RECyT
Year 26 / Nº 40 Supplement Nº 1 / 2024 / 24–25

From Sequencing to Variant Analysis: Exploring Sequencing Data Files in Clinical Genomics

Del Secuenciador al Análisis de Variantes: Un Recorrido por todos los Archivos de Secuenciación

Rodrigo F. E. Bogado[1, *]; Marcelo D. Gamarra1

1- Instituto de Genética Humana de Misiones (IGeHM). Área de Bioinformática. Par-que de la Salud de la Provincia de Misiones. Misiones, Argentina.

* E-mail: mdgamarraok@gmail.com

Received: 17/10/2023; Accepted: 10/11/2023

Abstract

In clinical genomics, the path from DNA sequencing to diagnosis involves a series of essential steps that require interdisciplinary collaboration, encompassing fields such as genetics, biochemistry, and *bioinformatics*. Each phase of this process is crucial for obtaining valuable genetic information and applying it to the diagnosis of genetically based diseases. The process begins with the generation of *DNA* reads from a biological sample using *next-generation sequencing (NGS) methods*. These reads are stored in files called FASTQ, containing nucleotide sequences and associated quality scores for each base detected during sequencing. The quality and accuracy of these reads are essential for the success of the entire process. Subsequently, the reads are aligned to the *human reference genome* to determine their precise location within it. The reference genome is a comprehensive and well-annotated representation of the human DNA, with several versions available, with the most recent being *GRCh38* (*Genome Reference Consortium human genome build 38*). Efficient access to specific genome regions is achieved computationally through "indexing." *Indexing* is a fundamental process in genomics and bioinformatics, involving the creation of an index or specialized data structure that enables rapid and efficient access to the reference genome sequence. Once aligned, *SAM* (*Sequence Alignment/Map*) or *BAM* (*Binary Alignment/Map*) files are generated, recording the position of each read in the genome. Following read alignment, *variant filtering* takes place. In this phase, discrepancies between the patient's genome and the reference genome are identified. This is where bioinformatics plays a crucial role, using tools and algorithms to detect these variants accurately and efficiently. Subsequently, variant annotation is performed, assigning functions and characteristics to the identified variants. This entails consulting *biological databases* such as *dbSNP, ClinVar, gnomAD, Uniprot,* and *Ensembl,* which contain information about previously documented variants and their associations with diseases. *Variant interpretation* is the final and critical stage of this process. Here, clinical specialists and bioinformaticians closely collaborate to determine whether any of the identified variants are relevant to the patient's disease or simply represent polymorphisms in the general population. This process involves a thorough evaluation of the clinical relevance of each variant, considering factors such as heritability, penetrance, clinical impact, and population frequency. Upon completing variant interpretation, a detailed report summarizing relevant findings and providing clinical recommendations is generated. This marks the initiation of the *genetic counseling phase* and the planning of a *personalized therapeutic approach*, if necessary, to deliver more precise and effective medical care to the patient. Given these considerations, DNA sequencing and its analysis in clinical genomics demand multidisciplinary collaboration, ranging from experts in genetics and clinical medicine to those with deep knowledge in computational tools. A comprehensive understanding of all stages of the sequencing process, from read generation to variant analysis, is fundamental and presents a challenge for advancing toward preventive, precise, and personalized medicine.

Keywords: Clinical Genomics, Bioinformatics, Next-Generation Sequencing.

Resumen

En genómica clínica, el camino desde la secuenciación del ADN hasta el diagnóstico implica una serie de pasos que requiere del trabajo interdisciplinario, implicando áreas como la genética y la bioinformática. Cada fase de este proceso es fundamental para obtener información valiosa y aplicarla en el diagnóstico de enfermedades de base genética.

El proceso comienza con la generación de lecturas de ADN a partir de una muestra mediante métodos de secuenciación de nueva generación *(NGS* por sus siglas en ingles). Estas lecturas se almacenan en *FASTQ*, que contienen las lecturas y calidades asociadas a la detección en la corrida. La calidad y precisión de estas son esenciales para el éxito de todo el proceso. Las lecturas se alinean contra el *genoma de referencia humano* (GRH) para determinar su ubicación exacta dentro del mismo. El GRH es una representación completa y anotada del ADN humano para el cual existen varias versiones siendo la más actual la versión *GRCh38* (*Genome Reference Consortium human genome build 38*). El acceso rápido y menos costoso computacionalmente a las regiones específicas del genoma se logra mediante la *indexación*, un proceso que implica la creación de un índice o una estructura especializada que permite un acceso eficiente a la secuencia del genoma. Luego, se obtienen los archivos *SAM* (*Sequence Alignment/Map*) o *BAM* (*Binary Alignment/Map*), que registran la posición de cada lectura. Una vez alineadas las lecturas, se procede al *filtrado de variantes*. En esta fase, se identifican las discrepancias entre el genoma del paciente y el GRH. Es aquí donde la bioinformática desempeña un papel crucial para detectar estas variantes de manera precisa y eficiente. Luego, se realiza la anotación de variantes, donde se asignan funciones y características a las variantes identificadas. Esto implica la consulta de bases de datos *biológicas*, como *CinVar, gnomAD, Uniprot* y *Ensembl*, las cuales contienen información sobre variantes previamente documentadas y su relación con enfermedades. La interpretación de las variantes representa la etapa final de este proceso. Aquí, especialistas clínicos y bioinformáticos colaboran estrechamente para determinar si alguna de las variantes identificadas guarda relación con la enfermedad del paciente o si se trata simplemente de un polimorfismo presente en la población general. Este proceso implica la evaluación exhaustiva de la relevancia clínica de cada variante, teniendo en cuenta factores como la heredabilidad, penetrancia y frecuencia poblacional. Una vez completada la interpretación de las variantes, se obtiene un informe detallado que resume los hallazgos relevantes y proporciona recomendaciones clínicas y se inicia la fase de asesoramiento genético y la planificación de un enfoque terapéutico personalizado para brindar una atención médica más precisa y efectiva al paciente. Así, la secuenciación del ADN y su análisis en genómica clínica demandan una colaboración multidisciplinaria, que abarca desde expertos en genética y clínica hasta profesionales informáticos. Comprender a fondo todas las etapas del proceso de secuenciación, desde la generación de lecturas hasta el análisis de variantes, resulta fundamental y plantea un desafío para el avance hacia una medicina preventiva, precisa y personalizada.

Palabras clave: Genómica Clínica, Bioinformática, NGS.