UNIVERSITAT DE BARCELONA

Sistema de paneles virtuales de genes para genómica clínica basado en NGS

Arnau Sellarès Rubio

Sistema de paneles virtuales de genes para genómica clínica basado en NGS

Arnau Sellarès Rubio

Sistema de paneles virtuales de genes para genómica clínica basado en NGS

UNIVERSITAT DE BARCELONA

Programa de Doctorado en Biomedicina - Universidad de Barcelona

---

Memoria presentada por **Arnau Sellarès Rubio**
para optar al grado de Doctor por la Universidad de Barcelona.

Realizada en la empresa Genomcore S.L.
bajo la dirección del Dr. Miquel Ràmia-Jesús.

Director de tesis

Dr. Miquel Ràmia-Jesús

Tutor de tesis

Dr. Josep Lluís Gelpí Buchaca

Arnau Sellarès Rubio

20 de septiembre de 2022

Para Imma, Xavier,
Berta y Luisa.

Agradecimientos

Estas líneas van dedicadas a toda la gente que ha hecho posible que este trabajo vea la luz. Antes que nada, me gustaría agradecer a Mike la infinidad de horas dedicadas a la elaboración y revisión de esta tesis, por sus aportaciones y conocimientos. A Laura y Oscar por sus consejos, gestiones, la oportunidad de trabajar en Genomcore, y toda la ayuda ofrecida, que no habrá sido fácil mientras cuidaban de Gabriel. A Andrea y Mònica por su paciencia y por tener siempre la respuesta a todas mis dudas. A Jordi, Darío, Geovanny, Héctor y todo el equipo de desarrolladores por sus ánimos y soporte, y la paciencia de aguantar hasta aquí. A Josep Lluís y Eduard por su ayuda y sus comentarios en las comisiones de seguimiento. A Júlia por el diseño de la cubierta. Finalmente, a mi familia, amigas y amigos, por estar a mi lado, ofrecerme su ayuda y hacerme reír.

Resumen

El diagnóstico de enfermedades humanas originadas a causa de variantes y mutaciones en el ADN de cada persona está experimentando un cambio de paradigma, desde la aproximación genética tradicional hacia una que aproveche todo el potencial de la genómica clínica mediante la *next-generation sequencing* o secuenciación de última generación (NGS), ayudando a poner fin a la odisea diagnóstica a la que se ven expuestos muchos pacientes, además de mejorar el rendimiento y la coste-eficiencia de las pruebas realizadas.

Sin embargo, la inmensa cantidad de datos generados actualmente mediante NGS implican un desafío computacional y de almacenamiento, además de suponer un obstáculo adicional con respecto a la correcta interpretación de las variantes identificadas por parte de los analistas clínicos. La detección de *Variants of Uncertain Significance* o variantes de significado incierto (VUS), *Unsolicited Findings* o descubrimientos no solicitados (UFs), y *Secondary Findings* o descubrimientos secundarios (SFs) ligadas a la naturaleza de la NGS incrementan la carga de trabajo interpretativo y el tiempo de respuesta para reportar los resultados al paciente, hecho que afecta directamente a la gestión de su tratamiento médico.

En este proyecto de tesis, se han diseñado dos productos integrados en el *pipeline* bioinformático de la empresa Genomcore, es decir, el conjunto de aplicaciones encadenadas que constituyen todo el proceso del análisis NGS, para ayudar a reducir la carga interpretativa de los analistas y el tiempo de respuesta para el reporte de resultados, que mejoran la automatización de todo el proceso: el *Virtual Panel Management System* o sistema de gestión de paneles virtuales (VPMS) y la *Report Generation Tool* o

herramienta de generación de informes (RGT).

El VPMS permite la creación de paneles virtuales de genes para realizar análisis genómicos dirigidos a regiones concretas del ADN, minimizando así la detección de VUS, UFs y SFs para facilitar la labor interpretativa de los analistas. La reutilización de los datos de secuenciación permite también al VPMS optimizar el proceso de reanálisis. La RGT proporciona una herramienta para la generación automatizada de informes clínicos de resultados, así como de salud personalizada mediante el uso de otros datos biomédicos.

De esta manera, la implementación de ambos sistemas proporciona a Genomcore una mejora en la integración, optimización y automatización del proceso de análisis de datos NGS y otros datos biomédicos en su plataforma bioinformática, desde la extracción de los datos en bruto hasta el informe final de resultados.

Abstract

The diagnostic of human diseases caused by variants and mutations in each individual's DNA is experiencing a paradigm shift from a traditional genetics approach towards one that exploits all the potential of clinical genomics through the use of next-generation sequencing (NGS), helping to put an end to the diagnostic odyssey many patients suffer, and also increasing the yield and cost-efficiency of the performed tests.

However, the huge amount of NGS data generated nowadays entails computational and storage challenges, also creating new barriers towards the correct interpretation of identified variants by clinical analysts. Detection of variants of uncertain significance (VUS), unsolicited findings (UFs), and secondary findings (SFs) that naturally arise from NGS testing increase diagnostic burden and response times when reporting results to the patients, therefore affecting medical treatment management overall.

In this dissertation, we designed two products that have been integrated into Genomcore's bioinformatics pipeline, i.e. the combination of all the applications that are part of the NGS analysis, to help reduce interpretive burden for analysts and response times for the reporting of results, which improve automation of the whole process: the Virtual Panel Management System (VPMS) and the Report Generation Tool (RGT).

The VPMS streamlines virtual gene panel creation in order to execute genomic analyses directed towards specific DNA regions, therefore minimizing the detection of VUS, UFs and SFs to ease the diagnostic burden onto analysts. Reusage of sequencing data also allows for the optimization of reanalysis processes by the VPMS. The RGT provides a tool to generate clinical reports of results in an automated fashion, as well as personalized health reports through the use of other biomedical data.

Thus, the implementation of both systems improves integration, optimization and automation of NGS and other biomedical analyses at Genomcore's bioinformatics platform, from data extraction to the reporting of results.

Índice

Índice de figuras

Índice de tablas

Índice de ficheros

Índice de abreviaciones

ACMG	*American College of Medical Genetics and Genomics* o Colegio Americano de Genética Médica y Genómica
ADNc	ADN complementario
AMP	*Association for Molecular Pathology* o Asociación de Patología Molecular
API	*application programming interface* o interfaz de programación de aplicaciones
APM	*Application Performance Monitoring*
AWS	*Amazon Web Services*
BAM	*Binary Alignment Map*
Bash	*GNU Bourne-Again Shell*
BCL	*binary base call*
BIMS	*Biomedical Information Management System* o sistema de gestión de información bioinformática
bp	pares de bases
CCDS	*consensus coding sequence*
CDS	*coding sequence* o región codificante de un gen
CGGD	*Clinical Grade Genomic Database* o base de datos genómicos de grado clínico
CGH	*Comparative Genomic Hybridization* o hibridación genómica comparativa
CGVR	*Clinical Genomic Variant Repository* o repositorio de variantes genómicas clínicas
CI/CD	*Continuous Integration and Continuous Delivery* o integración y entrega continuas
ClinGen	*Clinical Genome Resource*
CMA	*Chromosomal Microarray Analysis* o análisis de microarrays cromosómicos
CNV	*Copy Number Variation* o variación en el número de copias
CSV	*Comma-Separated Values* o valores separados por comas
CWL	*Common Workflow Language*
DTC	*direct-to-consumer*
ECR	*Elastic Container Registry*
EHR	*Electronic Health Record* o historial clínico electrónico
ENCODE	*Encyclopedia of DNA elements*

ESHG *European Society of Human Genetics* o Sociedad Europea de Genética Humana
FISH *Fluorescence In Situ Hybridization* o hibridación in situ fluorescente
FTP *File Transfer Protocol* o protocolo de transferencia de archivos
G2MC *Global Genomic Medicine Collaborative*
GA4GH *Global Alliance for Genomics and Health* o Alianza Global para Genómica y Salud
GATK *Genome Analysis Toolkit*
GIAB *Genome in a Bottle*
GMDR *Genomic Medical Data Repository* o repositorio de datos médicos genómicos
GMED *Genomic Medical Evidence Database* o base de datos de evidencia genómica
GRC *Genome Reference Consortium* o consorcio de referencia del genoma
gVCF *genomic variant call format*
GVF *genome variant format*
HGMD *Human Gene Mutation Database*
HGP *Human Genome Project* o Proyecto del Genoma Humano
HL7 *Health Level Seven*
HPO *Human Phenotype Ontology*
HQHSV *High-Quality Human Sequence/Variant* o secuencia/variante humana de alta calidad
HVP *Human Variome Project* o proyecto del varioma humano
IF *Incidental Finding* o descubrimiento incidental
IGV *Integrative Genomics Viewer*
indel inserción/deleción
iPOP *integrative Personal Omics Profile* o perfil ómico personal integral
LIMS *Laboratory Information Management System* o sistema de gestión de información de laboratorio
LOF *loss-of-function* o pérdida de función
LOINC *Logical Observation Identifier Names and Codes*
MGI *Medical Genome Initiative* o iniciativa del genoma médico
MLPA *Multiplex Ligation-dependent Probe Amplification*
MME *Matchmaker Exchange*
MoG *Made of Genes*
NGS *next-generation sequencing* o secuenciación de última generación
OGS *Opportunistic Genomic Screening* o cribado genómico oportunista
OMIM *Online Mendelian Inheritance in Man*
PCR *Polymerase Chain Reaction* o reacción en cadena de la polimerasa

PRS	*Polygenic Risk Score*
RefSeq	*Reference Sequence*
RGT	*Report Generation Tool* o herramienta de generación de informes
ROI	*Region Of Interest* o región de interés
RWD	*Real-World Data* o datos del mundo real
S3	*Amazon Simple Storage Service*
SF	*Secondary Finding* o descubrimiento secundario
SFMPP	*Société Française de Médecine Prédictive et Personnalisée* o Sociedad Francesa de Medicina Predictiva y Personalizada
SFTP	*Secure File Transfer Protocol* o protocolo seguro de transferencia de archivos
SMS	*Single-Molecule Sequencing*
SNP	*Single-Nucleotide Polymorphism* o polimorfismo de nucleótido único
SNV	*Single-Nucleotide Variant* o variante de nucleótido único
SV	*Structural Variant* o variante estructural
T2T	consorcio *Telomere-to-Telomere*
UF	*Unsolicited Finding* o descubrimiento no solicitado
VCF	*variant call format*
VEP	*Ensembl Variant Effect Predictor*
VPMS	*Virtual Panel Management System* o sistema de gestión de paneles virtuales
VUS	*Variant of Uncertain Significance* o variante de significado incierto
WES	*Whole-Exome Sequencing* o secuenciación del exoma completo
WGS	*Whole-Genome Sequencing* o secuenciación del genoma completo

CAPÍTULO 1

Introducción

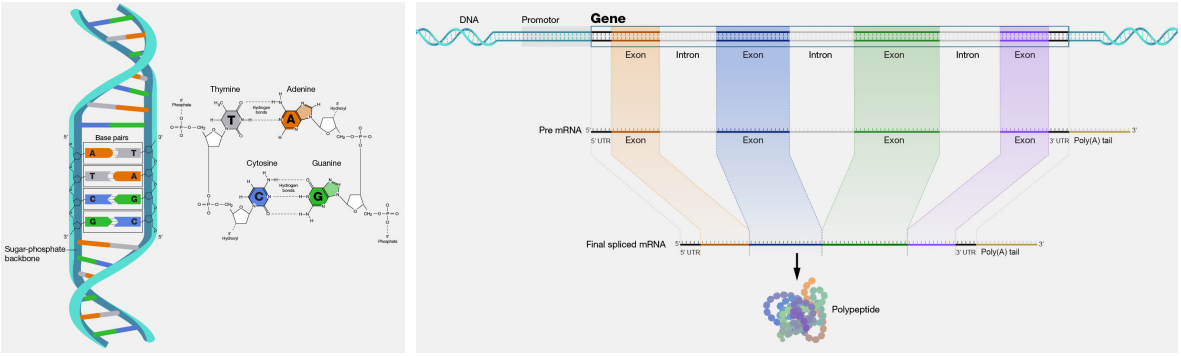
En este primer capítulo se hace una introducción de todos los conceptos utilizados en esta tesis, y un análisis del estado del arte para el sector de la genómica clínica en la interpretación de los resultados de secuenciación mediante el uso de aproximaciones en forma de paneles virtuales de genes, focalizados en proveer un diagnóstico definitivo para un paciente, así como incrementar su coste-eficiencia. Además, se muestran también los esfuerzos actuales para optimizar la translación de la gran cantidad de datos biológicos que se pueden obtener de una persona en información accionable y relevante para su salud, de una manera personalizada y accesible.

1.1 Genética y enfermedad

1.1.1 Los genes y sus productos

De tal palo, tal astilla; esta expresión popular encierra el concepto científico de **herencia**, un mecanismo que permite la transferencia de características de los padres y madres a sus hijos e hijas como, por ejemplo, el color de los ojos o del pelo, pero también, desafortunadamente, ciertas condiciones de salud, o predisposición a enfermedades (Gayon, 2016). Los primeros avances para dilucidar las causas de la herencia

no sucedieron hasta 1866 con la publicación de las memorias de Mendel sobre la hibridación en plantas, donde descubrió que existían *elementos* que determinaban los rasgos de un organismo y, además, se transmitían entre generaciones (Mendel, 1866; Hoßfeld *et al.*, 2017). Mendel, considerado actualmente el padre y fundador de la genética moderna, explicaba así como dichos rasgos eran heredados por las siguientes generaciones mediante un mecanismo predecible e inevitable a través de esos elementos desconocidos, en lo que actualmente conocemos como **herencia simple** o herencia mendeliana. El término **gen** no fue acuñado para esos elementos hasta años más tarde, junto con la definición de **fenotipo** (las características o rasgos visibles de una persona) y **genotipo** (la información genética que se hereda y determina el fenotipo) (Johannsen, 1914). Posteriormente, en 1953 se descubrió que los genes eran parte del ADN de una persona, una molécula compuesta por una secuencia de cuatro tipos diferentes de nucleótidos (fig. 1.1a), estructurada en forma de doble hélice (Watson y Crick, 1953; Maddox, 2003; Olby, 2003).



(A) *El ADN y los nucleótidos que lo componen. Imagen del NIH.*

(B) *Fabricación de proteínas a partir de los exones de los genes en el ADN. Imagen del NIH.*

FIGURA 1.1: *El ADN, los genes y las proteínas.*

Actualmente, sabemos que el gen es la unidad básica de herencia entre generaciones; son secuencias de ADN que codifican para uno o más ARNs, que a su vez pueden codificar para una o más **proteínas**, que ayudan al cuerpo a realizar sus funciones vitales y dirigir la actividad de las células (Portin y Wilkins, 2017). La estructura de un gen codificante consiste de varios elementos, de los cuales solamente una pequeña

parte, los **exones**, corresponden a la secuencia que acaba codificando para una proteína (Crick, 1979). Los otros elementos incluyen regiones promotoras, reguladoras, e **intrones** (fig. 1.1b). Por otro lado, existen también genes y regiones no codificantes, que ocupan, en la especie humana, la gran mayoría de la secuencia de nuestro ADN y participan en la expresión génica (Maher, 2012).

1.1.2 Genética y genómica

El término **genética** fue definido por primera vez en 1905 en una carta que el biólogo inglés W. Bateson le envió a su compañero A. Sedgwick para definir la ciencia de la herencia y la variación biológica (Gayon, 2016); es el estudio de los genes de manera individual, sus variaciones, su papel en la herencia, y sus efectos en los organismos para tratar de identificar su función y entender cómo su alteración puede causar una enfermedad. En este sentido, la **genética humana** es la ciencia de la variación biológica en los humanos (McKusick, 1975).

La **genómica** es un término más reciente, introducido en 1987 (McKusick y Ruddle, 1987), con una visión más global; estudia la información que contienen todos los genes de un organismo, es decir, su **genoma**, así como las interacciones entre ellos y con el entorno (Primrose y Twyman, 2003). El genoma es todo el material genético de una persona: el ADN que se transmite a las siguientes generaciones, incluyendo sus genes y otros elementos que controlan la actividad de éstos. Se estima que cada persona tiene alrededor de 20,000 genes que codifican para proteínas, de un total de 63,494 genes descritos hasta la fecha (Nurk *et al.*, 2022).

1.1.3 Variación genética y enfermedades

El genoma humano es prácticamente idéntico en todas las personas; las diferencias observadas en el genotipo de cada persona son las que definen las variantes o **variación genética humana**. El mecanismo último de generación de variación es la **mutación**,

término definido a partir de 1920 (Gayon, 2016), que provoca cambios aleatorios en la secuencia del ADN y, por lo tanto, puede producirse en casi todas las células de un organismo. De media, cada persona tiene 24-25 millones de bases afectadas por variaciones, es decir, casi el 1% de su ADN. Por tipología, la variación humana se divide en tres categorías principales (fig. 1.2): la mayoría de variaciones son *Single-Nucleotide Variants* o variantes de nucleótido único (SNVs) e inserciones/deleciones (indels), y solo el 0.01% corresponde a *Structural Variants* o variantes estructurales (SVs); por número de bases afectadas, las SVs representan aproximadamente el 80% de la variación (Roach *et al.*, 2010; The 1000 Genomes Project Consortium, 2015).

