Implementación de un banco de datos genéticos de variantes asociadas a muestras de pacientes clínicos en Colombia.

Jorge Díaz-Riaño[1], Daniel Mahecha[1], Yenny Gómez[1].

[1] Biotecnología y Genética S.A.S - Biotecgen S.A.S

- Email de contacto: biologiacomputacional@biotecgen.com.co

El acceso a la información genética a través de plataformas de secuenciación cada vez más precisas y económicas, ha provocado una ola de información *ómica* que ha normalizado el uso de conceptos computacionales como *Big Data,* y de herramientas bioinformáticas en escenarios novedosos donde su función es trascendental apoyando el desarrollo de nuevas técnicas de aprovechamiento de dicha información. En el campo de la salud pública, por ejemplo, existe la necesidad de estructurar bases de datos que permitan aplicar la información poblacional en la toma de decisiones y en la generación de diagnósticos clínicos más precisos.

Este trabajo pretende implementar un banco de datos genómicos para la información de variantes genéticas de pacientes colombianos con fines de aprovechamiento clínico, enfocada principalmente al diagnóstico de patologías asociadas a enfermedades genéticas, a la vez que propone una estructura de almacenamiento útil para el desarrollo de analítica de datos orientada a investigación médica y poblacional del país al incluir información de secuenciación y la *metadata* propia de los procesos clínicos asociados de manera anonimizada. La información genómica se recopiló a partir de las muestras de exomas, genes únicos y paneles de genes, de pacientes del laboratorio Biotecnología y Genética S.A.S quienes con autorización previa aceptaron el uso anonimizado de su información. Los datos asociados a dichas muestras consisten en datos crudos de secuenciación en formato *fastq.gz, a*rchivos de mapeo en formato *bam* y archivos de llamado de variantes anotados en formato *vcf* específicos por muestra junto a un *vcf* de toda la cohorte con información de frecuencias alélicas. Cada muestra tiene asociada una información de *metadata* anonimizada consistente en: sexo, fecha de nacimiento, lugar de nacimiento, consanguinidad parental, términos HPO asociados, variantes de importancia clínica e identificador anonimizado. El almacenamiento de los objetos se realiza en *buckets* de *Google Cloud* empleando el protocolo de cifrado *Google managed key*. El acceso se controla a través del servicio de autenticación y con la asignación de roles a usuarios específicos (*Identity and Access Management -IAM)*. A 31 de diciembre de 2021, contamos con un volumen total de archivos de 893Gb correspondientes a 600 muestras, abarcando un total de 22,497,664 SNPs (de los cuales 3,729,833 son *singletons*) y 3,954,984 Indels. El diseño y la publicación de este tipo de bases de datos conlleva la mejora de los procesos de diagnóstico clínico e investigación al contar con información estructurada de frecuencias alélicas y demás

métricas para poblaciones específicas . Es imperativo sistematizar, centralizar y facilitar el acceso al recurso genético humano en Colombia en concordancia con otras iniciativas internacionales, con el fin de optimizar los procesos de interpretación de patogenicidad de las variantes presentes en el acervo génico nacional. El desarrollo de nuestra base de datos provee una referencia de una cohorte amplia y en expansión para la realización de estudios genéticos de asociación y genética clínica en Colombia, con el fin de potenciar y facilitar las distintas áreas de investigación biomédica.