic: un recortador flexible para datos de secuencia de Illumina', Bioinformática, vol. 30, núm. 15, págs. 2114–2120, agosto de 2014, doi: 10.1093/bioinformatics/btu170.
[4] RK Patel y M. Jain, 'NGS QC Toolkit: Un conjunto de herramientas para el control de calidad de datos de secuenciación de próxima generación',Más uno, vol. 7, núm. 2, pág. e30619, febrero de 2012, doi: 10.1371/JOURNAL.PONE.0030619.
[5] r.Schmieder y R. Edwards, 'Control de calidad y preprocesamiento de conjuntos de datos metagenómicos', Bioinformática, vol. 27, núm. 6, págs. 863–864, marzo de 2011, doi: 10.1093/BIOINFORMATICS/BTR026.
[6] J.Goecks et al., 'Galaxy: un enfoque integral para apoyar la investigación computacional accesible, reproducible y transparente en las ciencias biológicas', Genome Biol, vol. 11, núm. 8, agosto de 2010, doi: 10.1186/gb-2010-11-8-r86.
[7] C. Hong, S.Manimaran y WE Johnson, 'PathoQC: preprocesamiento de lectura computacionalmente eficiente y control de calidad para conjuntos de datos de secuenciación de alto rendimiento', Cancer Inform, vol. 13s1, pág. CIN.S13890, enero de 2014, doi: 10.4137/CIN.S13890.
[8] S. Chen, T. Huang, Y. Zhou, Y. Han, M. Xu y J. Gu, 'AfterQC: filtrado, recorte, eliminación de errores y control de calidad automáticos para datos fastq', BMC Bioinformatics, vol. 18, núm. 3, págs. 91–100, marzo de 2017, doi: 10.1186/S12859-017-1469-3/FIGURES/9.
[9] S. Chen, Y. Zhou, Y. Chen y J. Gu, 'fastp: un preprocesador FASTQ ultrarrápido todo en uno', Bioinformática, vol. 34, núm. 17, págs. i884–i890, septiembre de 2018, doi: 10.1093/BIOINFORMATICS/BTY560.
[10] X. Liu, Z. Yan, C. Wu, Y. Yang, X. Li y G. Zhang, 'FastProNGS: preprocesamiento rápido de lecturas de secuenciación de próxima generación', BMC Bioinformatics, vol. 20, núm. 1, pág. 345, junio de 2019, doi: 10.1186/s12859-019-2936-9.
[11] SKChanumolu, M. Albahrani y HH Otu, 'FQStat: una arquitectura paralela para la evaluación de muy alta velocidad de métricas de calidad de secuenciación', BMC Bioinformatics, vol. 20, núm. 1, pág. 424, agosto de 2019, doi: 10.1186/s12859-019-3015-y.
[12] METRO.Mielczarek y J. Szyda, 'Revisión de algoritmos de alineación y llamada de SNP para datos de secuenciación de próxima generación', Journal of Applied Genetics, vol. 57, núm. 1. Springer Verlag, págs. 71–79, 1 de febrero de 2016. doi: 10.1007/s13353-015-0292-7.
[13] AD Smith et al., 'Actualizaciones del software de mapeo de lectura corta RMAP', Bioinformatics, vol. 25, núm. 21, págs. 2841–2842, noviembre de 2009,doi: 10.1093/BIOINFORMÁTICA/BTP533.
[14] R. Li, Y. Li, K. Kristiansen y J. Wang, 'SOAP: programa corto de alineación de oligonucleótidos', Bioinformática, vol. 24, núm. 5, págs. 713–714, marzo de 2008,doi: 10.1093/BIOINFORMÁTICA/BTN025.
[15] 'NovoAlign | Novocraft'. http://www.novocraft.com/products/novoalign/ (consultado el 15 de abril de 2021).
[dieciséis] SM Rumble, P.Lacroute, A. v. Dalca, M. Fiume, A. Sidow y M. Brudno, 'SHRiMP: mapeo preciso de lecturas cortas de espacio de color', PLoS Comput Biol, vol. 5, núm. 5, pág. e1000386, mayo de 2009, doi: 10.1371/journal.pcbi.1000386.
[17] N. Homer, B. Merriman y SF Nelson, 'BFAST: una herramienta de alineación para la resecuenciación del genoma a gran escala',Más uno, vol. 4, núm. 11, pág. e7767, noviembre de 2009, doi: 10.1371/JOURNAL.PONE.0007767.
[18] H. Li y R. Durbin, 'Alineación de lectura corta rápida y precisa con la transformada de Burrows-Wheeler', Bioinformática, vol. 25, núm. 14, págs. 1754-1760, julio de 2009,doi: 10.1093/bioinformática/btp324.