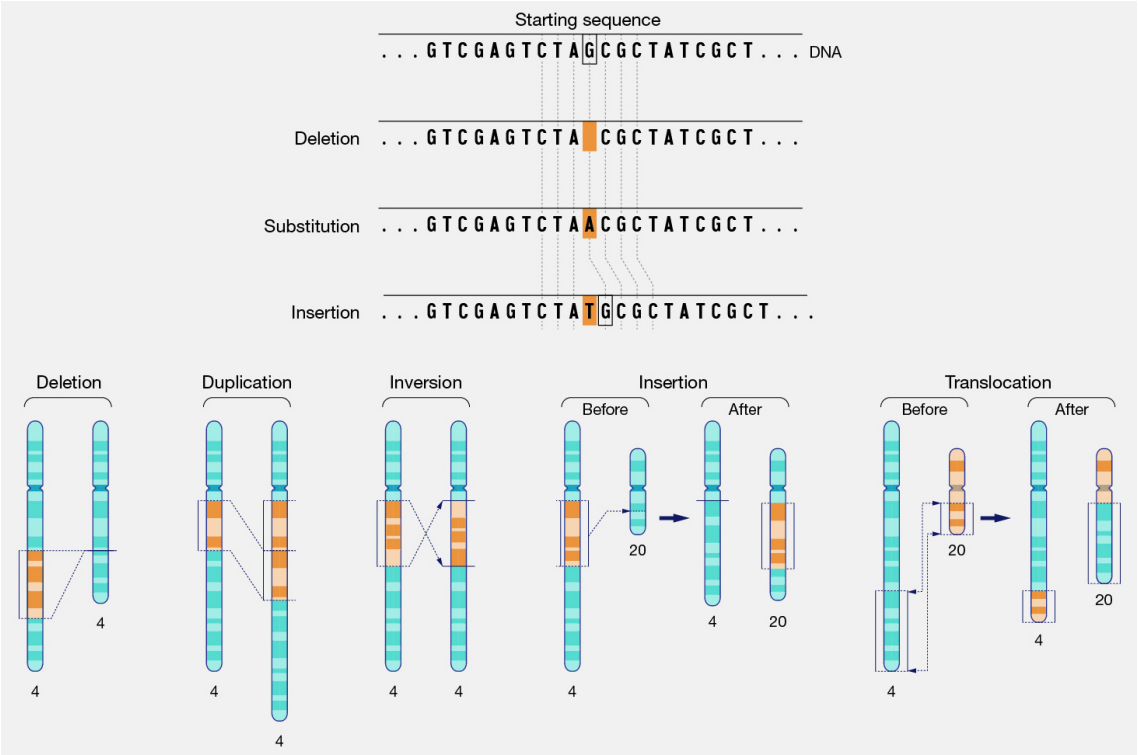


FIGURA 1.2: *Diferentes tipos de variación genética: una SNV es una sustitución nucleotídica en una posición del genoma, e incluye también los SNPs cuando su frecuencia poblacional es de un 1% o más (The International HapMap Consortium, 2003; Katsonis* et al.*, 2014); una indel define una inserción y/o deleción de un fragmento de ADN de menos de 1 kilobase (Sehn, 2015); una SV comprende regiones genómicas de más de 1 kilobase, y las hay de diferentes tipos, por ejemplo, duplicaciones, inversiones, y translocaciones (Feuk, Carson y Scherer, 2006). Imagen del NIH.*

Hasta la fecha, se han catalogado hasta un total de 88 millones de variantes humanas; 84.7 millones de *Single-Nucleotide Polymorphisms* o polimorfismos de nucleótido

único (SNPs), 3.6 millones de indels, y 60,000 SVs (The 1000 Genomes Project Consortium, 2015).

Las variantes genéticas humanas, que hacen que cada persona tenga un genotipo y fenotipo único, también pueden tener efectos en la salud: pueden provocar enfermedades, al generar proteínas malformadas que no pueden llevar a cabo sus funciones; alternativamente, pueden afectar a la respuesta a ciertos medicamentos, o a la probabilidad de desarrollar una enfermedad. De esta manera, contienen pistas importantes sobre las causas de las enfermedades genéticas que afectan a diferentes poblaciones, o a ciertas personas dentro de la misma población. Las enfermedades genéticas se pueden clasificar en **enfermedades mendelianas** o monogénicas si están causadas por genes individuales, o en **enfermedades complejas** o poligénicas si se deben a una combinación de factores genéticos y ambientales (fig. 1.3). La genética se utiliza para estudiar cómo se heredan entre generaciones ciertas enfermedades mendelianas, y la genómica para analizar las diferencias en múltiples genes que provocan enfermedades complejas.

Solo las enfermedades provocadas por mutaciones en las células de la **línea germinal** son hereditarias, ya que dichas mutaciones se transmiten a través de las células a la descendencia. Otras enfermedades no lo son, porque las mutaciones afectan solo a las **células somáticas** de un organismo, como en el caso del cáncer.

En cualquier caso, la genética y la genómica proporcionan herramientas para profundizar en la comprensión de las enfermedades, y ayudan a desarrollar diagnósticos y tratamientos dirigidos más tempranos. Esta tarea es de suma importancia en medicina; excepto en accidentes como caídas o envenenamiento, los factores genómicos juegan un papel clave en **nueve de cada diez causas principales de muerte** en los Estados Unidos (fig. 1.4), por ejemplo: cardiopatías y cáncer (Murphy *et al.*, 2021).

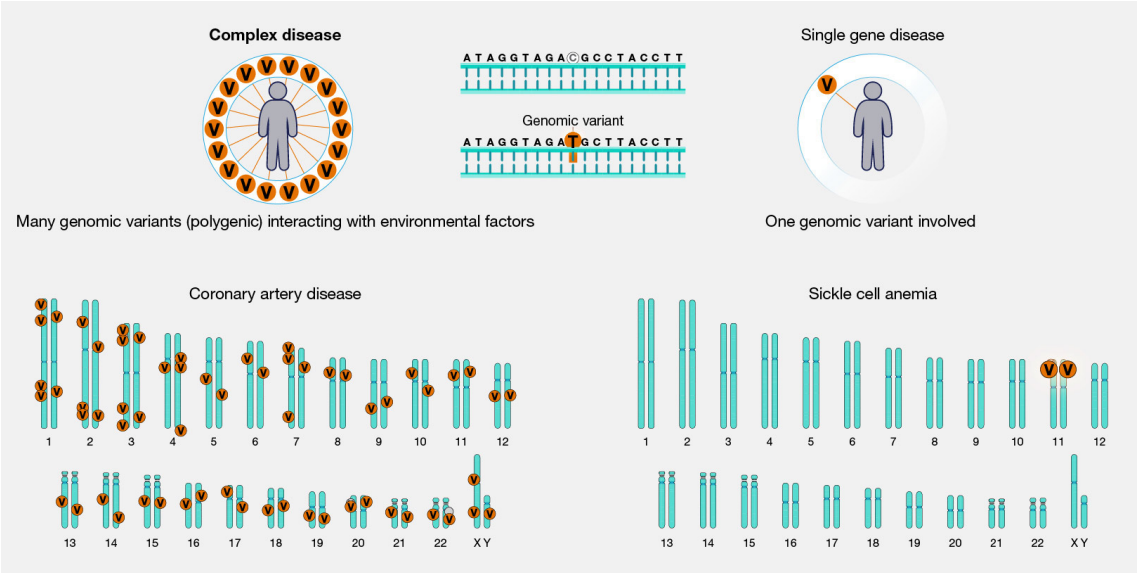


FIGURA 1.3: *Enfermedades complejas y mendelianas. Imagen del NIH.*

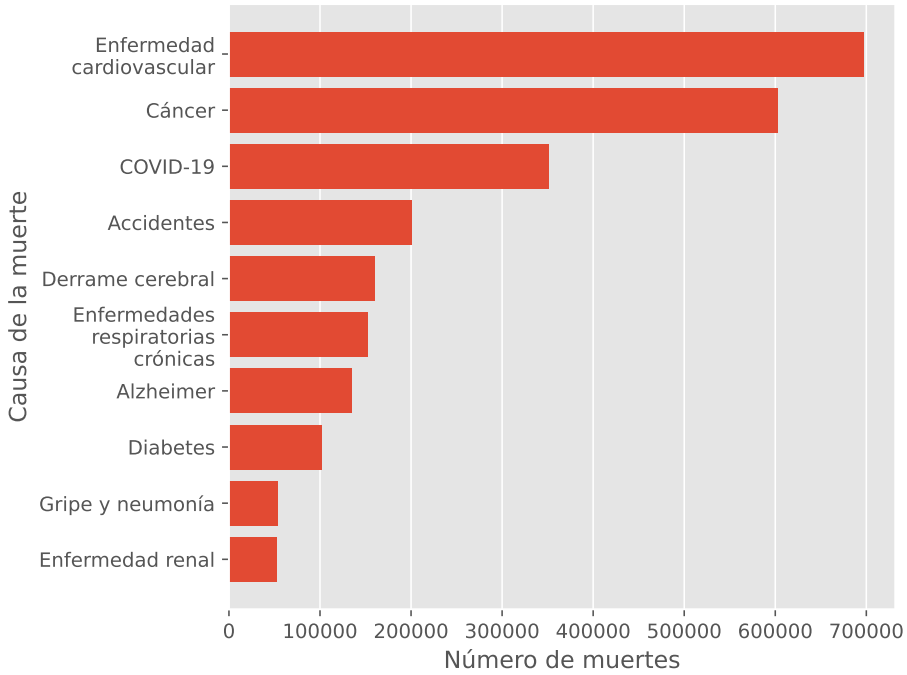


FIGURA 1.4: *Principales causas de muerte en Estados Unidos en 2020. Fuente: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm.*

1.2 Genética médica

El desarrollo de la especialidad clínica de la genética médica comenzó cuando los nuevos avances en genética fueron sistemáticamente aplicados en el campo de la medicina a finales de 1950. Se define como el estudio de la variación genética humana en relación con la salud y las enfermedades, y se centra en el diagnóstico, la prevención, y el tratamiento de los trastornos genéticos (McKusick, 1993; Passarge, 2021).

1.2.1 Subespecialidades

Actualmente, esta disciplina se divide en diferentes subespecialidades:

- La **citogenética**, desarrollada a finales de los años 1950 y definida como el estudio de la estructura cromosómica y la identificación de aberraciones genómicas que causan enfermedades (Speicher y Carter, 2005; Ferguson-Smith, 2015), se usó para describir los síndromes de Down (Lejeune, Gautier y Turpin, 1959), de Turner (Polani, Hunter y Lennox, 1954; Ford *et al.*, 1959) y de Klinefelter (Jacobs y Strong, 1959).
- La **genética de células somáticas**, el estudio de enfermedades monogénicas mediante la hibridación en cultivos celulares, apareció a mediados de los años 1960, y se usó sobre todo para el mapeado de genes en los cromosomas (Goldstein, 1971; Puck y Kao, 1982).
- La **genética molecular** estudia la variación genética a resolución de nucleótidos individuales. Se implementó a finales de los años 1970 y aplicó el mapeado de genes para diseñar sondas y marcadores de ADN específicos, además de demostrar por primera vez la existencia de polimorfismos (Kan y Dozy, 1978; McKusick, 1993).
- La **genética bioquímica**, el estudio de la base genética de las enfermedades metabólicas (Harris, 1953).

- La **inmunogenética**, el estudio de la base genética de la respuesta inmune, aplicada gracias al descubrimiento de los grupos sanguíneos (Avent y Reid, 2000).
- La **genética de poblaciones**, que describe e interpreta la variación genética en y entre poblaciones (Dobzhansky, 1937).
- La **genética mitocondrial**, que estudia las enfermedades originadas por variaciones en el ADN mitocondrial (Wallace, 2018).
- La **epigenética**, el estudio de las modificaciones que afectan directamente la expresión génica sin modificar la secuencia del ADN (Gayon, 2016).

La genética médica ha ido evolucionando a lo largo de los últimos 70 años, gracias a diversas aportaciones científicas adicionales: el análisis cromosómico para el consejo genético (Carter *et al.*, 1960), el descubrimiento de marcadores de ADN para el mapeo de asociaciones entre fragmentos de ADN (Solomon y Bodmer, 1979), la primera enfermedad mendeliana con un defecto bioquímico desconocido (enfermedad de Huntington) que se pudo mapear solo a partir de su asociación con un marcador genético (Gusella *et al.*, 1983), la farmacogenómica (Motulsky y King, 2016), etc. Todos los avances logrados están recopilados en la base de conocimientos *Mendelian Inheritance in Man* (McKusick, 1998), mantenida actualmente en la red como *Online Mendelian Inheritance in Man* (OMIM) (Amberger *et al.*, 2015).

La nueva disciplina de la genética médica impulsó también la creación de un nuevo puesto de trabajo, el del **genetista médico** o clínico. Su función principal es la de contestar a cuatro cuestiones principales: 1) qué está mal (**diagnóstico**), 2) qué pasará (**pronóstico**), 3) qué se puede hacer al respecto (**tratamiento**), y 4) por qué sucedió en primer lugar (**prevención**). Tiene que estar preparado para proporcionar un diagnóstico, consejo genético, y coordinación con los distintos servicios de salud.

1.2.2 Pruebas genéticas

Una prueba genética se define como el análisis de ADN, ARN, cromosomas, proteínas y ciertos metabolitos humanos para detectar mutaciones, genotipos, o fenotipos relacionados con enfermedades hereditarias con fines clínicos (Burke, 2002). Dichos fines se pueden clasificar en dos categorías (McPherson, 2006):

- Pruebas **predictivas** del riesgo a padecer una enfermedad. Incluyen las pruebas de identificación de portadores y de pronóstico de enfermedades.
- Pruebas **diagnósticas** para la confirmación o exclusión de una enfermedad genética. Incluyen las pruebas de diagnóstico clínico o prenatal.

La determinación de la validez y utilidad de una prueba genética viene dada principalmente por los siguientes indicadores (Altman y Bland, 1994; Burke, 2002; Mattocks *et al.*, 2010):

- La **especificidad** es la proporción de resultados negativos correctamente identificados por la prueba.
- La **sensibilidad** es la proporción de resultados positivos correctamente identificados por la prueba.
- La **penetrancia** es la proporción de personas con la variante analizada que manifiestan la enfermedad.

Las pruebas diagnósticas requieren de unos valores mínimos de especificidad y sensibilidad analíticas, determinados por los organismos competentes correspondientes (Mattocks *et al.*, 2010; FDA, 2021; Rehder *et al.*, 2021). No obstante, la realización de una prueba genética no siempre es necesaria; los reconocimientos médicos, el historial familiar, la hematología rutinaria, los estudios patológicos, y los exámenes radiológicos y electrofisiológicos también se utilizan para llegar a un diagnóstico genético (McPherson, 2006).

1.2.3 Genética molecular

La subespecialidad de la genética molecular empezó a desarrollarse a partir del descubrimiento en 1953 de la estructura de la molécula de ADN (Watson y Crick, 1953), hecho que desencadenó una serie de descubrimientos adicionales muy relevantes, como el código genético (Crick *et al.*, 1961; Yanofsky, 2007) y los modelos de expresión génica (Jacob y Monod, 1961; Gayon, 2016). Las primeras metodologías desarrolladas para su aplicación diagnóstica fueron la técnica de **secuenciación Sanger** (Sanger, Nicklen y Coulson, 1977), y la técnica de la ***Polymerase Chain Reaction* o reacción en cadena de la polimerasa (PCR)** de amplificación del ADN (Saiki *et al.*, 1986; Mullis y Faloona, 1987; Holland *et al.*, 1991).

Aunque la tecnología de secuenciación de Maxam-Gilbert de escisión química se desarrolló casi al mismo tiempo (Maxam y Gilbert, 1977), la técnica de Sanger se erigió como metodología por excelencia para validar un diagnóstico genético debido a su facilidad técnica y la exactitud de sus resultados, aunque su eficacia y rendimiento son limitadas en enfermedades de origen desconocido (Ng *et al.*, 2010; Worthey *et al.*, 2011; Jacob *et al.*, 2013), en enfermedades complejas (Kumar-Sinha y Chinnaiyan, 2018; Li *et al.*, 2022), o en enfermedades causadas por alteraciones en regiones genómicas demasiado grandes para ser secuenciadas por Sanger (Major *et al.*, 2013). Estas limitaciones se fueron paliando con la aparición de pruebas diagnósticas más modernas basadas en ***microarrays*** para genotipar polimorfismos humanos a lo largo de todo el genoma (sec. 1.2.3.1), y de tecnologías de ***next-generation sequencing* o secuenciación de última generación (NGS)** (ver sec. 1.2.3.3 más adelante). Estas últimas ofrecen mejores perspectivas, al permitir secuenciar en paralelo miles de genes, y también resolver el alineamiento de regiones genómicas con alta homología (Treangen y Salzberg, 2012; Mandelker *et al.*, 2016). La implementación clínica de la genética molecular también se vio favorecida con la creación de la primera secuencia de referencia del genoma humano (ver sec. 1.2.3.2 más adelante).

1.2.3.1 Pruebas de baja resolución

La implementación de las pruebas diagnósticas en la rutina clínica llegó a finales de la década de 1950 con el desarrollo de técnicas basadas en el análisis citogenético de los cromosomas humanos (Harper, 2007). Las primeras técnicas de diagnóstico se basaron en el análisis cromosómico de baja resolución mediante técnicas como el *banding* o el *Chromosomal Microarray Analysis* o análisis de microarrays cromosómicos (CMA), que usa sondas de ADN para detectar *Copy Number Variations* o variaciones en el número de copias (CNVs) mayores de 100,000 pares de bases (bp) (Speicher y Carter, 2005; Miller *et al.*, 2010). Esta aproximación derivó en el transcurso de los años hacia tecnologías moleculares basadas en *Fluorescence In Situ Hybridization* o hibridación in situ fluorescente (FISH) (Bauman *et al.*, 1980), como la técnica de la *Comparative Genomic Hybridization* o hibridación genómica comparativa (CGH) (Kallioniemi *et al.*, 1992), para finalmente desarrollar tecnologías de más alta resolución basadas en *microarrays*: *arrays* de ADN complementario (ADNc) (Pollack *et al.*, 1999), *arrays* de genoma completo (Snijders *et al.*, 2001), el *array painting* (Fiegler *et al.*, 2003), o *arrays* de SNPs, que tienen resolución de nucleótidos individuales (Zhao *et al.*, 2004). Cada una de estas técnicas tiene una utilidad concreta a la hora de detectar distintas tipologías de variaciones genómicas dependiendo de la cuestión biológica a resolver, y a menudo se pueden combinar para proporcionar un análisis completo de reordenamientos cromosómicos más complejos.

1.2.3.2 El genoma de referencia

Actualmente, nos encontramos en la edad de oro del estudio del genoma porque, al ser capaces de identificar la secuencia completa del genoma de una persona, se puede determinar la ubicación de los genes y sus interacciones. La localización de cada uno de los nucleótidos del ADN es posible gracias a la existencia de una **secuencia genómica de referencia**, que es una representación aceptada de la secuencia del ge-

noma humano utilizada como un estándar para la comparación con las secuencias de ADN generadas en estudios científicos y análisis clínicos. Esta secuencia de referencia no representa la secuencia genómica de una sola persona, sino una combinación de fragmentos procedentes de muestras de diferentes personas; el primer borrador fue desarrollado por el *Human Genome Project* o Proyecto del Genoma Humano (HGP) en 2001 (Lander *et al.*, 2001), y se generó mediante el ensamblaje *de novo* de lecturas largas de secuenciación (sec. 1.2.3.3), cubriendo alrededor del 99% de las posiciones cromosómicas conocidas con alta fidelidad; no obstante, estaba incompleta, y contenía muchos errores y más de 300 lagunas en la porción eucromática del genoma (Church *et al.*, 2011). Paralelamente a este esfuerzo público, la empresa Celera Genomics publicó también su referencia genómica humana (Venter *et al.*, 2001). La creación de una secuencia de referencia impulsó, además, el desarrollo de nuevos métodos de secuenciación basados en el mapeado de fragmentos cortos de ADN contra ésta (ver sec. 1.2.3.3 a continuación).

El ensamblaje genómico del HGP, considerado el primer genoma humano completo y que costó alrededor de 2.7 mil millones de dólares (National Institute of Health, 2016), se clasificó como "terminado" 3 años más tarde (International Human Genome Sequencing Consortium, 2004), pero ha sido mantenido y actualizado constantemente desde 2007 por el *Genome Reference Consortium* o consorcio de referencia del genoma (GRC) a lo largo de esta pasada década en el ensamblaje denominado GRCh37 (Church *et al.*, 2011). En 2013 el GRC publicó la nueva versión GRCh38 (Schneider *et al.*, 2017), que ha sido actualizada muy recientemente en 2022 como GRCh38.p14, en la que todavía falta casi el 8% de la secuencia completa del genoma humano (Genome Reference Consortium, 2022).

