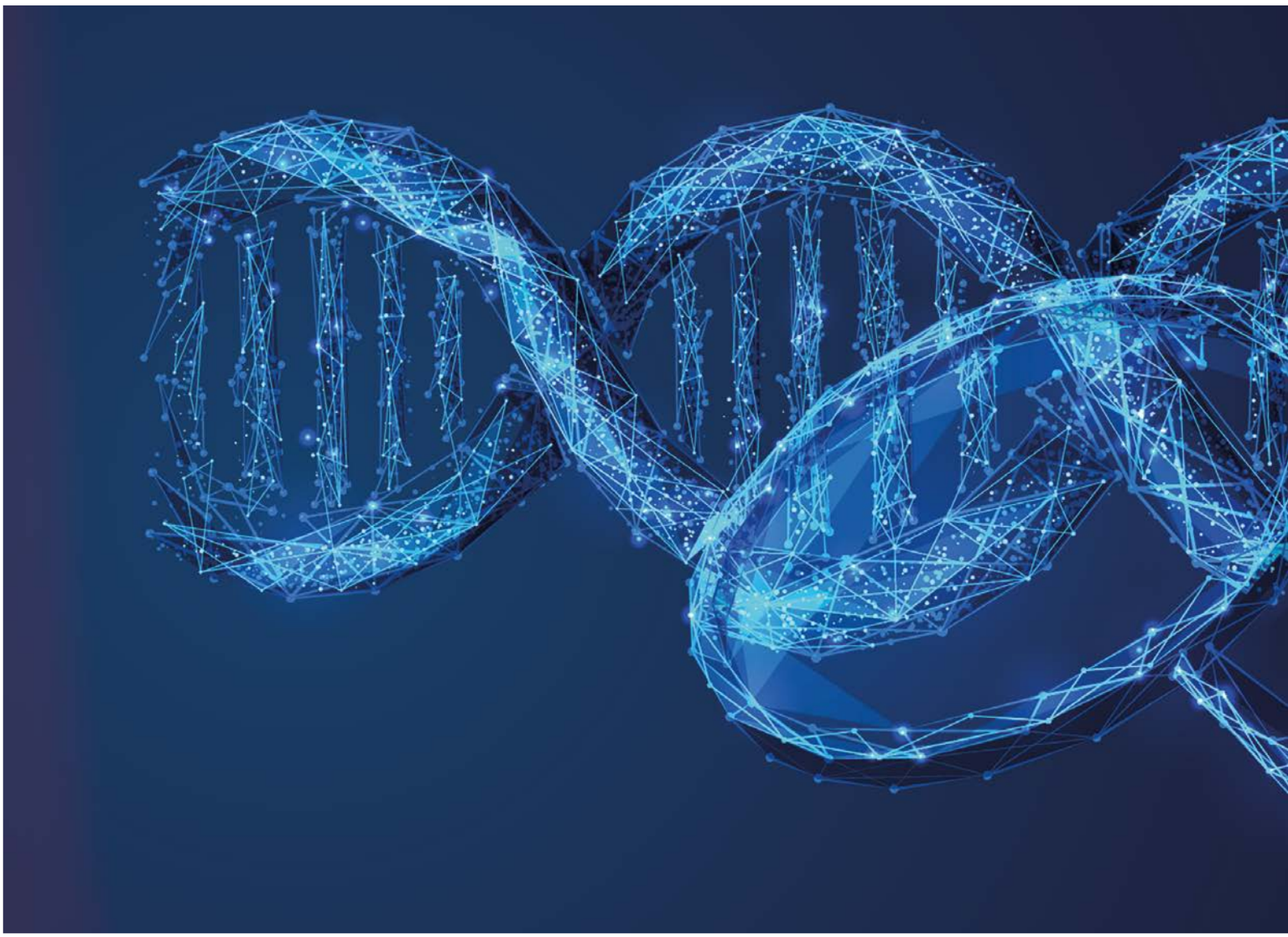

# ESTUDIO DE LAS ENFERMEDADES GENÉTICAS HUMANAS: LO QUE LA SECUENCIACIÓN MASIVA PUEDE OFRECERNOS

EL DESARROLLO DE LAS TECNOLOGÍAS DE SECUENCIACIÓN MASIVA O NEXT-GENERATION SEQUENCING (NGS) HA PERMITIDO AUMENTAR NUESTRO CONOCIMIENTO DEL GENOMA HUMANO HASTA NIVELES NUNCA ANTES VISTOS, GRACIAS A LA POSIBILIDAD DE SECUENCIAR UNA GRAN CANTIDAD DE INFORMACIÓN GENÉTICA DE MANERA RÁPIDA Y A UN PRECIO CADA VEZ MÁS REDUCIDO. NO OBSTANTE, TRABAJAR CON LOS ARCHIVOS RESULTANTES ES COMPLEJO Y COSTOSO COMPUTACIONALMENTE, POR LO QUE REQUIERE CONOCIMIENTOS BIOINFORMÁTICOS AVANZADOS, ASÍ COMO UNA INFRAESTRUCTURA INFORMÁTICA QUE DÉ SOPORTE AL PROCESAMIENTO Y ALMACENAJE DE LA GRAN CANTIDAD DE DATOS GENERADOS.

En los últimos años, las tecnologías NGS han logrado grandes avances tanto en investigación básica como en clínica, proporcionando conocimientos cada vez más profundos sobre la base genética de muchas enfermedades. Para poder comprender mejor lo que estas tecnologías nos ofrecen es importante tener una visión global de todo el proceso, desde la obtención de las muestras biológicas hasta los resultados finales que serán interpretados por los genetistas. Dicho proceso puede ser dividido en tres etapas: análisis primario, secundario y terciario.

## Análisis primario

Consiste en la extracción del ADN, secuenciación y obtención del archivo FASTQ. El ADN se obtiene a partir de muestras biológicas, las cuales pueden ser de distintos orígenes, como por ejemplo sangre, saliva, tejido fresco o parafinas.

Una vez extraído el ADN, se procede a la secuenciación. Generalmente consiste en un primer paso en el que se fragmenta el ADN en segmentos de 100-200 pares de bases. A continuación, a cada fragmento de ADN se le unirá un primer o adaptador que indica el punto de partida para la replicación. Posteriormente, cada fragmento se amplificará por PCR y, finalmente, se procederá a la lectura de cada uno de los fragmentos amplificados.