[19] B. Langmead y SLSalzberg, 'Alineación rápida de lectura separada con Bowtie 2', Nat Methods, vol. 9, núm. 4, págs. 357–359, abril de 2012, doi: 10.1038/nmeth.1923.
[20] R. Li et al., 'SOAP2: una herramienta ultrarrápida mejorada para la alineación de lecturas cortas', Bioinformatics, vol. 25, núm. 15, págs. 1966-1967, agosto de 2009,doi: 10.1093/BIOINFORMÁTICA/BTP336.
[21] M. Ruffalo, T. Laframboise y M.Koyutürk, 'Análisis comparativo de algoritmos para la alineación de lectura de secuenciación de próxima generación', Bioinformática, vol. 27, núm. 20, págs. 2790–2796, octubre de 2011, doi: 10.1093/bioinformatics/btr477.
[22] A. Cornualles y C.Guda, 'Una comparación de tuberías de llamadas variantes utilizando el genoma en una botella como referencia', Biomed Res Int, vol. 2015, 2015, doi: 10.1155/2015/456479.
[23] PAG.Danecek et al., 'Doce años de SAMtools y BCFtools', Gigascience, vol. 10, núm. 2 de febrero de 2021, doi: 10.1093/gigascience/giab008.
[24] F. García-Alcalde et al., 'Qualimap: evaluación de datos de alineación de secuenciación de próxima generación', Bioinformática, vol. 28, núm. 20, págs. 2678–2679, octubre de 2012, doi: 10.1093/BIOINFORMATICS/BTS503.
[25] BS Pedersen y AR Quinlan, 'Mos Depth: cálculo rápido de cobertura para genomas y exomas', Bioinformática, vol. 34, núm. 5, págs. 867–868, marzo de 2018, doi: 10.1093/BIOINFORMATICS/BTX699.
[26] 'GitHub -broadinstitute/picard: un conjunto de herramientas de línea de comandos (en Java) para manipular datos y formatos de secuenciación de alto rendimiento (HTS) como SAM/BAM/CRAM y VCF. https://github.com/broadinstitute/picard (consultado el 19 de abril de 2022).
[27] GA van derAuwera et al., 'De los datos fastQ a las llamadas de variantes de alta confianza: el canal de mejores prácticas del kit de herramientas de análisis del genoma', Curr Protoc Bioinformatics, no. SUPL.43, 2013, doi: 10.1002/0471250953.bi1110s43.
[28] corriente continuaKoboldt et al., 'VarScan 2: descubrimiento de mutación somática y alteración del número de copias en cáncer mediante secuenciación del exoma', Genome Res. vol. 22, núm. 3, págs. 568–576, marzo de 2012, doi: 10.1101/gr.129684.111.
[29] S. Hwang, E. Kim, I. Lee y EM Marcotte, 'Comparación sistemática de canales de llamadas de variantes utilizando variantes de exoma personal estándar de oro', Sci Rep, vol. 5 de diciembre de 2015,doi: 10.1038/srep17875.
[30] DE Larson et al., 'Somaticsniper: Identificación de mutaciones puntuales somáticas en datos de secuenciación del genoma completo, Bioinformática, vol. 28, núm. 3, págs. 311–317, febrero de 2012, doi: 10.1093/bioinformatics/btr665.
[31] CT Saunders, WSW Wong, S. Swamy, J.Becq, LJ Murray y RK Cheetham, 'Strelka: llamada precisa de variantes pequeñas somáticas a partir de pares secuenciados de muestras normales y de tumor', Bioinformatics, vol. 28, núm. 14, págs. 1811–1817, julio de 2012, doi: 10.1093/bioinformatics/bts271.
[32] K.Cibulskis et al., 'Detección sensible de mutaciones puntuales somáticas en muestras de cáncer impuras y heterogéneas', Nat Biotechnol, vol. 31, núm. 3, págs. 213–219, marzo de 2013, doi: 10.1038/nbt.2514.
[33] A.Rimmer et al., 'Integración de enfoques basados en mapeo, ensamblaje y haplotipos para llamar a variantes en aplicaciones de secuenciación clínica', Nat Genet, vol. 46, núm. 8, págs. 912–918, julio de 2014, doi: 10.1038/ng.3036.