No fue hasta muy recientemente que el consorcio *Telomere-to-Telomere* (T2T), formado el 2018, finalmente publicó una secuencia que contiene una representación completa del genoma para los 22 cromosomas autosómicos y los cromosomas sexuales X e Y (Nurk *et al.*, 2022). El **ensamblaje T2T-CHM13** (actualmente en su versión

2.0, https://github.com/marbl/CHM13 y https://www.ncbi.nlm.nih.gov/assembly/GCA_009914755.4) es el resultado de secuenciar una línea celular uniformemente homocigota, y su secuencia contiene un total de 3,117,275,501 bp (es decir, alrededor de 3 Gbp). El número total de bases con potenciales problemas es solo el 0.3% de la longitud total de la secuencia, comparado con el 8% en la versión GRCh38; en este sentido, se trata una referencia más completa, precisa y representativa para la detección de variantes en muestras humanas de todas las ascendencias (Aganezov *et al.*, 2022).

1.2.3.3 Secuenciación del ADN

Para determinar los orígenes de una gran mayoría de enfermedades humanas, es necesario analizar la secuencia del ADN en busca de respuestas a los fenotipos enfermos que se manifiestan en distintas personas. Hace ya más de 40 años que se descubrió como secuenciar genomas usando metodología Sanger (sec. 1.2.3); este método de secuenciación por electroforesis dio lugar rápidamente a la secuenciación ***shotgun*** para realizar ensamblajes *de novo* de un genoma (Staden, 1979; Shendure *et al.*, 2017).

El desarrollo de la secuencia de referencia del HGP (sec. 1.2.3.2) permitió investigar alternativas a la secuenciación mediante electroforesis, dando lugar así a **tecnologías de secuenciación de segunda generación** o NGS (Metzker, 2010; Dewey *et al.*, 2012; McCombie, McPherson y Mardis, 2019), también denominada secuenciación masiva en paralelo, en la que millones de moléculas de ADN en forma de lecturas cortas (*short-reads*) se alinean contra una secuencia de referencia; la primera tecnología de este tipo apareció en 2005 (Margulies *et al.*, 2005), e introdujo además el concepto de **multiplexación**, es decir, el análisis simultáneo de diversas muestras en una única secuenciación (Ronaghi *et al.*, 1996). Las ventajas de la NGS incluyen: poder identificar de manera precisa SNVs e indels, además de una selección de *software* de análisis reputado, disponibilidad de muestras control, datos asociados para validación, y la opción de secuenciar un genoma en menos de 48 horas (Rehder *et al.*, 2021).

La mayoría de las nuevas tecnologías NGS requieren de un proceso de amplificación de fragmentos de ADN, que puede ocasionar errores en la copia, sesgos dependientes de secuencia, pérdida de información (como por ejemplo la metilación), además de añadir complejidad y tiempo de ejecución. Para superar estos obstáculos, nació la **secuenciación de tercera generación**, también llamada *Single-Molecule Sequencing* (SMS) (Check Hayden, 2009). Esta tecnología produce lecturas mucho más largas (*long-reads*) que sus predecesoras, que permiten una mejor resolución de regiones repetitivas y complejas al permitir el ensamblaje *de novo*, así como de los haplotipos, es decir, las relaciones de fase entre variantes en un genoma diploide, aunque la proporción de errores de secuenciación es mucho mayor que en Sanger o NGS, del orden de un 15-30% (Goodwin, McPherson y McCombie, 2016).

Esta evolución de tecnologías ha permitido año tras año secuenciar genomas humanos a mucho menor coste por nucleótido secuenciado (fig. 1.5); el tiempo dedicado también ha disminuido (Service, 2006), y la cantidad de datos generados está aumentando (Karczewski y Snyder, 2018).

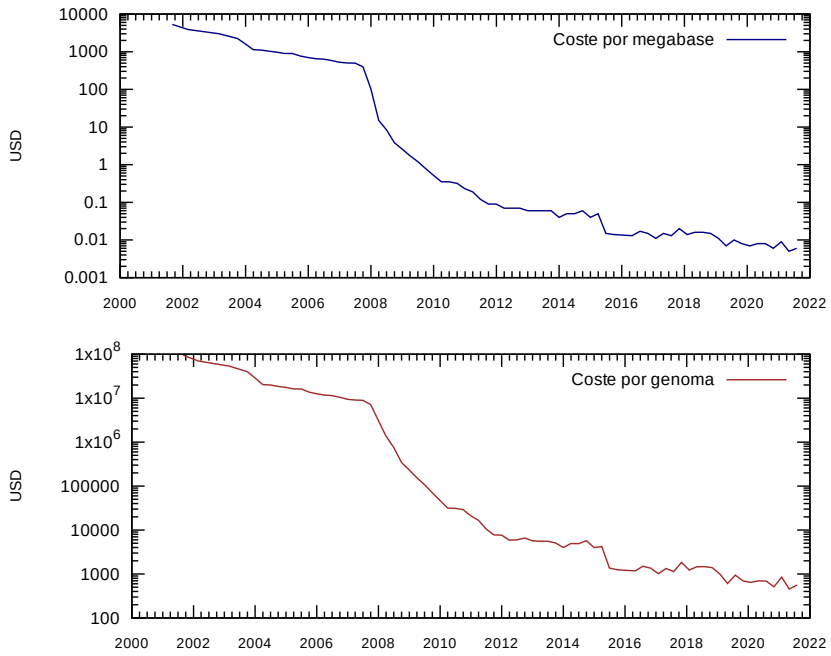


FIGURA 1.5: *Costes de secuenciación de 2001 a 2021. Datos extraídos y adaptados del NHGRI.*

1.2.3.4 Significado clínico de las variantes

A nivel clínico, una de las terminologías más usadas para reportar variantes identificadas mediante NGS en genes que causan enfermedades mendelianas, dependiendo del nivel de evidencia científica encontrado, es la publicada por el *American College of Medical Genetics and Genomics* o Colegio Americano de Genética Médica y Genómica (ACMG), que las clasifica en 5 categorías (Richards *et al.*, 2015):

- Patogénica.
- Probablemente patogénica.
- *Variants of Uncertain Significance* o variantes de significado incierto (VUS).
- Probablemente benigna.
- Benigna.

Existe además la terminología de *Clinical Genome Resource* (ClinGen), que las organiza según su nivel de penetrancia (ClinGen, 2022). Ambas aproximaciones se utilizan actualmente en ClinVar (ver sec. 1.3.1.3 más adelante) como términos estándar para reportar variantes a su base de datos (https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/). Generalmente, las variantes clínicamente relevantes se evalúan y clasifican de acuerdo con las mejores prácticas descritas en las guías del ACMG y la *Association for Molecular Pathology* o Asociación de Patología Molecular (AMP) (Green *et al.*, 2013; Richards *et al.*, 2015; Kalia *et al.*, 2017; Li *et al.*, 2017; Miller *et al.*, 2021).

1.2.3.5 Pruebas de alta resolución

Actualmente, la NGS permite diseñar pruebas diagnósticas de alta resolución, que se clasifican según su rango de detección (tbl. 1.1):

- Pruebas de **genotipado** para la detección de variantes conocidas asociadas a enfermedades o tratamiento.
- Pruebas de **paneles de genes** asociados a enfermedades o tratamientos.

- Pruebas de ***Whole-Exome Sequencing*** **o secuenciación del exoma completo (WES)**, para analizar las regiones codificantes para proteínas, aproximadamente el 1-2% del genoma.
- Pruebas de ***Whole-Genome Sequencing*** **o secuenciación del genoma completo (WGS)**, que analizan prácticamente toda la secuencia del ADN, incluyendo regiones no codificantes.

TABLA 1.1: *Comparación de plataformas de pruebas diagnósticas según su coste, capacidad de detección, tipos de variantes detectadas, descubrimientos incidentales o secundarios (sec. 1.3.5), dificultad interpretativa (sec. 1.3.1.3), y capacidad para descubrir nuevos genes. Adaptado de Rehm (2017).*

Variable	Genotipado	Paneles de genes	WES	WGS
Coste (US$)	<500	500-5,000	5,000-9,000	7,000-10,000
Detección	Baja	~5-50%	~25%	~25%
Variantes	Según diseño	SNVs y CNVs	SNVs	SNVs, CNVs y SVs
UFs y SFs	No	No	Sí	Sí
Interpretación	Fácil	Moderada	Desafiante	Desafiante
Nuevos genes	No	No	Sí	Sí

Los resultados de los paneles de genes son más fáciles de analizar e interpretar debido a su alcance limitado, aunque el hecho de que sea una lista de genes fijada no permite descubrir asociaciones de nuevos genes con la enfermedad a estudiar, como sí es posible con WES y WGS, que permiten detectar de miles a millones de variantes en un solo individuo. No obstante, cuando apareció la WGS, era demasiado costosa para la mayoría de grupos de investigación y clínicos como para adoptarla y escalar sus procesos con ella; este hecho propició el desarrollo de métodos de captura dirigidos (Hodges *et al.*, 2007) y de la WES más tarde (Ng *et al.*, 2009).

Los errores en el proceso de análisis NGS suelen ser más comunes que en los métodos de secuenciación tradicionales basados en Sanger, debido a la propia naturaleza

de la tecnología (Voelkerding, Dames y Durtschi, 2009); es por este motivo que se incrementa la **cobertura** (la profundidad de lectura) en estos experimentos, secuenciando cada posición genómica múltiples veces. La WES requiere del enriquecimiento de las regiones codificantes por métodos de captura o amplificación, y contiene la mayoría de SNVs y pequeñas indels actualmente identificadas como causantes de enfermedades mendelianas; no obstante, **la cobertura para WES no es uniforme**, por lo que la sensibilidad analítica puede ser inferior a algunos paneles de genes dirigidos a enfermedades concretas, aunque se asume que la WES tiene un mayor rendimiento diagnóstico a nivel global (Rehder *et al.*, 2021).

La WGS examina casi el genoma completo y, a diferencia de la WES, no requiere de métodos de enriquecimiento antes de la secuenciación, por lo que produce una cobertura más uniforme a través del exoma; además, los datos se pueden producir más rápidamente, y tiene mayor capacidad para detectar simultáneamente SNVs y CNVs, así como variantes complejas como SVs y expansiones repetitivas. Sin embargo, la cobertura es generalmente menor que en los paneles de genes de diagnóstico y en la WES, de modo que puede limitar la detección de mosaicismo, y el coste de la generación y almacenamiento de datos es también mayor (Rehder *et al.*, 2021).

1.2.3.6 Paneles de genes diagnósticos

Dentro de las pruebas de alta resolución, el panel de genes diagnóstico es un análisis diseñado para las características clínicas específicas de un paciente, que examina una lista de genes seleccionados por su asociación con un fenotipo particular. Su objetivo es el de maximizar la sensibilidad y especificidad clínicas (sec. 1.2.2), limitando el número de VUS identificadas, que son mucho más frecuentes en WES y WGS, minimizando así la carga clínica derivada de analizar genes irrelevantes o innecesarios para el diagnóstico (Bean *et al.*, 2020). La optimización de ambos indicadores depende en parte de la cobertura que se consiga en las regiones objetivo a estudiar y los tipos de variantes que se puedan detectar; al centrarse en un pequeño conjunto de genes, el

coste para lograr una cobertura genómica adecuada es reducido a través de la utilización eficiente de la capacidad de secuenciación y la disminución de los requerimientos en poder computacional y de almacenamiento de datos (Oliver, Hart y Klee, 2015). En consecuencia, esta cobertura supone una limitación en WES y especialmente en WGS, en las que la eficiencia y precisión de secuenciación es más variable, generando así regiones genómicas con insuficiente cobertura y, por lo tanto, incrementando la carga de validación de variantes identificadas en esas regiones para el profesional bioinformático y clínico (Yu *et al.*, 2012).

Además, la sensibilidad clínica se puede aumentar no solo evaluando las regiones codificantes y no codificantes de los genes objetivo, sino también incorporando al análisis pruebas auxiliares como la secuenciación Sanger, aunque se ha demostrado recientemente que en la gran mayoría de casos no es necesaria (Beck, Mullikin y NISC Comparative Sequencing Program, 2016; Arteche-López *et al.*, 2021). La especificidad clínica se puede maximizar limitando o excluyendo genes con evidencia limitada o incierta en relación con el fenotipo, minimizando así la detección de VUS (Rehder *et al.*, 2021).

La evolución de las tecnologías de secuenciación de ADN (sec. 1.2.3.3) ha hecho que los paneles de genes sean actualmente más flexibles y coste-eficientes. La NGS se ha implementado en laboratorios clínicos como principal modalidad de prueba diagnóstica en medicina genómica. En consecuencia, dichas pruebas están pasando de interrogar variantes patogénicas conocidas en uno o unos pocos genes (a través de la secuenciación Sanger o análisis de genotipado dirigido) a análisis en paralelo de conjuntos más grandes de genes usando NGS. Los laboratorios de diagnóstico molecular ofrecen hoy en día una gran variedad de pruebas, desde paneles de una sola enfermedad hasta paneles de múltiples enfermedades, que se pueden consultar en bases de datos centralizadas (Rubinstein *et al.*, 2013).

Aunque la NGS facilita mucho el desarrollo de pruebas genéticas, su mayor alcance genómico dificulta el proceso de validación desde una perspectiva diagnóstica. La

validación adecuada de estos paneles de genes mucho más grandes implica demostrar por parte del laboratorio la capacidad para identificar variantes clínicamente relevantes y de múltiples tipos sin importar las dificultades técnicas (Rehder *et al.*, 2021); para hacerlo, se deben evaluar todos los pasos de la prueba, desde las máquinas de extracción y secuenciación de ADN usadas para generar datos sin procesar (es decir, en bruto, la *raw data*), hasta los *pipelines* bioinformáticos utilizados para interpretar las variantes (Wilcox *et al.*, 2021).

Históricamente, las referencias utilizadas para validar las pruebas genéticas se han desarrollado pensando en pequeños paneles de genes, de manera que actualmente existe también este obstáculo para la implementación de pruebas genómicas y paneles grandes de genes; los conjuntos de datos existentes no cumplen con los criterios para validar variantes clínicamente relevantes debido al hecho de que las muestras se generaron a partir de individuos aparentemente sanos. En consecuencia, es poco probable que las muestras contengan variantes patogénicas en genes incluidos en análisis exhaustivos de NGS, hecho que las hace insuficientes para evaluar comparativamente la amplia gama de variantes normalmente incluidas en este tipo de paneles. Además, proporcionar suficientes muestras de referencia para variantes específicas en paneles grandes de genes que abarcan varias enfermedades puede ser **virtualmente imposible**, siendo como ya es un desafío para los paneles más pequeños (Wilcox *et al.*, 2021).

Los paneles de genes de diagnóstico concebidos como pruebas individuales exhaustivas pueden ser inadecuados para analizar enfermedades con extrema heterogeneidad genética. Para tales enfermedades y en ciertos escenarios, secuenciar todo el exoma o genoma como parte de una prueba inicial no diagnóstica, con la opción de recomendar más tarde paneles amplios, dirigidos a múltiples enfermedades con un objetivo diagnóstico, podría ser una mejor opción (Bean *et al.*, 2020).

1.2.4 Medicina personalizada

El desarrollo de las tecnologías NGS (sec. 1.2.3.3) han permitido la implementación en la rutina clínica de nuevas técnicas diagnósticas (sec. 1.2.3.5) que hacen uso de la información genómica de una persona como parte de su atención médica, favoreciendo así la aparición de la **medicina personalizada**, una disciplina que da un tratamiento especializado para cada paciente basándose en la disponibilidad de una gran cantidad de datos generados con diferentes tecnologías ómicas (ver sec. 1.2.4.3 más adelante), en contraposición al enfoque que tiene la medicina tradicional (Schork, 2015).

Para algunas enfermedades humanas, la información genómica se puede usar para ayudar a diagnosticarlas, predecir sus consecuencias, desarrollar tratamientos dirigidos, o para su prevención. Esa información **accionable** es solo una parte del rompecabezas de por qué algunas personas contraen una enfermedad y otras no, pero es una información que podemos analizar con mucha precisión; otros factores influyen, como los hábitos personales de cada uno, así como elementos ambientales posiblemente perjudiciales a las que estamos sometidos durante nuestra vida (Ashley *et al.*, 2010).

1.2.4.1 Ventajas y desafíos

La medicina personalizada dispone actualmente de una serie de ventajas relevantes sobre la medicina tradicional para su implementación en la rutina clínica:

- Proporciona un enfoque más **coste-eficiente y exhaustivo** que el diagnóstico tradicional, en el que se realizan múltiples pruebas individuales (Stark *et al.*, 2017; Payne *et al.*, 2018; Li *et al.*, 2020; Yeung *et al.*, 2020; Kosaki *et al.*, 2020).
- Permite efectuar nuevos diagnósticos mediante el **reanálisis de datos** procesados anteriormente (Wright *et al.*, 2018; Stark, Schofield, *et al.*, 2019); ver sec. 1.4.3 más adelante.
- Permite desarrollar una **medicina integral** combinando múltiples tecnologías ómicas en una sola persona para crear una visión holística de los efectos mole-

culares que conducen a ciertos fenotipos (ver sec. 1.2.4.3 más adelante).

- Permite generar **modelos predictivos de riesgo a enfermedades** enfocados a personas sanas mediante los *Polygenic Risk Score*s (PRSs) (Khera *et al.*, 2018).

No obstante, existen también obstáculos para su implementación generalizada:

- La dificultad en la **interpretación masiva** de los datos generados (Cheon, Mozersky y Cook-Deegan, 2014; Whiffin *et al.*, 2017).
- La capacidad de **almacenamiento** para la ingente cantidad de datos generados y la capacidad de **escalabilidad** y de **poder computacional** (Stein, 2010; Berger, Peng y Singh, 2013).
- La capacidad y habilidad del **personal clínico** gestionando la información generada (Rehder *et al.*, 2021).
- En algunos casos, una **menor especificidad y sensibilidad** comparadas con los métodos tradicionales (Rehder *et al.*, 2021).
- Problemas de **interferencia homóloga** en ciertas regiones genómicas (Rehm *et al.*, 2013; Mandelker *et al.*, 2014, 2016).
- La **falta de evidencia genómica con utilidad clínica**, ya que existe una menor disponibilidad de datos de secuenciación clínica en comparación con lo que existe en el ámbito de la investigación (Manolio *et al.*, 2015).
- La integración con ***Electronic Health Records* o historiales clínicos electrónicos (EHRs)** (Global Health Observatory, 2018a, 2019); ver sec. 1.5.3 más adelante.
- Aspectos éticos, como la **falta de diversidad e inclusividad de minorías**, ya que alrededor del 89% de participantes en estudios genómicos son de ascendencia europea (Landry *et al.*, 2018; Mills y Rahal, 2020).
- La **aceptación pública** y la **participación de los gobiernos** en su implementación y legislación (Global Health Observatory, 2018b, 2018c; Stark, Dolman, *et al.*, 2019).