Como resultado, se obtendrán millones de lecturas de fragmentos de ADN, recogidas en unos archivos llamado FASTQs, junto con el valor de calidad asignado por el secuenciador a cada base.

## Análisis secundario

En esta segunda etapa tiene lugar el procesamiento bioinformático de los archivos FASTQs para obtener un archivo VCF (*Variant Call Format*) que recogerá las variantes genéticas detectadas en la muestra. Generalmente, esta etapa puede dividirse a su vez en los siguientes pasos:

- *Preprocesamiento.* Consiste en la eliminación de los adaptadores y otras secuencias de baja calidad, y un control de calidad de los FASTQs para garantizar lo máximo posible la integridad y calidad de los datos. Algunos de los algoritmos más utilizados son Fastp o Fastqc.
- *Alineamiento.* Las lecturas contenidas en los ficheros FASTQs se mapean frente al genoma humano de referencia. El archivo obtenido recibe el nombre de BAM (*Binary Aligned Mapped*), y contiene cada una de las lecturas ordenadas y mapeadas, es decir, con información de su localización dentro de la referencia utilizada. Uno de los algoritmos más utilizados es BWA-mem. Dependiendo de las características del experimento, será recomendable eliminar lecturas duplicadas del archivo BAM.
- *Llamada de variantes.* Identificación de variaciones genómicas presentes en el genoma del paciente con respecto al genoma humano de referencia. Las variantes pueden ser, por ejemplo, cambios en un solo nucleótido (SNVs, o *Single Nucleotide Variation*), o inserciones y deleciones de varios nucleótidos (llamadas conjuntamente INDELs). Las variantes identificadas se recogerán en un fichero VCF, donde se indica la coordenada genómica y el cambio concreto, junto con algunos valores estadísticos que permitirán al genetista interpretar la veracidad de la variante desde un punto de vista más técnico. Algunos de los algoritmos más utilizados son Haplotypecaller, BCFtools o VarScan2.