[34] K. Chen y otros, 'BreakDancer: un algoritmo para el mapeo de alta resolución de la variación estructural genómica', Nat Methods, vol. 6, núm. 9, págs. 677–681, agosto de 2009, doi: 10.1038/nmeth.1363.
[35] A.Abyzov, AE Urban, M. Snyder y M. Gerstein, 'CNVnator: un enfoque para descubrir, genotipar y caracterizar CNV típicas y atípicas a partir de la secuenciación del genoma familiar y poblacional', Genome Res. vol. 21, núm. 6, págs. 974–984, junio de 2011, doi: 10.1101/gr.114876.110.
[36] T. Rausch, T.Zichner, A. Schlattl, AM Stütz, V. Benes y JO Korbel, 'DELLY: descubrimiento de variantes estructurales mediante análisis integrado de lectura dividida y de pares pares', Bioinformática. vol. 28, núm. 18, págs. 333–339, septiembre de 2012, doi: 10.1093/bioinformatics/bts378.
[37] ST Sherry, M. Ward y K.Sirotkin, 'dbSNP: base de datos para polimorfismos de nucleótidos únicos y otras clases de variación genética menor', Genome Res, vol. 9, núm. 8, págs. 677–679, agosto de 1999, doi: 10.1101/GR.9.8.677.
[38] A.Auton et al., 'Una referencia global para la variación genética humana', Nature, vol. 526, núm. 7571. Nature Publishing Group, págs. 68–74, 30 de septiembre de 2015. doi: 10.1038/nature15393.
[39] A. González-Pérez y N. López-Bigas, 'Mejora de la evaluación del resultado de SNV no sinónimos con una puntuación de nocividad por consenso,Condel', Am J Hum Genet, vol. 88, núm. 4, págs. 440–449, abril de 2011, doi: 10.1016/j.ajhg.2011.03.004.
[40] I.Adzhubei, DM Jordan y SR Sunyaev, Predicción del efecto funcional de mutaciones sin sentido humanas utilizando PolyPhen-2, no. SUPL.76. 2013. doi: 10.1002/0471142905.hg0720s76.
[41] P. Kumar, S.Henikoff y PC Ng, 'Predicción de los efectos de la codificación de variantes no sinónimos en la función de las proteínas mediante el algoritmo SIFT', Nat Protoc, vol. 4, núm. 7, págs. 1073–1082, 2009, doi: 10.1038/nprot.2009.86.
[42] S. Roy et al., 'Estándares y directrices para validar tuberías bioinformáticas de secuenciación de próxima generación: una recomendación conjunta de la Asociación de Patología Molecular y el Colegio de Patólogos Estadounidenses', Journal of Molecular Diagnostics, vol. 20, núm. 1. Elsevier BV, págs. 4 a 27, 1 de enero de 2018.doi: 10.1016/j.jmoldx.2017.11.003.
[43] K. Wang, M. Li y H.Hakonarson, 'ANNOVAR: anotación funcional de variantes genéticas a partir de datos de secuenciación de alto rendimiento', Nucleic Acids Res, vol. 38, núm. 16, págs. e164–e164, septiembre de 2010, doi: 10.1093/NAR/GKQ603.
[44] JR Grant, ASArantes, X. Liao y P. Stothard, 'Anotación en profundidad de SNP que surgen de proyectos de resecuenciación utilizando NGS-SNP', Bioinformatics, vol. 27, núm. 16, págs. 2300–2301, agosto de 2011, doi: 10.1093/bioinformatics/btr372.
[45] PAG.Cingolani et al., 'Un programa para anotar y predecir los efectos de polimorfismos de un solo nucleótido, SnpEff: SNP en el genoma de la cepa w1118 de Drosophila melanogaster; iso-2; iso-3', Fly (Austin), vol. 6, núm. 2, págs. 80–92, 2012, doi: 10.4161/fly.19695.
[46] W. McLaren y otros, 'TheEnsembl Variant Effect Predictor', Genome Biol, vol. 17, núm. 1, págs. 1 a 14, 2016, doi: 10.1186/s13059-016-0974-4.
[47] 'Programa Atlas del Genoma del Cáncer - Instituto Nacional del Cáncer'. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (consultado el 22 de abril de 2022).
[48] JG Tate et al., 'COSMIC: el catálogoDe mutaciones somáticas en el cáncer', Nucleic Acids Res, vol. 47, núm. D1, págs. D941–D947, enero de 2019, doi: 10.1093/NAR/GKY1015.