1.2.4.2 Implementación actual

La aplicación clínica de la NGS para la resolución de diagnósticos sin resultados concluyentes empezó a ser una realidad hace ya más de 10 años (Choi *et al.*, 2009; Ashley *et al.*, 2010; Worthey *et al.*, 2011); uno de los primeros casos de éxito fue el de Nic Volker (ver sec. 1.4.1 más adelante). **En el ámbito público**, varios países han implementado iniciativas para secuenciar los genomas clínicos de grandes poblaciones (Stark, Dolman, *et al.*, 2019), como es el caso del proyecto 100,000 genomas impulsado por *Genomics England* (Peplow, 2016), y se prevé que más de 60 millones de personas en todo el mundo tendrán su genoma secuenciado para 2025 (Birney, Vamathevan y Goodhand, 2017); organizaciones como la *Global Genomic Medicine Collaborative* (G2MC) y el Foro Económico Mundial están facilitando colaboraciones globales respecto la implementación de pruebas genómicas en entornos clínicos (Ginsburg, 2019; World Economic Forum, 2020).

La integración en aumento de la genómica en los sistemas de salud pública se refleja en la explosión del uso de la genómica en el **sector privado**, particularmente en los Estados Unidos. El proyecto *MyCode* de *Geisinger Health System*, que comenzó como una colaboración con *Regeneron Pharmaceuticals* para realizar la secuenciación del exoma en 100,000 pacientes de *Geisinger* y utilizar los resultados para el descubrimiento de fármacos y la atención clínica (Carey *et al.*, 2016), se ha ampliado recientemente a todos los pacientes que dan su consentimiento (Stark, Dolman, *et al.*, 2019). También han surgido empresas de pruebas *direct-to-consumer* (DTC), como *23andMe* y *Ancestry*, que capturan información genómica importante relacionada con la salud, aunque la respuesta del público y de los médicos con respecto a su utilidad clínica han sido polémicas (Levenson, 2016; Roberts *et al.*, 2017; Tandy-Connor *et al.*, 2018).

1.2.4.3 Tecnologías integrales ómicas

Las diferentes aplicaciones ómicas de las tecnologías NGS (transcriptoma, proteoma, metaboloma, microbioma, etc.) permiten usar los datos generados de manera combinada para realizar análisis integrales en un paciente, que vayan más allá del análisis de su secuencia genómica (Karczewski y Snyder, 2018). Esta aproximación integral permite capturar la complejidad de las enfermedades humanas debido a que múltiples tecnologías ómicas pueden establecer una cadena de causalidad que una sola tecnología no puede (Chen *et al.*, 2012; Kremer *et al.*, 2017; Cummings *et al.*, 2017; Piening *et al.*, 2018; Rodgers y Collins, 2020). Sin embargo, una de sus limitaciones más importantes en la actualidad es la falta de conjuntos de datos grandes y estructurados que faciliten su aplicación clínica (Karczewski y Snyder, 2018).

Uno de los primeros estudios en realizar un *integrative Personal Omics Profile* o perfil ómico personal integral (iPOP) de un individuo a lo largo de etapas vitales sanas y enfermas se publicó en 2012 (Chen *et al.*, 2012); combinó perfiles genómicos, transcriptómicos, proteómicos, metabolómicos e inmunológicos de un voluntario de 54 años de edad durante un período de 14 meses, y reveló varios riesgos médicos, incluida la diabetes de tipo 2. En general, la observación de sus cambios dinámicos moleculares ofreció una **prueba piloto de medicina personalizada**. Posteriormente, otro estudio del mismo grupo de investigación remarcó la importancia de realizar perfiles multiómicos a nivel individual, para generar una gran cantidad de datos personales y longitudinales que permitan la creación de un mapa detallado de los cambios moleculares individuales que ocurren en respuesta a diferentes fenotipos (Piening *et al.*, 2018).

1.3 Genómica y bioinformática

Gracias a la NGS, el desafío actual de la genómica clínica ya no es la generación de datos, sino el poder computacional de la bioinformática para analizarlos, visualizarlos, interpretarlos de forma masiva e integrarlos en los sistemas hospitalarios (Rehm, 2017; Stark, Dolman, *et al.*, 2019). La bioinformática es una disciplina que desarrolla y aplica herramientas computacionales avanzadas para analizar e interpretar datos biológicos de grandes dimensiones. Es un campo en continuo desarrollo debido a la propia naturaleza y velocidad de aparición de tecnologías NGS (Stein, 2010; Berger, Peng y Singh, 2013); es común que las herramientas computacionales que se diseñan para tales fines sean rápidamente desfasadas a la hora de analizar las grandes cantidades de datos genómicos generados y las diferentes tipologías de variantes que se pueden detectar. Actualmente existen multitud de herramientas, tanto de código abierto como soluciones comerciales, para procesar e interpretar datos NGS e intentar mantener el constante ritmo de innovación de la tecnología (Oliver, Hart y Klee, 2015), como el *Genome Analysis Toolkit* (GATK) por ejemplo (McKenna *et al.*, 2010; DePristo *et al.*, 2011).

Las tecnologías NGS definen un amplio conjunto de métodos para llevar a cabo la preparación de librerías genómicas previa a la secuenciación, la generación masiva de lecturas cortas de ADN, su alineación y ensamblaje contra la secuencia de referencia, y la identificación de variantes que difieren de ésta (Dewey *et al.*, 2012). Estos procesos se pueden clasificar en tres etapas analíticas: análisis primario, secundario y terciario (fig. 1.6), detalladas en las secs. 1.3.1.1, 1.3.1.2, 1.3.1.3; cada una de ellas engloba un paso necesario para transformar señales a datos de secuenciación en bruto, datos en bruto a información interpretable, e información a conocimientos clínicamente accionables.

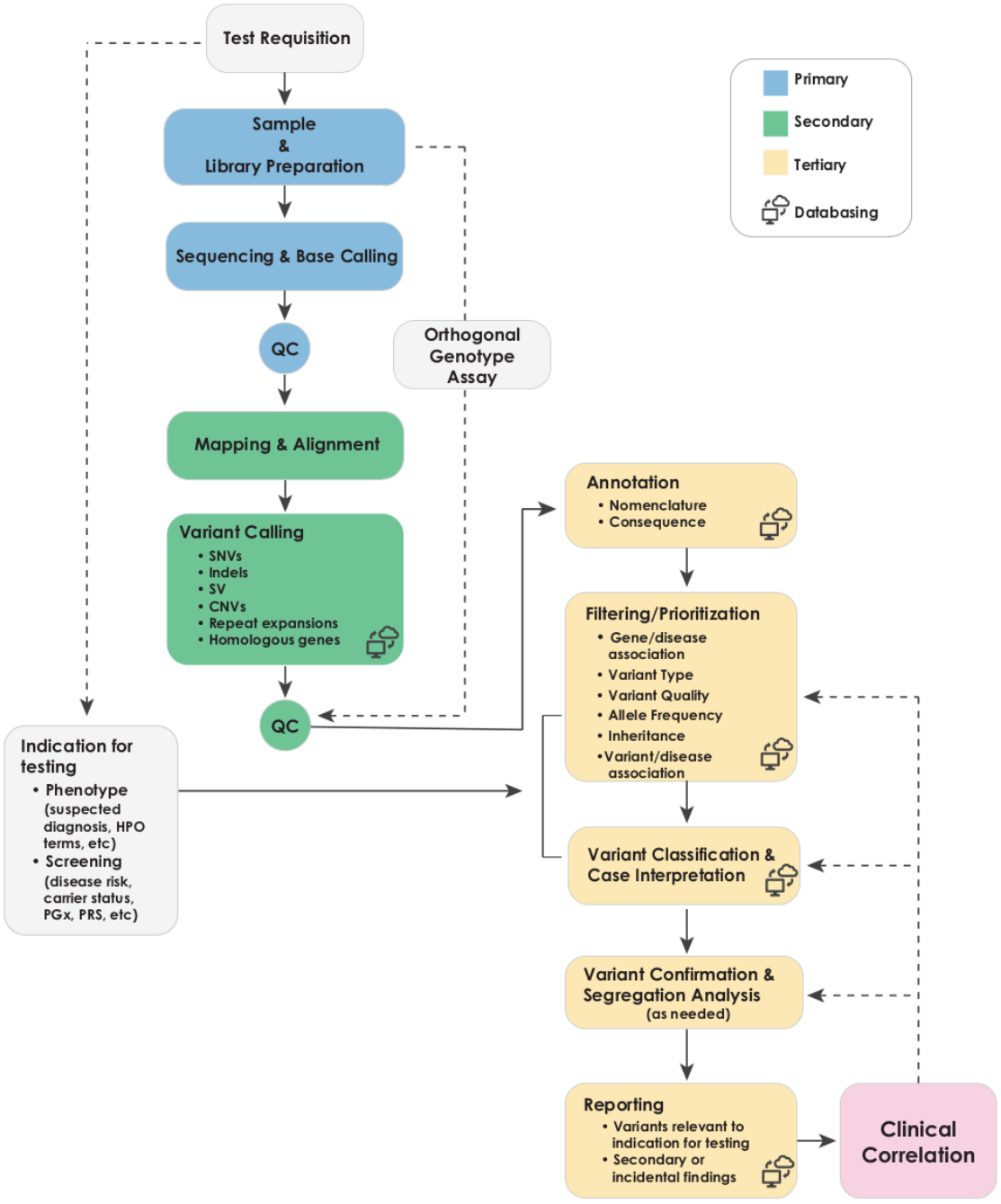


FIGURA 1.6: *Etapas de secuenciación de un genoma clínico completo. Fuente: Marshall* et al. *(2020).*

1.3.1 Etapas analíticas

1.3.1.1 Análisis primario

El análisis primario (fig. 1.6) incluye el aspecto técnico de la secuenciación: extracción de ADN, preparación de librerías, generación de lecturas y control de calidad de datos mediante la asignación de puntuaciones de calidad a cada base identificada (Oliver, Hart y Klee, 2015). Este análisis es un proceso muy integrado a las máquinas de secuenciación y al *software* comercial que usan cada una de ellas. El *software* transforma las señales generadas por las máquinas, que se guardan en forma de ficheros de formato *binary base call* (BCL) (Liu *et al.*, 2012), en bases nucleotídicas que llevan una puntuación de calidad asociada, y como datos de salida generan lecturas cortas de ADN en un fichero FASTQ, el formato más común de entrada para *pipelines* bioinformáticos de análisis NGS (Roy *et al.*, 2018). En la mayoría de casos el análisis primario incluye también la **demultiplexación**, la asociación computacional de esas lecturas con una muestra concreta después de la secuenciación multiplexada.

1.3.1.2 Análisis secundario

El análisis secundario (fig. 1.6) está constituido por el proceso bioinformático de ensamblaje de las regiones genómicas secuenciadas, que incluye la alineación de las lecturas obtenidas contra un genoma de referencia, la detección de variantes genómicas a partir de las puntuaciones de calidad de cada fragmento, y más operaciones de control de calidad para corregir el sesgo técnico de las máquinas de secuenciación:

- El **alineamiento** de las lecturas contra el genoma de referencia se almacena en ficheros en formato *Binary Alignment Map* (BAM) (Li *et al.*, 2009); existen además varios pasos de refinamiento que incluyen el marcaje o filtraje de lecturas duplicadas y el realineamiento, que usa la visión colectiva de las lecturas en torno a los sitios putativos de indels para minimizar la alineación errónea de los extremos de cada lectura (Oliver, Hart y Klee, 2015). Existen también

errores derivados de la máquina de secuenciación, que se pueden anticipar y corregir para priorizar aquellas variantes con un mayor grado de confianza (Glenn, 2011).

- La **identificación de variantes o *variant calling*** es la comparación de las lecturas alineadas con la referencia para identificar las variaciones genómicas, cuya información se guarda en ficheros en formato *variant call format* (VCF), que soportan el almacenado de múltiples muestras (Danecek *et al.*, 2011), o en formato *genome variant format* (GVF), cuyo uso está menos extendido (Reese *et al.*, 2010). Dependiendo del protocolo empleado, la detección de variantes de diferentes tipologías se producirá a nivel del genoma entero, del exoma o de un panel de genes dirigido (sec. 1.2.3.5), mediante el uso de ficheros en formato BED que limiten la amplitud del análisis (Quinlan, 2014).

1.3.1.3 Análisis terciario

El análisis terciario (fig. 1.6) involucra todo el proceso bioinformático interpretativo, que da contexto a la información generada en un experimento NGS, y está constituido por los procesos de anotación, filtraje, priorización y clasificación de variantes, y la interpretación de las variantes clínicamente relevantes (Oliver, Hart y Klee, 2015):

- La **anotación** de las variantes identificadas (Stein, 2001) implica definir el impacto previsto a nivel genético de las variantes según una nomenclatura estandarizada (Den Dunnen *et al.*, 2016; Wagner *et al.*, 2021). Este proceso determina qué variantes se seleccionan para la revisión de analistas durante el proceso posterior de clasificación, y se encarga de priorizarlas funcionalmente consultando diversas bases de datos (Oliver, Hart y Klee, 2015):
  - Las **anotaciones basadas en frecuencia poblacional** se usan para identificar aquellas variantes con mayor frecuencia alélica en la población (polimorfismos), hecho que las hace potencialmente benignas y, por lo tanto, menos

proclives a tener relevancia clínica. Para eliminar los polimorfismos de las listas de variantes candidatas, se usan **umbrales de frecuencia o *cutoffs***.

- Las **anotaciones estructurales** se utilizan para asignar el efecto de la variante a los transcritos y proteínas codificadas en ese gen concreto, basándose en el cambio de aminoácidos producido; por ejemplo, las mutaciones sin sentido (*nonsense*) se clasifican como de gran impacto y, por lo tanto, se priorizan en la lista de variantes candidatas.
- Las **anotaciones predictivas** se usan para inferir el impacto de una variante en el producto resultante de la secuencia afectada, a partir de cambios de nucleótidos o aminoácidos junto con información adicional contextual (por ejemplo, conservación evolutiva o impacto en estructuras proteicas simuladas en 3D), aunque la especificidad y sensibilidad de este tipo de aproximaciones es reducida (Thusberg, Olatubosun y Vihinen, 2011).
- Las **anotaciones basadas en evidencia científica** tienen una alta especificidad y sensibilidad, ya que se derivan de la literatura existente y otros datos históricos, como los catalogados y curados en la *Human Gene Mutation Database* (HGMD) (Stenson *et al.*, 2008), OMIM (Amberger *et al.*, 2015), dbSNP (Sherry, Ward y Sirotkin, 1999) y dbVar (Lappalainen *et al.*, 2013), y centralizadas en ClinVar, una base de datos dinámica (Landrum *et al.*, 2014); es una base de datos crítica utilizada por los laboratorios para descargar archivos estáticos como instantáneas en el tiempo que luego se utilizan en los procesos de anotación.

- El **filtraje y priorización** de las variantes consiste en reducir la lista de variantes con características que tienen más probabilidades de causar una enfermedad genética, y priorizarlas a partir de una serie de criterios que definan el orden en que los analistas las revisarán para evaluar su relevancia clínica (Rehder *et al.*, 2021). En este sentido, existen aproximaciones basadas tanto en el genotipo

como en el fenotipo para optimizar este equilibrio:

  - Los **análisis basados en el genotipo** pueden aplicar filtros adicionales para reducir aún más la lista de variantes que se tienen que reportar, utilizando herramientas como OMIM (Amberger *et al.*, 2015) y la *Gene Curation Coalition* (https://thegencc.org/) para centrarse en variantes dentro o cerca de los genes vinculados a una enfermedad genética. Los *cutoffs* de frecuencia alélica también se pueden usar para el mismo propósito, usando datos de frecuencia poblacional de bases de datos de referencia como *gnomAD* para excluir variantes demasiado comunes para causar enfermedades genéticas raras (Karczewski *et al.*, 2020).
  - En los **análisis basados en el fenotipo** es más probable que las variantes identificadas se clasifiquen como benignas o VUS, ya que carecen de evidencia previa de patogenicidad ni de predicción de un impacto en la *loss-of-function* o pérdida de función (LOF), que surgirían más de los análisis basados en el genotipo (Rehder *et al.*, 2021).

- La **clasificación** de variantes es el proceso mediante el cual los laboratorios inician la toma de decisiones con respecto a la relevancia de las variantes seleccionadas en relación con la indicación clínica principal para el análisis (sec. 1.2.3.4); su objetivo es limitar aún más la lista de variantes reportables que cumplan los criterios para el informe final de resultados y, por lo general, implica una revisión por parte de un panel de expertos de la evidencia disponible que respalde o niegue la patogenicidad de la variante y su asociación con la enfermedad y el fenotipo del paciente (Rehder *et al.*, 2021).
- La **interpretación** de variantes constituye el proceso por el cual se procede a realizar un análisis menos automatizado y en profundidad de genes y variantes relevantes, antes de tomar decisiones médicas que afecten al paciente. Este paso representa el más costoso en cuanto a tiempo y recursos, con un **promedio de**

7.3 horas por cada análisis realizado (Austin-Tse *et al.*, 2022). El uso de bases de datos de variantes de grado clínico facilita mucho la correcta interpretación de resultados por parte de los analistas clínicos (ver sec. 1.3.3 más adelante).

- Finalmente, existe el proceso de **reporte de resultados** del análisis al paciente o médico involucrado mediante un informe clínico (ver sec. 1.5.2 más adelante).

1.3.2 Validación clínica del pipeline bioinformático

La mayoría de los pasos del análisis clínico de una WGS o WES requieren de procesos bioinformáticos que permiten el análisis masivo de los datos generados y su automatización. La implementación de un *pipeline* bioinformático que encapsule cada una de las etapas analíticas es compleja, debido a los requerimientos de *hardware* y personal especializado que sepa manejar y evaluar la correcta parametrización de todos los pasos involucrados. Actualmente, existe una gran variedad de soluciones comerciales y de código abierto para cada una de las etapas, con sus ventajas e inconvenientes dependiendo de la plataforma de secuenciación y el protocolo seleccionado (Moorthie, Hall y Wright, 2013; Pabinger *et al.*, 2014); a causa también del alto grado de personalización que se puede lograr, se ha observado que la concordancia entre *pipelines* alternativos no es muy alta (O'Rawe *et al.*, 2013). La disponibilidad tan variable de soluciones y la falta de estándares establecidos dificultan la selección de las herramientas adecuadas para cada análisis clínico específico. En consecuencia, como en todo proceso con un resultado clínico, existe una validación de los *pipelines* bioinformáticos en las que, además, el laboratorio debe documentar todo el *hardware*, *software*, bases de datos, incluidas las versiones utilizadas, y sistemas adicionales desarrollados internamente, e incluir cualquier modificación o versiones que se correspondan para una correcta trazabilidad, que se utilizan para validar el *pipeline* entero (Rehm *et al.*, 2013; Roy *et al.*, 2018; Rehder *et al.*, 2021); esta validación refuerza que todo el proceso analítico esté estandarizado y sea reproducible, e incluye la determinación de la sensi-

bilidad, especificidad, y precisión de todos los tipos de variantes identificados. Existen diferentes instituciones que definen estos requerimientos clínicos y garantizan su cumplimiento para desarrollar unas recomendaciones de buenas prácticas y aseguren unas pruebas clínicas precisas y seguras para los pacientes (Oliver, Hart y Klee, 2015). Para validar el rendimiento y eficiencia de la detección de variantes de cada *pipeline*, se pueden usar conjuntos de datos sintéticos como control, aunque se recomienda el uso de muestras biológicas, como las establecidas por el proyecto internacional HapMap (The International HapMap Consortium, 2003) o las del proyecto *Genome in a Bottle* (GIAB), considerado uno de los mejores estándares de referencia (Zook *et al.*, 2014).