## Análisis terciario

Este último paso es el que va a dotar de importancia biológica a los datos obtenidos. Una vez que tenemos las variantes detectadas en el fichero VCF, se realiza la anotación, que consiste en consultar en bases de datos cada una de las variantes detectadas y añadir la información al VCF.

Estas bases de datos pueden ser:

- Funcionales (ej. RefSeq, Pfam).
- Poblacionales (ej. 1000 Genomes, dbSNP, ESP6500, ExAC, gnomAD).
- De predicción de impacto funcional *in silico* (ej. dbNSFP, dbscSNV).
- Clínicas (ej. ClinVar, HPO).

**LA NGS SE ESTÁ CONVIRTIENDO EN UNA HERRAMIENTA INDISPENSABLE, TANTO A NIVEL DE INVESTIGACIÓN TRASLACIONAL COMO A NIVEL DIAGNÓSTICO**

De esta manera obtendremos el VCF final anotado, a partir del cual se podrán realizar las interpretaciones correspondientes por parte del genetista. La correcta interpretación de los resultados va a depender, en gran medida, de la información disponible para cada una de las variantes identificadas. Una vez revisadas, las variantes seleccionadas podrán ser validadas por otros métodos de laboratorio, y finalmente informadas al personal médico o investigador.

Dependiendo del caso, es posible que haya miles de variantes detectadas, y la revisión de cada una de ellas desde el propio VCF es una tarea que se puede volver muy tediosa. Por este motivo existen programas que vuelcan la información del VCF en tablas navegables, filtrables y con múltiples opciones que facilitan en gran medida la labor de revisión (Figura 1).

## Principales tipos de análisis genómicos

### *Genoma*

La secuenciación del genoma completo o WGS (*Whole Genome Sequencing*) consiste en la secuenciación de todo el genoma del paciente. Esta tecnología ofrece una cobertura uniforme, con una profundidad media de 30-60 lecturas por posición (30-60X). Esto permite tanto la detección de variantes puntuales (SNVs o INDELs), como de variantes estructurales, por ejemplo CNVs (*Copy Number Variations*).

Es el método más indicado para detectar mutaciones en genes previamente no asociados a la enfermedad de estudio, especialmente cuando éstas están ubicadas en regiones no codificantes.

En contraposición a toda la información que produce, es un método costoso, tanto a nivel económico como computacional, lo que hace que, hoy en día, no sea la aproximación prioritaria en muchos laboratorios, especialmente en el ámbito clínico donde se prioriza la coste-efectividad de las pruebas utilizadas.

### *Exoma*

La secuenciación de exoma completo o WES (*Whole Exome Sequencing*) consiste en la secuenciación de las regiones codificantes, o exones, de los más de 20.000 genes que componen el genoma humano.

El WES resulta especialmente interesante para aquellos casos que ya han pasado por pruebas genéticas previas más específicas que no fueron concluyentes para establecer un diagnóstico, para casos que presenten un fenotipo clínico heterogéneo que no corresponda con ninguna enfermedad conocida o que corresponda a varias enfermedades genéticas, o para aquellas enfermedades genéticas que carezcan de un test genético específico.