[49] MJ Landrum et al., 'ClinVar: mejorar el acceso a interpretaciones variantes y evidencia de respaldo', Nucleic Acids Res, vol. 46, núm. D1, págs. D1062–D1067, enero de 2018, doi: 10.1093/NAR/GKX1153.
[50] B. Mair et al., 'Las mutaciones de ganancia y pérdida de función en el gen GATA3 del cáncer de mama dan como resultado una sensibilidad diferencial a los fármacos',PLoS Genet, vol. 12, núm. 9, pág. e1006279, septiembre de 2016, doi: 10.1371/JOURNAL.PGEN.1006279.
[51] Dakota del SurGoenka et al., 'Identificación acelerada de variantes causantes de enfermedades con secuenciación ultrarrápida del genoma de nanoporos', Nature Biotechnology 2022, págs. 1–7, marzo de 2022, doi: 10.1038/s41587-022-01221-5.
[52] J.Weischenfeldt, O. Symmons, F. Spitz y JO Korbel, 'Impacto fenotípico de la variación estructural genómica: conocimientos desde y para las enfermedades humanas', Nature Reviews Genetics 2013 14:2, vol. 14, núm. 2, págs. 125-138, enero de 2013, doi: 10.1038/nrg3373.
[53] Y. Wang, Y. Zhao, A.Bollas, Y. Wang y KF Au, 'Tecnología de secuenciación de nanoporos, bioinformática y aplicaciones', Nature Biotechnology 2021, págs. 1–18, noviembre de 2021, doi: 10.1038/s41587-021-01108-x.
[54] J. Zeng, H. Cai, H. Peng, H. Wang, Y. Zhang y T.Akutsu, 'Causalcall: llamada base de nanoporos utilizando una red convolucional temporal', Front Genet, vol. 10, pág. 1332, enero de 2020, doi: 10.3389/FGENE.2019.01332/BIBTEX.
[55] v.Bože, B. Brejová y T. Vinař, 'DeepNano: redes neuronales recurrentes profundas para llamadas de bases en lecturas MinION Nanopore', PLoS One, vol. 12, núm. 6, pág. e0178751, junio de 2017, doi: 10.1371/journal.pone.0178751.
[56] Z. Xu y otros, 'Fast-bonito: un llamador de base basado en aprendizaje profundo más rápido para la secuenciación de nanoporos', Inteligencia artificial en las ciencias biológicas, vol. 1, pág. 100011, diciembre de 2021, doi: 10.1016/J.AILSCI.2021.100011.
[57] y.Fukasawa, L. Ermini, H. Wang, K. Carty y MS Cheung, 'LongQC: una herramienta de control de calidad para datos de lectura larga de secuenciación de tercera generación', Genes G3|Genomas|Genética, vol. 10, núm. 4, págs. 1193–1196, abril de 2020, doi: 10.1534/G3.119.400864.
[58] NJ Loman y AR Quinlan, 'Poretools: un conjunto de herramientas para analizar datos de secuencia de nanoporos', Bioinformatics, vol. 30, núm. 23, págs. 3399–3401, diciembre de 2014, doi: 10.1093/BIOINFORMATICS/BTU555.
[59] M. Watson y otros, 'poRe: un paquete R para la visualización y análisis de datos de secuenciación de nanoporos', Bioinformatics, vol. 31, núm. 1, págs. 114-115, enero de 2015, doi: 10.1093/BIOINFORMATICS/BTU590.
[60] RM Leggett, D. Cielos, M.Caccamo, MD Clark y RP Davey, 'NanoOK: análisis de alineación de referencias múltiples de datos de secuenciación de nanoporos, perfiles de calidad y error', Bioinformatics. vol. 32, núm. 1, págs. 142-144, enero de 2016, doi: 10.1093/BIOINFORMATICS/BTV540.
[61] J.Tarraga, A. Gallego, V. Arnau, I. Medina y J. Dopazo, 'Poro HPG: un marco eficiente y escalable para datos de secuenciación de nanoporos', BMC Bioinformatics. vol. 17, núm. 1, págs. 1 a 7, febrero de 2016, doi: 10.1186/S12859-016-0966-0/FIGURES/3.
[62] W. deCoster, S. D'Hert, DT Schultz, M. Cruts y C. van Broeckhoven, 'NanoPack: visualización y procesamiento de datos de secuenciación de lectura larga', Bioinformática, vol. 34, núm. 15, págs. 2666–2669, agosto de 2018, doi: 10.1093/BIOINFORMATICS/BTY149.
[63] 'GitHub -rrwick/Filtlong: herramienta de filtrado de calidad para lecturas largas'. https://github.com/rrwick/Filtlong (consultado el 25 de abril de 2022).