1.3.3 Bases de datos genómicos de grado clínico

Los datos NGS se generan y analizan de una manera muy diferente en comparación con las pruebas tradicionales de laboratorio, lo que implica nuevos obstáculos para analizarlos, interpretarlos y reportarlos. Una vez los datos son procesados (sec. 1.3.1), la falta de información curada por profesionales para respaldar la toma de decisiones clínicas y la gran cantidad de datos generados dificultan la estandarización de la interpretación de las variantes identificadas. Este problema requiere de la necesidad de un soporte en forma de base de datos que ayude a los profesionales del laboratorio y a los médicos involucrados en el proceso de toma de decisiones; específicamente, se requiere de una ***Clinical Grade Genomic Database*** **o base de datos genómicos de grado clínico (CGGD)**, para asegurarse de que toda la información contenida en ella se haya obtenido de resultados NGS generados bajo estándares de calidad clínica, agilizando así la implementación de la NGS en laboratorios clínicos (Yohe *et al.*, 2015). Idealmente, la CGGD contiene ***High-Quality Human Sequences/Variants*** **o secuencias/variantes humanas de alta calidad (HQHSVs)**: secuencias y/o variantes producidas a partir de muestras humanas en un laboratorio que cumpla con los estándares de calidad clínica para el análisis que las genera. La información contenida en

la CGGD se puede clasificar en 3 capas o categorías:

1. Los *Clinical Genomic Variant Repositories* o repositorios de variantes genómicas clínicas (CGVRs), que contienen datos de secuenciación. Un ejemplo de este tipo de base de datos la encontramos en el proyecto 1000 Genomas (The 1000 Genomes Project Consortium, 2010).
2. Los *Genomic Medical Data Repositories* o repositorios de datos médicos genómicos (GMDRs), con información clínica y fenotípica, excluyendo la ocurrencia y los resultados entre pacientes con las mismas variantes. Un ejemplo de base de datos que incluye este tipo de información es COSMIC (Bamford *et al.*, 2004).
3. Las *Genomic Medical Evidence Databases* o bases de datos de evidencia genómica (GMEDs), que contienen información de clasificación o asociación. Esta es la capa más útil con respecto a la interpretación de los resultados NGS de un paciente. ClinVar (Landrum *et al.*, 2014) sería un ejemplo de base de datos que incluye este tipo de información; sin embargo, carece de garantías sobre la calidad de los datos de las secuencias generadas, así como de los estándares utilizados para la evidencia médica de utilidad clínica (Yohe *et al.*, 2015).

Los datos enviados a una CGGD pueden tener principalmente tres formatos diferentes: FASTQ, BAM, y VCF. Para su recuperación se recomienda el uso de una interfaz web con un formato estandarizado para permitir la interacción con otras bases de datos, y de un mecanismo para automatizar la correlación de la información de secuenciación en la CGGD con datos del laboratorio o del sistema de EHRs, y para recuperar automáticamente información de todas las variantes detectadas en un paciente (Yohe *et al.*, 2015). En cuanto a privacidad y seguridad, la CGGD debe cumplir con los requisitos para el cumplimiento de las leyes locales o regionales del país donde se almacena, aunque pueda contener datos de pacientes de otros países; muchas bases de datos actuales se desarrollaron sin tener en cuenta estos requisitos. Además, los

datos solo deben enviarse a la CGGD si el laboratorio dispone del consentimiento informado del paciente (Yohe *et al.*, 2015).

1.3.4 Reclasificación de variantes y reanálisis de datos

Con la base de conocimientos de variantes en rápida evolución, se han informado altas tasas de reclasificación de variantes identificadas (Deignan *et al.*, 2019; Machini *et al.*, 2019), por lo que existen recomendaciones con respecto a que los laboratorios ofrezcan el reanálisis de datos de pruebas genéticas previamente reportadas. Dichas recomendaciones incluyen (Rehder *et al.*, 2021):

- Una descripción de los procedimientos de reclasificación con respecto a la variante y reanálisis en cuanto al estudio.
- Una indicación sobre si se aplican cargos adicionales.
- Una declaración sobre el reanálisis en la descripción de la prueba y en los informes individuales.
- La petición por parte de los laboratorios a los proveedores de atención médica de realizar consultas periódicas para determinar si el conocimiento sobre una variante reportada anteriormente ha cambiado y ha dado lugar a una reclasificación, como una VUS reclasificada como patogénica.

1.3.5 Descubrimientos adicionales

Durante un análisis WES o WGS de una persona, los laboratorios pueden reportar dos tipos de descubrimientos adicionales:

- Los ***Unsolicited Findings* o descubrimientos no solicitados (UFs)**, también llamados *Incidental Findings* o descubrimientos incidentales (IFs) (Van Ness, 2008), definen eventos **inesperados** durante el análisis; la *European Society of Human Ge-*

netics o Sociedad Europea de Genética Humana (ESHG) ha sugerido que UF es un término descriptivo más apropiado que IF (El *et al.*, 2013).

- Los ***Secondary Findings* o descubrimientos secundarios (SFs)** definen eventos que no están relacionados con la indicación principal de la prueba genética (Johnston *et al.*, 2012), pero que pueden ser clínicamente relevantes para la salud del paciente, es decir, variantes que se espera que causen una enfermedad y, en consecuencia, se intentan identificar **de manera activa**. Esta última situación también es definida por la ESHG como *Opportunistic Genomic Screening* o cribado genómico oportunista (OGS) (Wert *et al.*, 2021), un concepto más tradicional entendido como la búsqueda deliberada de variantes genéticas no relacionadas con la pregunta diagnóstica principal (Wilson y Jungner, 1968; Andermann *et al.*, 2008), que conlleva cierta polémica actualmente (ver sec. 1.5.4 más adelante).

1.3.6 Flujo del diagnóstico clínico

Actualmente, el proceso por el que pasa cada paciente al que se le realiza un análisis NGS para el diagnóstico clínico de una enfermedad de origen genético está formado por una serie de etapas que, de manera general, se describen a continuación (fig. 1.7):

- El médico encargado del caso solicita una sesión de consejería genética con el paciente para que se le informe sobre las ventajas y consecuencias del análisis NGS que se le realizará (Green *et al.*, 2013; Bowdin *et al.*, 2016).
- Una vez firmado el consentimiento informado por parte del paciente, se encarga al laboratorio la prueba diagnóstica específica para su situación (sec. 1.2.3.5).
- El paciente se dirige al laboratorio correspondiente para que le realicen la extracción de muestras necesarias para el análisis genómico.
- Se realiza el procesamiento bioinformático de los datos generados durante la secuenciación (secs. 1.3.1.1, 1.3.1.2).

- Se realiza la curación e interpretación de las variantes detectadas (sec. 1.3.1.3).
- Se reportan los resultados clínicamente relevantes al médico (sec. 1.5.2); en caso de que no sean concluyentes, **se repite el circuito entero** hasta dar con un diagnóstico definitivo para la pregunta clínica principal.

FIGURA 1.7: *Flujo diagnóstico de los datos NGS, desde la sesión de consejería genética para obtener el consentimiento informado del paciente, hasta el reporte de resultados clínicamente relevantes.*

1.4 Paneles virtuales de genes

1.4.1 La odisea diagnóstica

La aproximación tradicional de diagnóstico genético no siempre es eficiente, especialmente en el caso de enfermedades raras en las que un diagnóstico molecular tarda en llegar; durante el proceso de identificar las variantes causantes de una enfermedad, se llega a someter al paciente a múltiples pruebas diagnósticas hasta identificar el gen o genes con la variante o variantes patogénicas relacionadas con su fenotipo, poniendo fin a una **odisea diagnóstica** que puede durar años (Lazaridis *et al.*, 2016; Sawyer *et al.*, 2016; Thevenon *et al.*, 2016; Lavelle *et al.*, 2022). Seguramente, uno de los precedentes más conocidos sea el de Nic Volker, un niño de 4 años con una enfermedad rara, un síndrome intestinal autoinmune; después de pasar por más de 100 operaciones quirúrgicas y mientras el personal médico consideraba un trasplante de médula, se le diagnosticó con éxito una mutación en el gen `XIAP` del cromosoma X mediante WES, que permitió modificar su tratamiento a un trasplante de sangre del cordón umbilical que curó definitivamente sus síntomas intestinales (Jacob *et al.*, 2013).

Muchos estudios también explican cómo la implementación temprana de la WGS se asocia con tiempos de respuesta diagnóstica más cortos y mayores rendimientos diagnósticos en comparación con las pruebas genéticas tradicionales, como los paneles de genes y los microarrays cromosómicos (Clark *et al.*, 2018; Farnaes *et al.*, 2018; Gubbels *et al.*, 2020; Wang *et al.*, 2020; Freed *et al.*, 2020; Lunke *et al.*, 2020; The NICUSeq Study Group, 2021). La implementación de la **medicina de precisión rápida (RPM) basada en la secuenciación rápida del genoma completo (rWGS)** ha demostrado mejores resultados clínicos y una reducción en los costes de atención médica de los pacientes en situaciones reales, por ejemplo con el proyecto *Baby Bear*, que mejoró los resultados clínicos de los recién nacidos al **incrementar la tasa de diagnóstico en un tiempo medio de respuesta de 3 días**, además de ser coste-eficiente al reducir

considerablemente los gastos hospitalarios, en comparación con las pruebas genéticas tradicionales (Dimmock *et al.*, 2021).

1.4.2 Análisis dirigidos en todas las etapas vitales

En el ámbito médico, realizar diagnósticos usando los resultados de una WES o WGS resulta problemático debido a la gran cantidad de variantes identificadas con un significado clínico incierto, que deben analizarse en un tiempo razonable para el tratamiento del paciente (sec. 1.3.5); tampoco resulta ideal la aproximación tradicional de realizar decenas o centenares de pruebas diagnósticas más dirigidas, como en el caso de Nic Volker. En una combinación de ambas aproximaciones, el proyecto *Baby Bear* ha demostrado la utilidad de la rWGS para el diagnóstico pediátrico, resultando más coste-eficiente y reduciendo significativamente la odisea diagnóstica de los pacientes, al utilizar la NGS para ir realizando análisis digitales de los datos generados hasta dar con un diagnóstico concluyente (sec. 1.4.1). Esta aproximación virtual se puede aplicar más allá del diagnóstico pediátrico de enfermedades raras, por ejemplo, a lo largo de toda la vida de una persona a través de consultas clínicas de su genoma (fig. 1.8).

En este sentido, un **panel virtual de genes** es un filtro fenotípico que define la consulta digital dirigida de un análisis NGS, en la que el uso de WES y WGS hace innecesaria la ejecución de un panel diagnóstico tradicional (sec. 1.2.3.6); estos paneles priorizan genes con mayor probabilidad de estar asociados con los síntomas de una persona (Wert *et al.*, 2021). Existen ya estudios que demuestran su eficacia diagnóstica como herramienta clínica, usándolos como metodología para acceder a ventanas de información genómica de manera controlada y limitada, como si cada una fuera un estudio, pero sin el coste económico y humano de los análisis tradicionales recurrentes; además, esta aproximación dispone también de ventajas a la hora de reanalizar los datos y reclasificar resultados de una manera automatizada (ver sec. 1.4.4 más adelante), y favorece la implementación de la medicina personalizada (sec. 1.2.4).

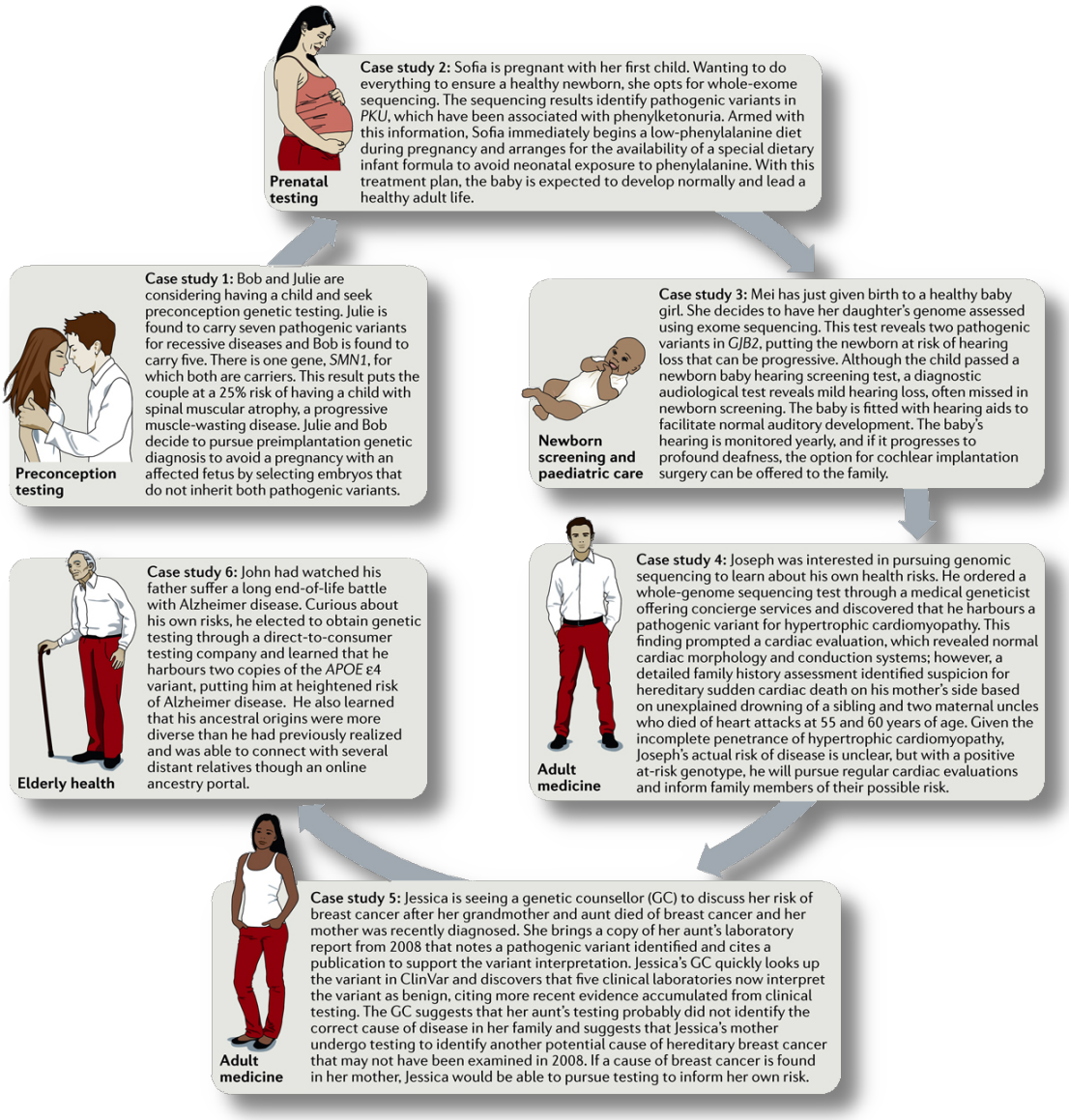


FIGURA 1.8: *El uso de la genómica a lo largo de la vida de una persona. Casos de uso de la genómica para la atención del paciente en diferentes etapas de su vida. Fuente: Rehm (2017).*

La implementación de los paneles virtuales de genes permite además evitar la detección de UFs y SFs que puedan aumentar la carga interpretativa del personal médico, así como la ansiedad para los pacientes (sec. 1.3.5). El diseño de un panel virtual permite priorizar las variantes accionables relacionadas con la indicación primaria para el test diagnóstico, a muy bajo coste humano y temporal, dando opciones para reajustar y reanalizar los datos de manera digital en caso de obtener un diagnóstico poco concluyente, aumentando potencialmente la tasa diagnóstica de una sola WGS o WES (Wang *et al.*, 2019).

1.4.3 Reanálisis

Los datos generados a partir de WES o WGS se almacenan digitalmente en varios formatos (secs. 1.3.1.1, 1.3.1.2, 1.3.1.3), lo que permite a los laboratorios, en caso de no poder establecer un diagnóstico con los datos actuales, volver a analizar esos mismos datos cuando se publique nueva evidencia en estudios futuros. Reanalizar la información genómica de un paciente del que se sospecha que tiene una afección genética subyacente puede mejorar el rendimiento diagnóstico de la prueba de secuenciación, lo que podría proporcionar beneficios significativos para el paciente y sus familiares, así como para el sistema de salud, por lo que se recomienda que los laboratorios proporcionen opciones de reanálisis para casos WES o WGS finalizados (Costain *et al.*, 2018; Machini *et al.*, 2019). Es importante también el desarrollo de herramientas que permitan capturar y recuperar anotaciones de variantes introducidas por un usuario anteriormente, y que permitan curar resultados locales y recuperarlos posteriormente para una posible toma de decisiones futura, aumentando así la eficiencia del reanálisis de datos (Oliver, Hart y Klee, 2015). Se recomienda que los laboratorios consideren un mínimo de 2 años para almacenar un tipo de fichero primario (por ejemplo, archivos BAM o FASTQ con todas las lecturas incluidas); esto permite el reanálisis de datos con *pipelines* analíticos mejorados (Rehder *et al.*, 2021). Además, los laboratorios

deben considerar el almacenamiento del VCF y el informe final del análisis durante 5 años, dada la probabilidad de una futura solicitud de reanálisis, y el establecimiento de una **base de datos interna de variantes** analizadas y clasificadas utilizando los estándares del ACMG y la AMP (sec. 1.2.3.4).

La producción de información genética adicional también proporciona mecanismos para que los clínicos caractericen mejor el genoma de un paciente; por ejemplo, pasar de paneles de genes a WGS proporciona una visión más completa del genoma del paciente, y ejerce su mayor influencia en la reidentificación de variantes. Las solicitudes para pruebas diagnósticas pueden abordarse de manera escalonada, con el personal médico inicialmente solicitando que se analice un panel restringido de genes en caso de que el fenotipo del paciente pueda indicar una posible condición genética (Robertson *et al.*, 2022).

Uno de los primeros estudios en examinar la coste-eficiencia del reanálisis del exoma consideró su coste-eficiencia dentro de una cohorte de bebés con sospecha de trastornos monogénicos, demostrando que el reanálisis realizado **18 meses después del informe original** fue más coste-eficiente que el realizado cada 6 meses, con un ahorro de costes incremental de 1058.74 dólares australianos (aproximadamente 710€) por diagnóstico adicional en comparación con la vía de atención médica estándar (Stark, Schofield, *et al.*, 2019).