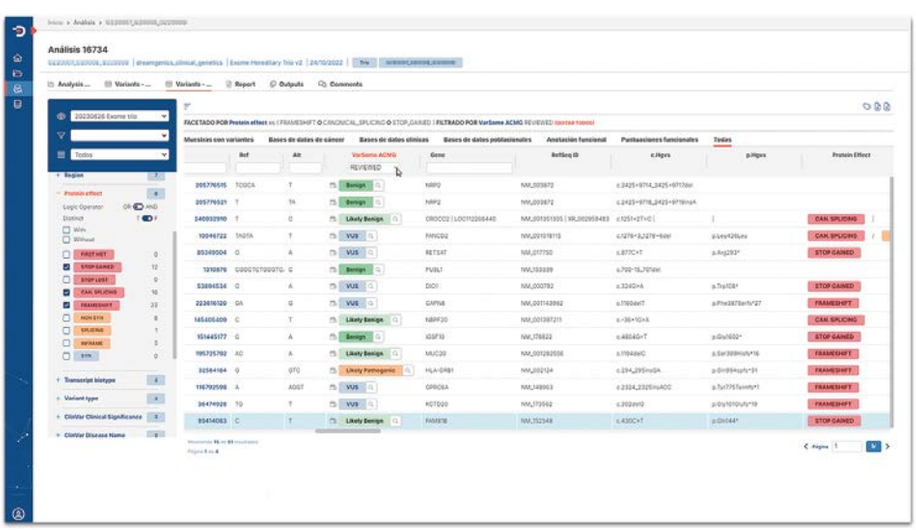


***Figura 1. Índice de variantes visualizado en la plataforma Genome One Reports***

Al centrarse en una pequeña parte del genoma (en torno al 2%), esta tecnología permite una mayor profundidad, de unas 100-150x, permitiendo detectar variantes a más baja frecuencia que en el caso del WGS. Además, supone un menor coste en análisis bioinformático y almacenamiento de datos.
Como principal limitación, no se tendrán datos de regiones no codificantes. Además, la cobertura a lo largo de la secuencia será muy variable, dificultando por ejemplo la detección de variantes estructurales como CNVs. Por este motivo, hay que destacar la importancia de valorar bien el método de secuenciación para el estudio, y asegurarse de que las regiones o variantes de interés pueden ser cubiertas por la tecnología elegida.

***Panel***

La secuenciación de un panel de genes constituye una buena alternativa para el estudio de un set de genes relacionados con una patología específica, obteniendo una secuenciación más dirigida y a un menor coste.
Estos paneles utilizan sondas específicamente diseñadas para asegurar la secuenciación de las principales variantes, tanto en regiones codificantes como intrónicas, que estén relacionadas con la patología de estudio. La profundidad media alcanzada es de unos 500-1000x, lo que permite detectar variantes a muy baja frecuencia (detectando mosaicismo en variantes germinales, o eventos clonales en análisis somático). Pese a la mayor rapidez y menor coste, es una aproximación muy dirigida, que requiere de un elevado conocimiento de la patología para el diseño de la sonda, o en su defecto, de la disponibilidad de kits comerciales acordes a la enfermedad de estudio.

## Conclusiones

La NGS se está convirtiendo en una herramienta indispensable, tanto a nivel de investigación traslacional como a nivel diagnóstico, proporcionando resultados cada vez más precisos, rápidos y baratos, y desplazando progresivamente a otras tecnologías. La reciente publicación de la cartera común de pruebas genéticas en el Sistema Nacional de Salud pone de manifiesto la creciente importancia de esta tecnología y el papel protagonista que va a jugar en los próximos años. Así mismo, el desarrollo de herramientas bioinformáticas cada vez más avanzadas resulta fundamental para que seamos capaces de extraer la máxima información posible a los datos obtenidos en la secuenciación masiva. La combinación de ambas permitirá incrementar nuestro conocimiento de las bases moleculares de las enfermedades, pudiendo conocer las características individuales de éstas en cada paciente, ayudando en el diseño de estrategias terapéuticas personalizadas. ✚

Por: **Alicia Gómez y Sara Bonilla**
Bioinformatics Scientists
en Dreamgenics

**LA RECIENTE PUBLICACIÓN DE LA CARTERA COMÚN DE PRUEBAS GENÉTICAS EN EL SNS PONE DE MANIFIESTO LA CRECIENTE IMPORTANCIA DE LA NGS**

**Referencias:**
1. Xuan J., Yu Y., Qing T. et al. Next-generation sequencing in the clinic: promises and challenges. Cancer letters 340, 2 (2013)
2. Koboldt, D.C. Best practices for variant calling in clinical sequencing. Genom. Med. 12, 91 (2020).
3. Austin-Tse, C.A., Jobanputra, V., Perry, D.L. et al. Best practices for the interpretation and reporting of clinical whole genome sequencing. Genom. Med. 7, 27 (2022).