[64] S.Koren, BP Walenz, K. Berlin, JR Miller, NH Bergman y AM Phillippy, 'Canu: ensamblaje de lectura larga escalable y preciso mediante ponderación adaptativa de k-mer y separación repetida', Genome Res. vol. 27, núm. 5, págs. 722–736, mayo de 2017, doi: 10.1101/GR.215087.116.

[sesenta y cinco] I.Salmela, R. Walve, E. Rivals, E. Ukkonen y C. Sahinalp, 'Autocorrección precisa de errores en lecturas largas utilizando gráficos de Bruijn', Bioinformática, vol. 33, núm. 6, págs. 799–806, marzo de 2017, doi: 10.1093/BIOINFORMATICS/BTW321.

[66] S. Goodwin, J.Gurtowski, S. Ethe-Sayers, P. Deshpande, MC Schatz y WR McCombie, 'Secuenciación de nanoporos de Oxford, corrección de errores híbridos y ensamblaje de novo de un genoma eucariótico', Genome Res. vol. 25, núm. 11, págs. 1750–1756, noviembre de 2015, doi: 10.1101/GR.191395.115.

[67] G. Holley y otros, 'Ratatosk: la corrección de errores híbrida de lecturas largas permite llamar y ensamblar variantes con precisión', Genome Biol, vol. 22, núm. 1, pág. 28 de diciembre de 2021, doi: 10.1186/s13059-020-02244-4.

[68] JR Wang, J. Holt, L. McMillan y CD Jones, 'FMLRC: corrección de errores de lectura larga híbrida utilizando un índice FM', BMC Bioinformatics, vol. 19, núm. 1, págs. 1 a 11, febrero de 2018,doi: 10.1186/S12859-018-2051-3/TABLAS/4.

[69] S. Fu, A. Wang y KF Au, 'Una evaluación comparativa de métodos híbridos de corrección de errores para lecturas largas propensas a errores', GenomeBiol, vol. 20, núm. 1, págs. 1 a 17, febrero de 2019, doi: 10.1186/S13059-018-1605-Z/FIGURES/6.