1.4.3.1 Tipos de reanálisis

La forma en que se realizan los reanálisis puede variar en función de dos categorías principales: los procesos de reanálisis pueden ser **reactivos** (el paciente o el proveedor solicitante lo encarga) o **proactivos** (el laboratorio reanaliza directamente los casos terminados según sus políticas internas). La mayoría de los laboratorios actualmente realizan reanálisis reactivos; sin embargo, implementar análisis proactivos es un paso importante para maximizar la utilidad clínica de las pruebas NGS (Austin-Tse *et al.*, 2022).

1.4.3.2 Reanálisis en diferentes etapas de la NGS

El proceso de reanálisis no es específico de la genómica clínica, sino que ocurre, por ejemplo, cada vez que un médico revisa el historial médico de un paciente; reanalizar la misma información clínica con el contexto proporcionado por nuevos descubrimientos científicos puede resolver casos previamente no concluyentes. En el caso de la genómica, usar paneles virtuales para reanalizar la misma información clínica con el contexto proporcionado por nuevos descubrimientos genéticos ayuda a resolver casos sin un diagnóstico concluyente (Wenger *et al.*, 2017; Tan *et al.*, 2020; Robertson *et al.*, 2022). Existen diversos mecanismos por los cuales el reanálisis puede incrementar el rendimiento diagnóstico (Robertson *et al.*, 2022):

- El más común de ellos es el **descubrimiento de nuevas asociaciones** gen-enfermedad y variante-enfermedad.
- El **acceso a la información** sobre nuevas asociaciones gen-enfermedad o variante-enfermedad y variantes reclasificadas en bases de datos públicas.
- La **mejora de los métodos de filtraje** de variantes.
- La **recopilación de información fenotípica adicional** del paciente.
- La **reclasificación de variantes** previamente caracterizadas.
- La colaboración internacional mediante el **intercambio de datos** (ver sec. 1.5.3 más adelante).
- La generación de **información genómica adicional** en comparación con el análisis original.
- El **refinamiento del proceso de curación** de variantes.
- El uso de **genomas de referencia y modelos genéticos actualizados** para realinear las lecturas secuenciadas (Wright *et al.*, 2018).

1.4.4 Estado del arte

Actualmente, existen varios estudios que han utilizado paneles virtuales de genes para aumentar la tasa diagnóstica de casos previamente no resueltos, por ejemplo para enfermedades mitocondriales (Wortmann *et al.*, 2015), neuropatías (Walsh *et al.*, 2017; Wang *et al.*, 2019), displasias esqueléticas (Chandler *et al.*, 2018), y muerte súbita (Schön *et al.*, 2021). La actualización de estos paneles virtuales a partir de términos fenotípicos estándar definidos por la *Human Phenotype Ontology* (HPO), que se encargue de añadir o eliminar genes candidatos de manera automatizada basándose en la evidencia científica de ese momento, y los clasifique por orden de prioridad a partir del número de fenotipos afectados, también resulta muy prometedora a la hora de aumentar la tasa diagnóstica y la rapidez del test, al reducir el número de genes a analizar y acortar así el tiempo de análisis y anotación, además de eliminar el proceso manual de seleccionar e ir actualizando los genes específicos para el panel virtual (Saklatvala, Dand y Simpson, 2018; Wang *et al.*, 2019). El hecho de que sea virtual facilita mucho su actualización periódica, y permite diseñar análisis individualizados de bajo coste para cada paciente y sus condiciones fenotípicas concretas.

Para que el proceso de curación y mantenimiento de los paneles virtuales de genes sea escalable a las necesidades actuales de la genómica clínica, se han desarrollado a nivel internacional esfuerzos comunitarios y colaborativos para definir y actualizar dichos paneles, como es el caso de ***PanelApp***, una herramienta web de colaboración abierta que permite que los paneles se compartan y evalúen conjuntamente por la comunidad científica (Martin *et al.*, 2019); la herramienta sirve como una base de datos de conocimiento público de paneles virtuales de genes que están curados y, además, en constante revisión, debido a la complejidad de interpretar correctamente las variantes identificadas con la literatura y la evidencia clínica actuales (Stark *et al.*, 2021).

1.5 Reporte de resultados

Todo este proceso de análisis y reanálisis de los datos generados por NGS puede producir grandes cantidades de variantes clínicamente relevantes según la evidencia científica del momento, con lo que resulta vital desarrollar procesos que prioricen toda esa información para poder asistir el diagnóstico de la pregunta clínica principal por la cual se encargó la prueba genética para un paciente. Este proceso implica disponer de herramientas que permitan la visualización de todos los datos generados, su incorporación a un informe de resultados, el intercambio de dichos resultados entre instituciones, y la limitación de descubrimientos adicionales que puedan entorpecer la labor del analista clínico que se encarga de su interpretación.

1.5.1 Visualización

La visualización de los datos originados mediante un análisis NGS se ha vuelto más compleja, ya que requiere de herramientas capaces de integrar datos de diferentes tipos (clínicos, moleculares, fenotípicos, ambientales, etc.) y diferentes tejidos a lo largo del tiempo de vida de una persona, para proporcionar información de una manera flexible, rápida y eficiente, cuando anteriormente se limitaba a representar el genoma en una sola dimensión (Rehm, 2017; Stark, Dolman, *et al.*, 2019). Estas herramientas necesitan de una infraestructura computacional adecuada, incluyendo almacenamiento y capacidad de procesamiento suficientes para realizar análisis de conjuntos de datos grandes y complejos depositados en repositorios estables y accesibles, a la vez que mantiene los estándares de privacidad de los países en los que se alojan, para poder visualizar distintos tipos de datos de una manera consolidada; todos estos procesos requieren de profesionales formados en biología, informática, ciencias de la computación, matemáticas, estadística y/o ingeniería (Green y Guyer, 2011).

1.5.2 Informe de resultados

Una vez se han validado las variantes identificadas, se procede a informar de los resultados al paciente, generalmente mediante un informe estático. El contenido del informe debe cumplir con los criterios de las guías redactadas por instituciones como el ACMG, la ESHG y la *Medical Genome Initiative* o iniciativa del genoma médico (MGI) (Matthijs *et al.*, 2016; Rehder *et al.*, 2021; Austin-Tse *et al.*, 2022), aunque entre ellas adoptan diferentes enfoques con respecto a las VUS, los SFs y las variantes con una relevancia ambigua para el paciente (sec. 1.3.5). Al comunicar los resultados finales al paciente, las políticas de reporte del laboratorio deben maximizar el potencial diagnóstico del análisis al tiempo que reducen la comunicación de variantes que puedan causar una carga innecesaria para el médico o ansiedad adicional para el paciente. Las recomendaciones estándar más recientes son (Rehder *et al.*, 2021):

- Las variantes deben priorizarse según su relevancia para el fenotipo.
- El informe de resultados debe indicar claramente las variantes relevantes para la indicación principal de la prueba y distinguirlas de los SFs u otros tipos de variantes. En este aspecto, el ACMG recomienda que los descubrimientos primarios aparezcan en el informe escrito como un resultado interpretativo breve y conciso al inicio del informe, indicando la presencia o ausencia de variantes consistentes con el fenotipo reportado.
- Se debe considerar la redacción de informes de resultados para audiencias de diferentes orígenes, ya que un informe para un profesional de la salud puede ser diferente del informe para un público no especializado (Recchia *et al.*, 2020).
- Los laboratorios pueden optar por utilizar declaraciones como *Positivo*, *Anormal* o *Descubrimiento clínicamente relevante* para describir la detección de una variante que explique los descubrimientos clínicos principales o una variante clínicamente accionable; *Negativo* indicaría que no se identificaron variantes relevantes para el fenotipo; *Incierto* o *Ver informe* indicaría que existe incertidumbre con respecto

a la asociación entre el fenotipo y las variantes reportadas.

- Al reportar sobre un gen asociado a un trastorno genético tratable, el laboratorio debe considerar agregar una referencia al tratamiento en el informe.
- Todos los informes deben incluir una lista de variantes identificadas clínicamente relevantes, anotadas según la nomenclatura de la *Human Genome Variation Society* (Den Dunnen *et al.*, 2016) y clasificadas según las directrices del ACMG y la AMP (sec. 1.2.3.4).
- Los nombres de los genes deben ajustarse a la nomenclatura aprobada por el *HUGO Gene Nomenclature Committee* (https://www.genenames.org/).
- Los siguientes elementos deben incluirse para cada variante dentro de un gen:
  - Coordenada genómica con la versión del genoma.
  - Nombre del gen.
  - Transcrito de referencia.
  - Cigosidad.
  - Nomenclatura del ADNc.
  - Cambio de nucleótidos.
  - Nomenclatura para el impacto proteico previsto o conocido.
  - Clasificación de la variante.
- Los siguientes elementos deben incluirse para cada variante fuera de las regiones codificantes:
  - Coordenada genómica con la versión del genoma.
  - Cambio de nucleótidos.
  - Cigosidad.
  - Clasificación de la variante.
- El informe debe incluir al final un resumen de la metodología validada y todas las limitaciones de la prueba, incluida la versión concreta de las bases de datos y los *pipelines* bioinformáticos usados.

1.5.3 Intercambio de datos

La labor interpretativa para clasificar correctamente variantes genéticas es ardua, de alta complejidad y requiere de revisiones periódicas para evitar decisiones médicas erróneas. Poder compartir datos estandarizados y estructurados de variantes entre instituciones facilita esta tarea enormemente, al poder centralizar el conocimiento global para una variante concreta en una base de datos común, en vez de disponer de diferentes evidencias clínicas en bases de datos aisladas; a su vez, este intercambio de datos puede facilitar el desarrollo de protocolos automatizados de procesamiento de esos datos para su reanálisis periódico conforme se vaya introduciendo y curando información nueva. Las implementaciones actuales incluyen el desarrollo de sistemas federados que reduzcan los silos de datos y permitan la interoperabilidad de éstos, con el objetivo último de mejorar directamente la salud de los pacientes, al mismo tiempo que se sigue permitiendo a las bases de datos locales mantener su autonomía y legislación local (Global Alliance for Genomics and Health, 2016). Esta aproximación federada ha sido propuesta por la *Global Alliance for Genomics and Health* o Alianza Global para Genómica y Salud (GA4GH) y desarrollada en el proyecto ***Beacon*** (Global Alliance for Genomics and Health, 2018); otras iniciativas internacionales incluyen las desarrolladas por ClinGen (Rehm *et al.*, 2015) y *Matchmaker Exchange* (MME) (Philippakis *et al.*, 2015), que han facilitado un incremento muy significativo de nuevos descubrimientos, como se puede observar en la cantidad de nuevas entradas agregadas a las bases de datos clínicamente relevantes como OMIM (Amberger *et al.*, 2015) y ClinVar (Landrum *et al.*, 2014).

Los EHRs juegan un papel importante en el intercambio y la estandarización de datos, como demuestra un estudio que encontró muchas asociaciones clínicamente relevantes que no se habían reportado anteriormente (Vujkovic *et al.*, 2020). La integración de los EHRs en la rutina clínica todavía está muy lejos de ser algo común (Rasmussen *et al.*, 2016), como también pasa con el intercambio de implementaciones

y experiencias de evaluación (Wolf *et al.*, 2018). Actualmente, el desarrollo del formato ***Health Level Seven*** **(HL7)** se está estandarizando para la transferencia segura de datos clínicos entre distintas instituciones y proveedores de servicios de salud, para mejorar la interoperabilidad e integración de los distintos sistemas EHR existentes (Saripalle, Runyan y Russell, 2019).

1.5.4 Reporte de descubrimientos adicionales

La inclusión de descubrimientos adicionales (sec. 1.3.5) en el informe de resultados conlleva actualmente cierta polémica entre diferentes instituciones sobre si se deberían reportar al paciente aunque no estén relacionados con la indicación clínica principal, ya que podrían resultar relevantes para las perspectivas de salud o las opciones reproductivas de los pacientes o sus familias. Declaraciones de la ESHG (El *et al.*, 2013; Wert *et al.*, 2021) con respecto al OGS recomiendan a día de hoy que:

- El análisis del genoma se limite a la indicación principal para la prueba, destinada a la identificación de la etiología genética subyacente de una enfermedad.
- Se aplique un enfoque cauteloso para el OGS, es decir, no buscar activamente SFs.
- El análisis genómico sea lo más específico y dirigido posible.

Estas recomendaciones de no realizar OGS se sostienen con varios argumentos a favor de una aproximación dirigida: los programas de cribado tienden a tener una financiación limitada de las instituciones públicas, y las decisiones políticas para iniciar estas nuevas actividades generalmente exigen de la compensación del presupuesto extra en atención médica con otros presupuestos estatales. El ESHG establece que, al menos por el momento, hay demasiadas incógnitas y preocupaciones para afirmar con seguridad que las propuestas actuales de OGS cumplen claramente con estos criterios, y mucho menos que definan el estándar de atención médica; además, un

escenario en el que el OGS desplace los recursos destinados a pruebas genómicas basadas en indicaciones clínicas sigue siendo motivo de preocupación en los sistemas de atención médica financiados públicamente (Wert *et al.*, 2021). La necesidad de una base de conocimientos de interpretación de variantes mejor curada para clasificar correctamente esas variantes adicionales también es importante (Amendola *et al.*, 2015). Otras instituciones han publicado recomendaciones sobre SFs en la misma línea: el Colegio Canadiense de Genetistas Médicos, la Sociedad Alemana de Genética Humana, el Consejo de Salud de los Países Bajos, la Agencia Francesa de Biomedicina, y otros programas clínicos y de investigación, ninguno de ellos recomendando el análisis intencional de los SFs, sugiriendo reportar solo aquellos genes con una asociación conocida entre el genotipo aberrante y la patología (Boycott *et al.*, 2015; Matthijs *et al.*, 2016; Wert *et al.*, 2021).

Por otro lado, organizaciones como el ACMG, la *Société Française de Médecine Prédictive et Personnalisée* o Sociedad Francesa de Medicina Predictiva y Personalizada (SFMPP) y Genomics England con el proyecto 100,000 Genomas, adoptan un enfoque totalmente opuesto y recomiendan el OGS de un conjunto predefinido de variantes genómicas accionables y altamente penetrantes en genes candidatos, independientemente de la indicación principal para la prueba y de la edad del paciente (Green *et al.*, 2013; Pujol *et al.*, 2018; Miller *et al.*, 2021; Genomics England, 2022). Argumentan que hay que aprovechar la oportunidad para buscar SFs de manera rutinaria y sistemática, es decir, realizar OGS en pacientes que se someten a pruebas genómicas; en el caso de Genomics England, de hecho, ya se están reportando (Genomics England, 2022). Una propuesta de lista de SFs ha sido diseñada por el ACMG (https://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/); se han publicado varias revisiones, y actualmente está compuesta por 78 genes (Green *et al.*, 2013; Richards *et al.*, 2015; Kalia *et al.*, 2017; Miller *et al.*, 2021, 2022). La SFMPP emitió una declaración en 2018 recomendando informar sobre los SFs en 36 genes accionables relacionados con formas específicas de cáncer en adultos (Pujol *et al.*, 2018). Sus recomendaciones

se limitan al OGS en adultos, a la espera de un futuro debate sobre la aceptabilidad del OGS para genes relacionados con cáncer en menores.

Aunque los beneficios más evidentes del OGS son médicos, un estudio de 2019 publicó evidencia de que **solo el 2.7%** de las personas sanas evaluadas con la lista de 59 genes del ACMG (Kalia *et al.*, 2017) demostró tener una variante dominante y clínicamente accionable (Haer-Wigman *et al.*, 2019).

1.6 Genomcore S.L.

Como parte de un doctorado industrial, el presente proyecto se ha desarrollado en Genomcore S.L., una empresa tecnológica fundada en 2015 con el objetivo de permitir la implementación efectiva de la medicina personalizada en la sociedad, ubicada en Esplugues de Llobregat (Barcelona). Trabaja bajo dos marcas diferentes en función de si se dirige al segmento profesional o al del usuario final:

- **Genomcore** ofrece un *Biomedical Information Management System* o sistema de gestión de información bioinformática (BIMS), una plataforma bioinformática para la gestión de datos personales biosanitarios en entornos profesionales, además de servicios de consultoría asociados, que permite gestionar grandes volúmenes de datos personales de salud de diferente naturaleza (genética, bioquímica, historiales clínicos, etc.) de forma integrada (https://genomcore.com). Su principal diferenciación es que, aparte de trabajar con datos agregados y anonimizados, posibilita la gestión de éstos de forma asociada al paciente (es decir, datos personales), ofreciendo el marco legal y tecnológico para que los clientes puedan implementar diferentes casos de uso sobre una misma plataforma. Esta plataforma, a su vez, se divide en tres unidades dependiendo de su ámbito de aplicación:
  - BIOMED (https://genomcore.com/biomed/), una infraestructura para

satisfacer las necesidades bioinformáticas de los sectores de la salud, la investigación biomédica y la farmacéutica, en un entorno centralizado que facilita el análisis colaborativo.

  - FRONTDESK (https://genomcore.com/frontdesk/), una aplicación de autoservicio única y personalizable para usuarios finales con la finalidad de conectar a los pacientes con los proveedores de atención médica, que ofrece una experiencia centrada en el paciente.
  - IVF (https://genomcore.com/ivf/), una plataforma para el cribado genético y el estudio de la compatibilidad entre donantes y receptores de biobancos de gametos y clínicas de reproducción asistida.

- ***Made of Genes*** **(MoG)** es un servicio de evaluación y gestión de la salud de forma personalizada a través de estudios moleculares integrados de ADN, analíticas de sangre y *Real-World Data* o datos del mundo real (RWD), para usuarios finales y grandes canales (https://madeofgenes.com/). Utiliza la tecnología desarrollada en Genomcore, además de técnicas de inteligencia artificial curadas por expertos biosanitarios, para dar un sentido nuevo a las analíticas tradicionales, proporcionando guías personalizadas de salud. Su objetivo no es el diagnóstico de enfermedades, sino la mejora y la preservación de la salud a través del autoconocimiento del usuario que permite promover un cambio de hábitos en el día a día.

CAPÍTULO 2

Objetivos

Los objetivos planteados inicialmente para este proyecto de tesis fueron los siguientes:

1. Implementar un *Virtual Panel Management System* o sistema de gestión de paneles virtuales (VPMS) que permita al usuario o profesional clínico realizar, de manera intuitiva y personalizada, análisis automatizados y estandarizados para las regiones genómicas que escoja, a partir de datos NGS del paciente.
2. Implementar una *Report Generation Tool* o herramienta de generación de informes (RGT) que permita utilizar los resultados del VPMS para su uso en medicina personalizada, que sean consistentes, rutinarios y estandarizados según las necesidades del mercado clínico actual.
3. Validar en entornos clínicos el VPMS y la RGT, mediante una prueba piloto que permita validar uno o varios paneles virtuales de genes con el servicio de medicina genética y genómica de un hospital o una entidad del sector de la salud.