[70] SMKiełbasa, R. Wan, K. Sato, P. Horton y MC Frith, 'Comparación de secuencia genómica domesticada de semillas adaptativas', Genome Res. vol. 21, núm. 3, págs. 487–493, marzo de 2011, doi: 10.1101/GR.113985.110.

[71] I.Sović, M. Šikić, A. Wilm, SN Fenlon, S. Chen y N. Nagarajan, 'Mapeo rápido y sensible de lecturas de secuenciación de nanoporos con GraphMap', Nat Commun, vol. 7 de abril de 2016, doi: 10.1038/ncomms11307.

[72] H. Li, 'Minimap2: alineación por pares para secuencias de nucleótidos', Bioinformática, vol. 34, núm. 18, págs. 3094–3100, septiembre de 2018,doi: 10.1093/BIOINFORMÁTICA/BTY191.

[73] FJSedlazeck et al., 'Detección precisa de variaciones estructurales complejas mediante secuenciación de una sola molécula', Nature Methods 2018 15:6, vol. 15, núm. 6, págs. 461–468, abril de 2018, doi: 10.1038/s41592-018-0001-7.

[74] J.Marić, I. Sović, K. Križanović, N. Nagarajan y M. Šikić, 'Graphmap2: mapeador de RNA-seq con reconocimiento de empalme para lecturas largas', bioRxiv, p. 720458, julio de 2019, doi: 10.1101/720458.

[75] A. Zhou, T. Lin y J. Xing, 'Evaluación de tuberías de procesamiento de datos de secuenciación de nanoporos para la identificación de variaciones estructurales', GenomaBiol, vol. 20, núm. 1 de noviembre de 2019, doi: 10.1186/s13059-019-1858-1.

[76] METRO.Cretu Stancu et al., 'Mapeo y fase de variación estructural en genomas de pacientes mediante secuenciación de nanoporos', Nature Communications 2017 8:1, vol. 8, núm. 1, págs. 1 a 13, noviembre de 2017, doi: 10.1038/s41467-017-01343-4.

[77] L. Gong et al., 'Picky detecta exhaustivamente variantes estructurales de alta resolución en lecturas largas de nanoporos', Nature Methods 2018 15:6, vol. 15, núm. 6, págs. 455–460, abril de 2018,doi: 10.1038/s41592-018-0002-6.

[78] CYTham et al., 'NanoVar: caracterización precisa de las variantes estructurales genómicas de los pacientes mediante secuenciación de nanoporos de baja profundidad', Genome Biol, vol. 21, núm. 1, págs. 1–15, marzo de 2020, doi: 10.1186/S13059-020-01968-7/FIGURES/3.

[79] K.Cleal y DM Baird, 'Dysgu: llamada de variante estructural eficiente mediante lecturas cortas o largas', Nucleic Acids Res, enero de 2022, doi: 10.1093/NAR/GKAC039.

[80] K.Shafin et al., 'La llamada de variantes con reconocimiento de haplotipos con PEPPER-Margin-DeepVariant permite una alta precisión en lecturas largas de nanoporos', Nature Methods 2021, págs. 1–11, noviembre de 2021, doi: 10.1038/s41592-021-01299-w.

[81] J. Zou, M. Huss, A. Abid, P. Mohammadi, A.Torkamani y A. Telenti, 'Introducción al aprendizaje profundo en genómica', Nat Genet, vol. 51, núm. 1, págs. 12 a 18, enero de 2019, doi: 10.1038/s41588-018-0295-5.

[82] C.Curnin et al., 'Detección de inserciones y eliminaciones en el genoma humano basada en aprendizaje automático', bioRxiv, p. 628222, diciembre de 2019, doi: 10.1101/628222.

[83] R. Poplin et al., 'Un universalLlamador de variantes snp y de indel pequeño utilizando redes neuronales profundas', Nat Biotechnol, vol. 36, núm. 10, pág. 983, noviembre de 2018, doi: 10.1038/nbt.4235.

[84] PYMESahraeian, R. Liu, B. Lau, K. Podesta, M. Mohiyuddin y HYK Lam, 'Redes neuronales convolucionales profundas para una detección precisa de mutaciones somáticas', Nat Commun, vol. 10, núm. 1 de diciembre de 2019, doi: 10.1038/s41467-019-09027-x.

[85] BJ Ainscough et al., 'Un enfoque de aprendizaje profundo para automatizar el refinamiento de la llamada de variantes somáticas a partir de datos de secuenciación del cáncer', Nat Genet, vol. 50, núm. 12, págs. 1735-1743, diciembre de 2018,doi: 10.1038/s41588-018-0257-y.

[86] R. Luo, FJSedlazeck, TW Lam y MC Schatz, 'Una red neuronal profunda convolucional multitarea para llamadas de variantes en secuenciación de una sola molécula', Nat Commun, vol. 10, núm. 1 de diciembre de 2019, doi: 10.1038/s41467-019-09025-z.

[87] J. Zhou y OGTroyanskaya, 'Predicción de los efectos de variantes no codificantes con un modelo de secuencia basado en aprendizaje profundo', Nat Methods, vol. 12, núm. 10, págs. 931–934, septiembre de 2015, doi: 10.1038/nmeth.3547.

[88] I.Korvigo, A. Afanasyev, N. Romashchenko y M. Skoblov, 'Generalizar mejor: aplicar el aprendizaje profundo para integrar puntuaciones de predicción deletéreas para estudios SNV de exoma completo', PLoS One, vol. 13, núm. 3 de marzo de 2018, doi: 10.1371/journal.pone.0192829.

[89] Y. Yuan et al., 'Deepgene: un clasificador avanzado de tipos de cáncer basado en aprendizaje profundo y mutaciones puntuales somáticas, BMC Bioinformatics, vol. 17, núm. Suplemento 17, diciembre de 2016, doi: 10.1186/s12859-016-1334-9.

[90] K.Wolstencroft et al., 'La suite de flujo de trabajo Taverna: diseño y ejecución de flujos de trabajo de servicios web en el escritorio, la web o la nube', Nucleic Acids Res, vol. 41, núm. Edición del servidor web, 2013, doi: 10.1093/nar/gkt328.

[91] J.Köster y S. Rahmann, 'Snakemake: un motor de flujo de trabajo bioinformático escalable', Bioinformática, vol. 28, núm. 19, págs. 2520–2522, 2012, doi: 10.1093/bioinformatics/bts480.

[92] P. di Tommaso, M.Chatzou, EW Floden, PP Barja, E. Palumbo y C. Notredame, 'Nextflow permite flujos de trabajo computacionales reproducibles', Nature Biotechnology 2017 35:4, vol. 35, núm. 4, págs. 316–319, abril de 2017, doi: 10.1038/nbt.3820.

[93] F. daVeiga Leprevost et al., 'BioContainers: un marco de código abierto e impulsado por la comunidad para la estandarización del software', Bioinformatics, vol. 33, núm. 16, págs. 2580–2582, agosto de 2017, doi: 10.1093/BIOINFORMATICS/BTX192.

[94] C. Dong et al., 'Comparación e integración de métodos de predicción deletéreos para SNV no sinónimos en estudios de secuenciación del exoma completo', Hum Mol Genet, vol. 24, núm. 8, págs. 2125-2137, 2015,doi: 10.1093/hmg/ddu733.
[95] D. Quang, Y. Chen y X.Xie, 'DANN: Un enfoque de aprendizaje profundo para anotar la patogenicidad de variantes genéticas', Bioinformática, vol. 31, núm. 5, págs. 761–763, 2015, doi: 10.1093/bioinformatics/btu703.
[96] Y. Tian et al., 'REVEL yBayesDel supera a otros metapredictores in silico para la clasificación de variantes clínicas', Sci Rep, vol. 9, núm. 1, págs. 1 a 6, 2019, doi: 10.1038/s41598-019-49224-8.