CAPÍTULO 3

Metodología

Este capítulo detalla la metodología utilizada para el desarrollo de esta tesis, explicando cómo se realiza la gestión de los datos que se obtienen de diferentes laboratorios y colaboradores externos, su tratamiento y organización dentro de la plataforma genómica de Genomcore, su traducción a resultados accionables, y la revisión y validación de todos los procesos involucrados.

3.1 Sistema de gestión de paneles virtuales de genes

3.1.1 Genomcore BIOMED

Genomcore BIOMED (fig. 3.1) permite la implementación de la medicina personalizada con la visión de ofrecer el marco tecnológico y legal necesario para aplicar su uso en la atención primaria (sec. 1.6). Esta plataforma bioinformática constituye una solución flexible para almacenar todo tipo de datos biomédicos, desde datos NGS y otras tecnologías ómicas a cualquier tipo de información estructurada de salud, y permite procesarlos mediante flujos de trabajo analíticos automatizados para la generación de

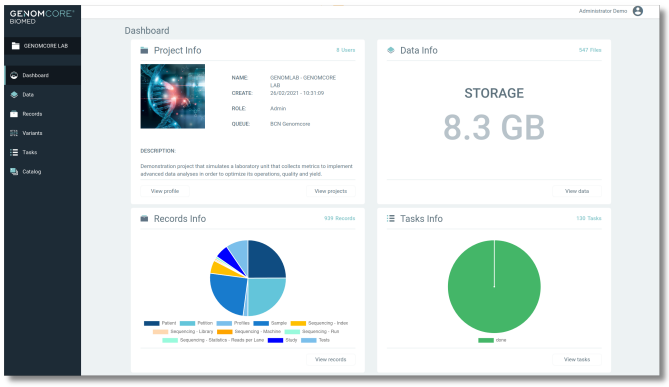
resultados accionables; dichos análisis se realizan en diversos nodos de computación, disponibles bajo demanda en un centro de datos local, como también en la nube mediante *Amazon Web Services* (AWS), estableciendo así una plataforma analítica capaz de realizar computaciones masivas y en paralelo ajustadas a las necesidades de cada usuario.

Los resultados de los análisis se almacenan en la plataforma de forma segura, además de disponer de funcionalidades para su intercambio dentro y fuera de ésta. Principalmente, existen dos maneras de almacenar datos:

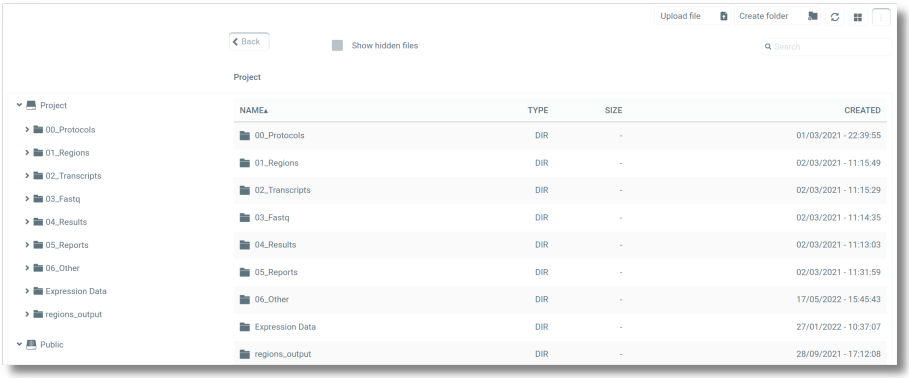
- Mediante ficheros convencionales que se almacenan en el módulo de ficheros de la plataforma (sec. 3.1.6).
- Mediante **Records**, entidades tipadas y estructuradas en formato JSON para almacenar información en bases de datos de manera flexible, diseñadas exclusivamente para su uso en la plataforma (sec. 3.1.7).

3.1.2 Infraestructura

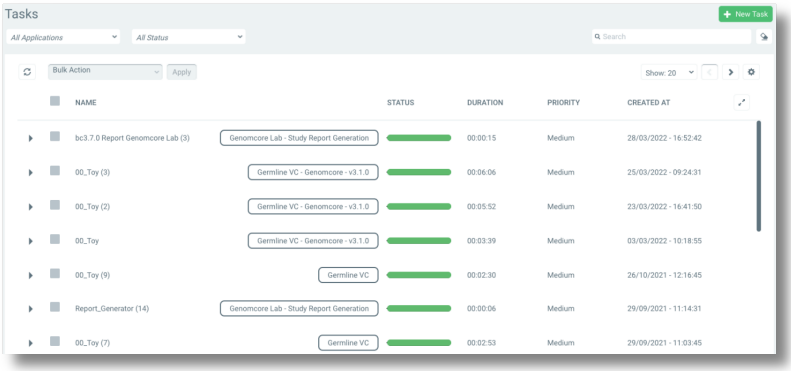
La unidad básica de trabajo en BIOMED son los **proyectos**, y cada proyecto forma parte de una **organización**; el administrador de cada organización gestiona los permisos de los usuarios a sus proyectos, estableciendo así diferentes niveles de acceso por usuario (administrador, analista, editor, o de solo lectura) para definir las acciones que puede realizar. El control de acceso a la plataforma se realiza mediante un sistema de credenciales basado en un correo electrónico, una contraseña, y el identificador de la organización a la que pertenece ese usuario. Esa misma organización es la que permite también definir los permisos de cada proyecto a la hora de utilizar las diferentes aplicaciones y flujos de trabajo o ***workflows*** de los que dispone la plataforma (secs. 3.1.3.1, 3.1.3.2). Además, BIOMED tiene diversas ***application programming interfaces* o interfaces de programación de aplicaciones (APIs)** que proporcionan una infraestructura pública de gestión de los distintos servicios ofrecidos:



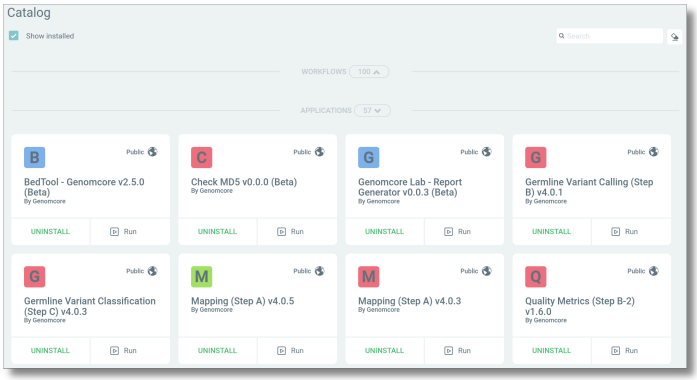
(A) *Tablero general de un proyecto.*



(B) *Módulo de ficheros.*



(C) *Módulo de tareas.*



(D) *Catálogo de aplicaciones y workflows.*

FIGURA 3.1: *Genomcore BIOMED y los diferentes módulos de trabajo.*

- Autenticación de usuarios (*Auth API*).
- Subida y descarga de ficheros (*File Service API*).
- Ejecución de tareas (*Tasks API*).
- Consulta y almacenamiento de Records (*Records API*).
- Envío de notificaciones automatizadas en correos electrónicos (*Email API*).
- Servicios generales, por ejemplo, creación de aplicaciones (*BIOMED API*)

3.1.3 Gestión de aplicaciones y workflows

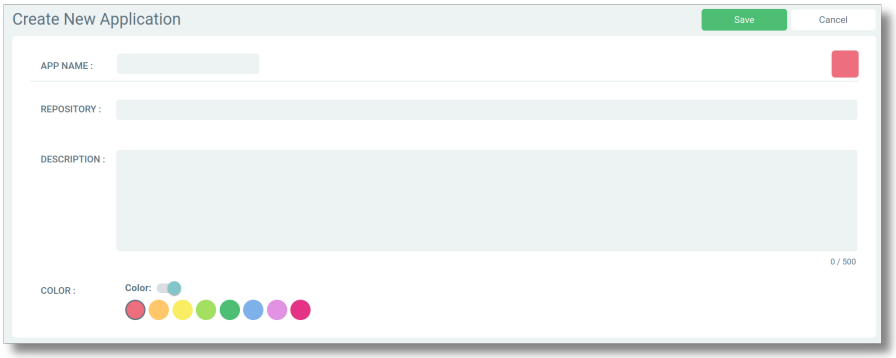
3.1.3.1 Aplicaciones y versiones

Los análisis ejecutados en la plataforma se gestionan mediante la creación de aplicaciones y versiones, y su concatenación en forma de *workflows* (sec. 3.1.3.2). Cada **aplicación** (fig. 3.2a) es un programa que se usa para realizar una tarea concreta, por ejemplo: transformar los datos en bruto en formato BCL de la secuenciación a ficheros FASTQ, o realizar la detección de variantes. Los cambios aplicados a la funcionalidad de una aplicación durante su desarrollo y vida útil se organizan en **versiones** (fig. 3.2b). Existen 3 categorías para cada versión que se genera (fig. 3.3):

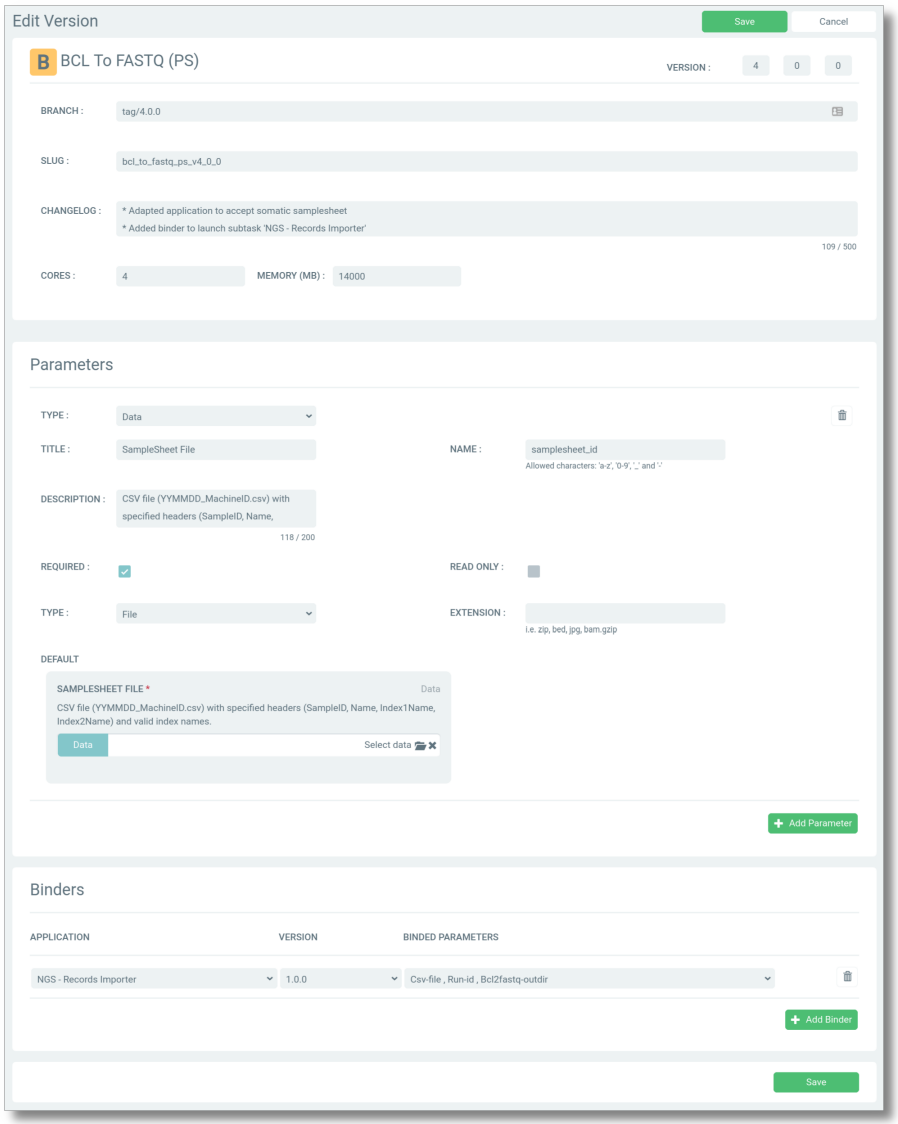
- *Beta*: la versión que está en desarrollo; solo puede existir una versión *Beta*.
- *Latest*: la versión estable y funcionando en el entorno de producción. Solo puede existir una versión con esta categoría.
- *Deprecated*: versiones anteriores que se encuentran desactualizadas. Pueden existir múltiples versiones con esta categoría.

Estas categorías disponen además de las siguientes características y restricciones:

- Cada nueva versión en desarrollo (fig. 3.2b) se categoriza por defecto como *Beta*, y solo puede ser modificada a *Latest* una vez ha completado las pruebas correspondientes (sec. 3.1.3.4).



(A) *Creación de una aplicación.*



(B) *Creación de una versión, con un apartado de parámetros de entrada y otro de 'binders', que permiten configurar conexiones a distintas versiones de manera flexible (sec. 3.1.3.2).*

FIGURA 3.2: *Creación de aplicaciones y versiones en BIOMED.*

- Cuando una versión *Beta* se actualiza a *Latest*, la versión anterior deja ser *Latest* y se categoriza automáticamente a *Deprecated*.
- Solo se puede crear una versión *Beta* si la última versión existente es *Latest*.
- Cada versión tiene seis acciones disponibles desde el panel de control (fig. 3.3).

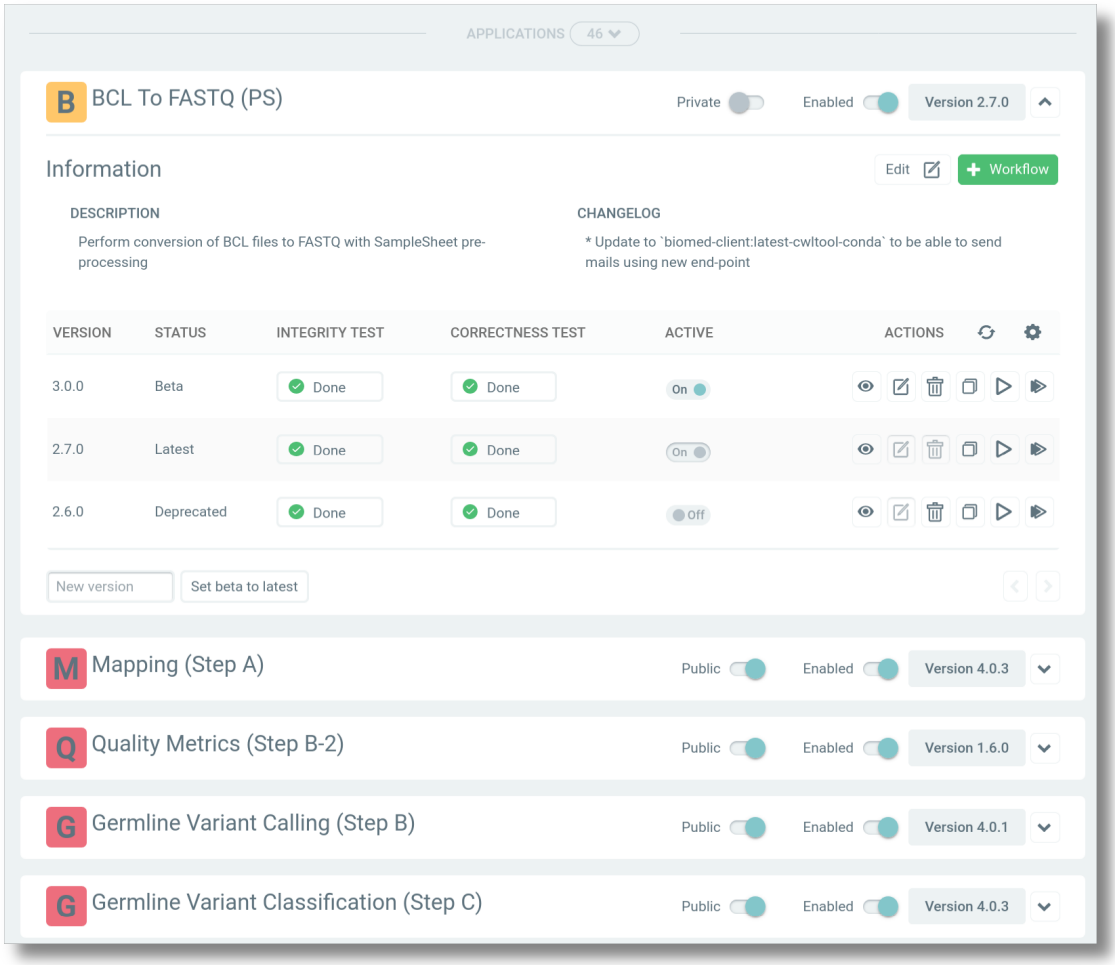


FIGURA 3.3: *Aplicaciones y versiones disponibles desde el Panel de Control. Los 6 botones de acciones corresponden, de izquierda a derecha: 1) visualizar los parámetros de la versión, 2) editarlos, 3) eliminar la versión, 4) crear una nueva aplicación a partir de la versión, 5) ejecutar la prueba de integridad, o 6) ejecutar la prueba de exactitud.*

Para los usuarios de la plataforma, cada versión no es más que una "caja negra" (figs. 3.2b, 3.3) a la que se le proporcionan unos datos de entrada y genera unos de salida que se guardarán, por ejemplo, en forma de ficheros en la plataforma; internamente, está enlazada en realidad a una imagen Docker (https://www.docker.com/)

con el código de las instrucciones de ejecución específicas para esa aplicación, generada mediante un fichero `Dockerfile` (fch. 3.1), que se almacena en un ***Elastic Container Registry*** **(ECR)** de AWS, un repositorio privado de imágenes Docker de la empresa, listas para ser descargadas y ejecutadas (sec. 3.1.5).

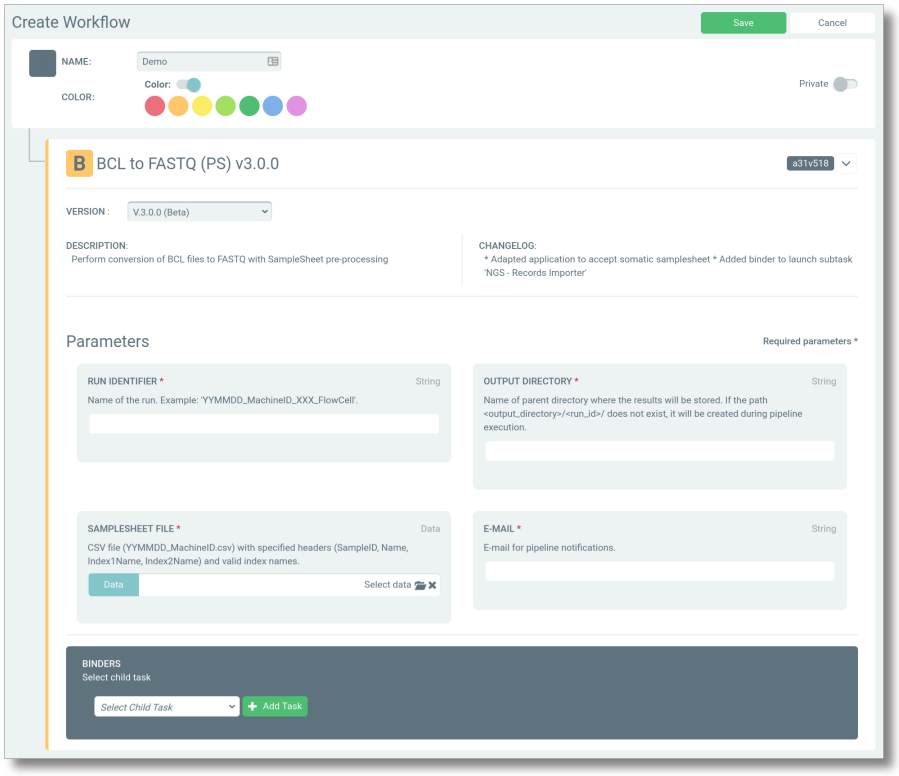
FICHERO 3.1: *Contenido de un Dockerfile de ejemplo.*

```
# Start from a public Python image and install wget
FROM python:3.8-slim-buster
RUN apt-get update && apt-get install -y --no-install-recommends wget
# Set clock and system language to spanish
RUN ln -sf /usr/share/zoneinfo/Europe/Madrid /etc/localtime
RUN echo "es_ES.UTF-8 UTF-8" >> /etc/locale.gen && locale-gen
# Create a user without admin permissions
RUN useradd --create-home pyuser
WORKDIR /home/pyuser
USER pyuser
# Setup a Python virtual environment
ENV VIRTUAL_ENV=/home/pyuser/env
RUN python -m venv $VIRTUAL_ENV
ENV PATH="$VIRTUAL_ENV/bin:$PATH"
RUN python -m pip install -U pip wheel
```

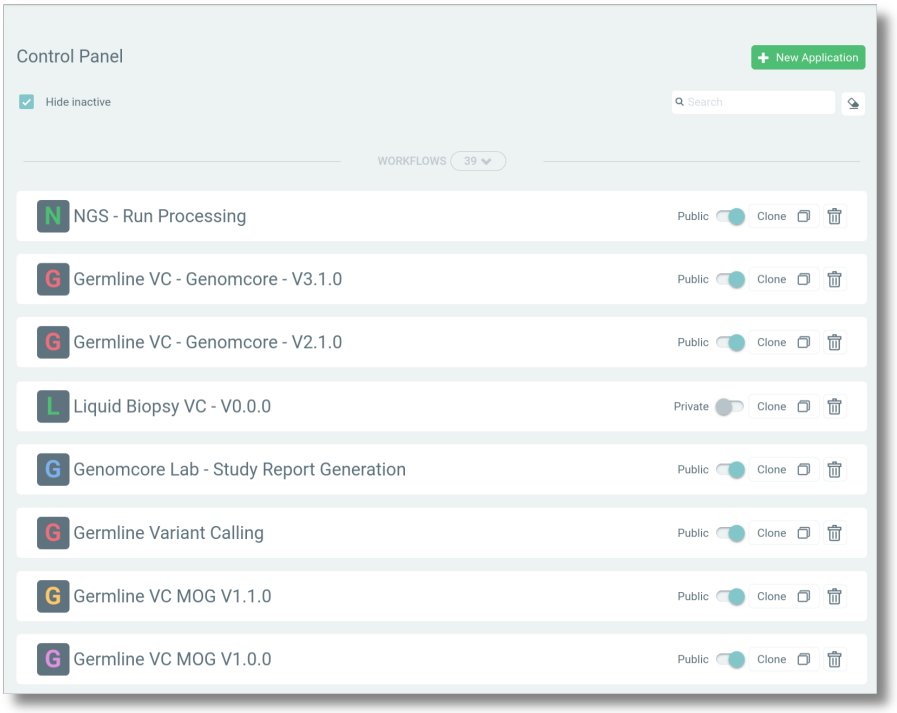
3.1.3.2 Workflows

Distintas aplicaciones se pueden encadenar entre sí formando un *workflow* (fig. 3.4), de manera que los datos de salida de la primera aplicación se usen como datos de entrada de la segunda, y así sucesivamente para cuantos procesos requiera el *workflow* entero. Esta concatenación se consigue mediante los ***binders***, parámetros que se configuran desde la versión (fig. 3.2b) y que se pueden añadir en el momento de creación de un *workflow* (fig. 3.4a) para conectar los datos de salida de una versión con los datos de entrada de otra.

Además, los *binders* permiten la **creación de *workflows* que no sean lineales**, es decir, con tantas ramificaciones como *binders* se configuren, debido a que se pueden configurar para apuntar a diferentes versiones desde una misma aplicación (fig. 3.5).



(A) *Creación de un workflow.*



(B) *Workflows disponibles desde el Panel de Control. Los tres botones a la derecha de cada fila permiten: seleccionar si el workflow es accesible a todos los usuarios de la plataforma; duplicar el workflow para crear uno nuevo; eliminar el workflow.*

FIGURA 3.4: *Creación y gestión de workflows en BIOMED.*

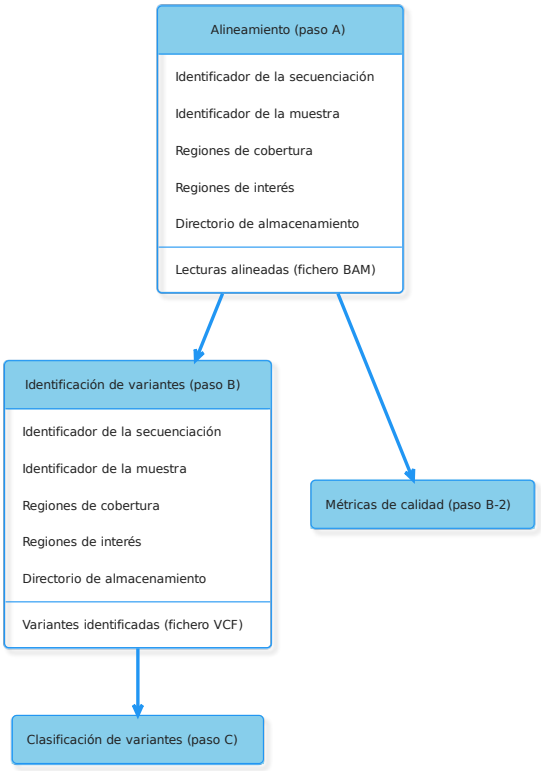


FIGURA 3.5: *Esquema del workflow de análisis NGS de identificación de variantes de Genomcore BIO-MED. Cada aplicación (en azul) le envía los parámetros de salida (en blanco) a la siguiente en forma de 'binders' conectando las aplicaciones entre sí, formando así un workflow. El parámetro de salida aislado en el paso A y el paso B corresponde, respectivamente, al fichero BAM y VCF generado por esa aplicación. El resto de parámetros no se generan en esa aplicación; se reciben como datos de entrada, se usan en la ejecución y, al finalizar, se envían a la siguiente aplicación. El workflow no es lineal, ya que la aplicación del paso A ejecuta dos aplicaciones cuando finaliza: el paso B y el B-2.*

3.1.3.3 Almacenamiento

La información sobre las aplicaciones, versiones y workflows que se crean en BIO-MED se almacenan mediante peticiones contra la *BIOMED API* (sec. 3.1.2) en una base de datos relacional de tipo PostgreSQL (https://www.postgresql.org/), que enlaza cada *workflow* con sus correspondientes aplicaciones y versiones. Además, también almacena la relación entre cada versión y su imagen Docker correspondiente (sec. 3.1.5).

3.1.3.4 Pruebas de integridad y exactitud

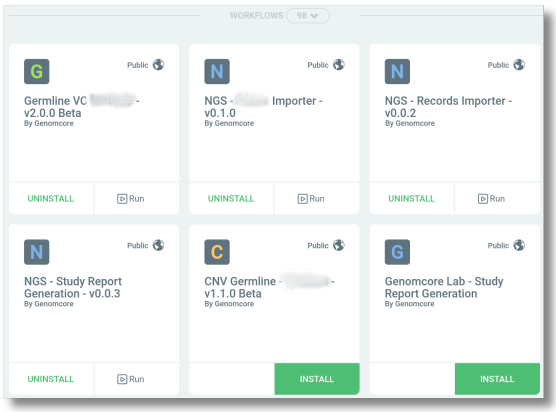
Las versiones de cada aplicación desarrollada disponen de dos modos de prueba en la plataforma, programadas por los mismos desarrolladores de éstas (fig. 3.3):

- La **prueba de integridad** comprueba que las interfaces de los diferentes componentes de la versión sean consistentes entre ellos y que el resultado de su integración permita realizar las funcionalidades esperadas; por ejemplo, que se puedan descargar y subir ficheros a BIOMED, que las llamadas a las distintas APIs se realicen sin errores, que la ejecución finalice correctamente, etc.
- La **prueba de exactitud** es la encargada de comprobar que la nueva versión proporciona los resultados esperados en un conjunto de datos de referencia; esto garantiza que los resultados sean reproducibles siempre que se utilice exactamente la misma versión de la aplicación.

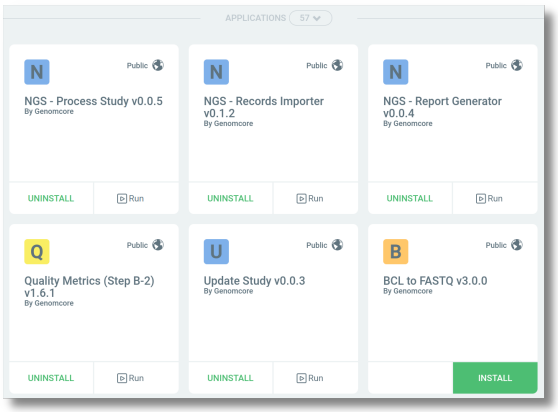
Una vez ambas pruebas han sido ejecutadas correctamente, la plataforma permite a un administrador modificar la categoría de la versión de *Beta* a *Latest*, de manera que los usuarios de la plataforma a partir de ese momento van a visualizar la nueva versión como la más estable y actualizada para realizar sus análisis.

3.1.3.5 Instalación

Una vez creados los *workflows*, aplicaciones y versiones correspondientes, la plataforma permite gestionar también su instalación en los proyectos especificados por el administrador de cada organización (fig. 3.6), de manera que los usuarios con acceso a ese proyecto y suficientes permisos pueden ver la disponibilidad de las aplicaciones que se les permite ejecutar.



(A) *Instalación de workflows. Los nombres de colaboradores se han tapado por motivos de confidencialidad.*



(B) *Instalación de aplicaciones.*

FIGURA 3.6: *Instalación de workflows y aplicaciones de BIOMED.*

3.1.4 FLUJO DE DESARROLLO

Todo el código desarrollado para el presente proyecto se ha almacenado en Bitbucket (https://bitbucket.org), un servicio de alojamiento de repositorios de código fuente, accesible de manera global pero protegido con credenciales privadas para cada desarrollador, basado en `git`, un sistema de control de versiones (Chacon y Straub, 2014). Nuestro flujo de trabajo como desarrolladores de código se basa en el modelo de ramas de `gitflow` (Driessen, 2010); la gestión de las diferentes versiones de cada aplicación en los diferentes entornos de implementación de BIOMED se basa en tres ramas diferentes:

- `production`: es la rama principal de cada repositorio, y contiene el código que está actualmente funcionando en el entorno de producción.
- `staging`: contiene el código de la versión desarrollada en un entorno de preproducción, y parte de la rama `production`. Esta rama se fusiona con la de `production` cuando la nueva versión desarrollada está validada y estable; en el momento de la fusión de ramas se despliega en el entorno de producción.
- `development`: contiene el código de la versión en desarrollo, y parte de la rama `staging`. Cuando esta rama se fusiona con la de `staging`, los diferentes

encargados del equipo de Producto de la empresa validan la funcionalidad de la nueva versión en el entorno de preproducción de BIOMED.

Las fusiones de cada rama despliegan automáticamente la imagen Docker correspondiente al entorno específico de BIOMED mediante el sistema automatizado de integración e implementación continua (*CI/CD* por sus siglas en inglés) de Bitbucket (https://bitbucket.org/product/features/pipelines), configurado para todos los repositorios de la empresa. Las tres ramas están protegidas para que ningún desarrollador pueda subir código directamente contra ninguna de ellas, de manera que todo el código nuevo se sube y organiza usando las siguientes tipologías de ramas:

- `release`: contiene el código de las versiones candidatas y oficiales por publicar, y parte de la rama `development`. Esta rama se revisa por el equipo de desarrollo antes de fusionarse con la rama `development` para comprobar en el entorno de pruebas de BIOMED que funciona correctamente.
- `feature`: contiene el código para una característica concreta por implementar, y parte de la rama `release`. Una nueva versión o `release` puede contener múltiples ramas `feature` fusionadas; una para cada característica nueva.
- `hotfix`: es un espejo de la rama `release` y se usa para aplicar parches directamente contra la rama `production` que arreglen errores urgentes detectados en el entorno de producción.

3.1.5 Ejecución de tareas

La ejecución de tareas en BIOMED se gestiona mediante un gestor de colas, SLURM (https://slurm.schedmd.com/), que se encarga de distribuirlas entre los diferentes nodos de computación disponibles (fig. 3.7); estos nodos están configurados para acceder al AWS ECR privado de la empresa, de manera que, durante la asignación de la tarea a su cola, se encargan de descargarse la imagen Docker necesaria para su ejecución.

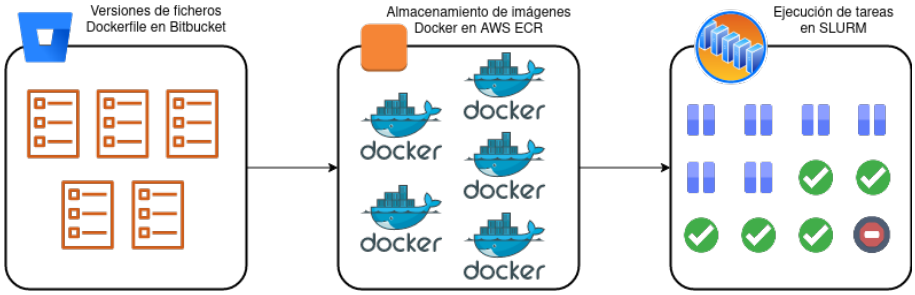


FIGURA 3.7: *Creación de imágenes Docker y ejecución en SLURM. El código de las aplicaciones se encapsula en imágenes Docker, creadas con instrucciones de un fichero Dockerfile y almacenadas en un ECR. De allí, se descargan bajo demanda en los nodos de computación gestionados por SLURM.*

Las aplicaciones y *workflows* instaladas en un proyecto determinado ya pueden ser ejecutadas finalmente por los usuarios de la plataforma; para ello, disponen de un módulo que les permite seleccionar la tarea a ejecutar y rellenar sus parámetros de entrada con valores para el análisis específico a realizar (ver fig. 4.4 más adelante). Las tareas en ejecución y finalizadas se pueden visualizar en formato de lista en BIOMED, ordenadas de más recientes a más antiguas, y con un botón para acceder a un sub-menú de opciones que permiten visualizar datos adicionales para cada tarea ejecutada (fig. 3.8).

3.1.6 MÓDULO DE FICHEROS

Los ficheros necesarios para los análisis se pueden almacenar directamente en BIOMED en módulo de ficheros, que está integrado con el de ejecución de tareas y agiliza la comunicación entre las imágenes Docker ejecutándose en los nodos y los ficheros de entrada necesarios, así como la transferencia de los ficheros de salida también al mismo módulo; esa comunicación se realiza mediante una API diseñada con este fin, a la que llamamos `File Service API`. Internamente, el módulo almacena los ficheros en un CEPH (https://docs.ceph.com), un sistema de almacenamiento distribuido y escalable, capaz de guardar los ficheros de gran tamaño que se utilizan para un análisis NGS. Además, este módulo dispone de dos divisiones: una privada para el proyecto concreto de BIOMED, al que solo tienen acceso los usuarios con permisos para

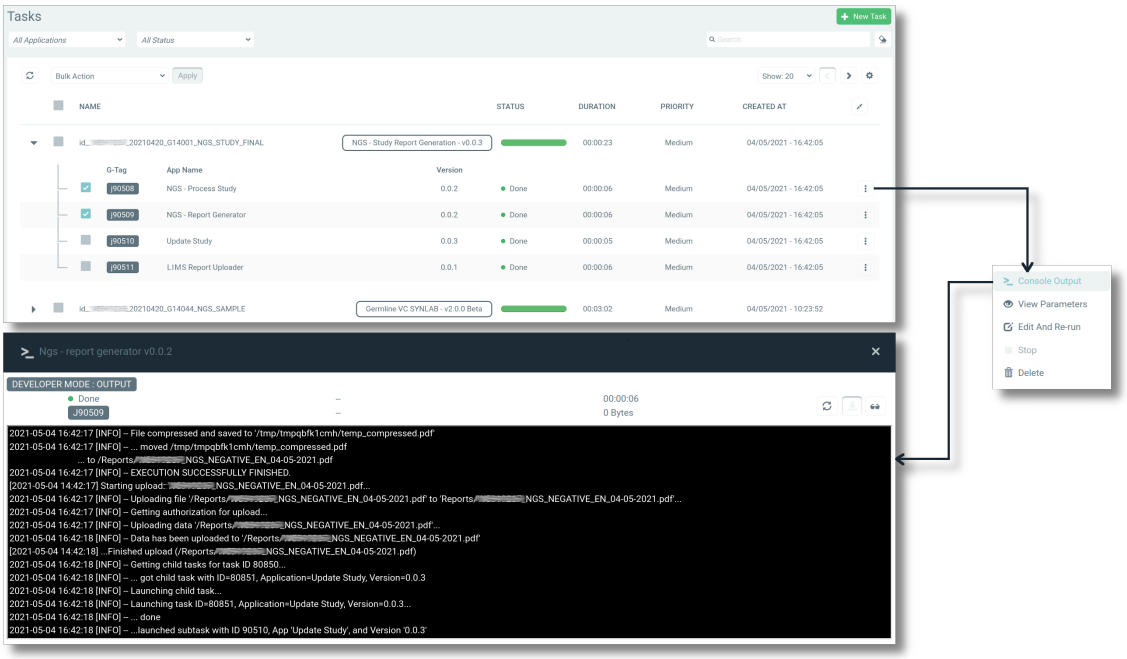


FIGURA 3.8: *Imagen superior: visualización de tareas ejecutadas. Imagen inferior: consulta del registro de ejecución de una aplicación. Derecha: submenú de opciones para una tarea para 1) ver el registro de la ejecución, 2) ver sus parámetros de entrada, 3) editar la tarea y volverla a ejecutar, 4) parar la ejecución de la tarea en caso de que siga en curso, o 5) eliminarla de la lista de tareas ejecutadas. El identificador de la muestra se ha eliminado por motivos de confidencialidad.*

ese proyecto, y otra pública, a la que tienen acceso los usuarios dentro de la misma organización de la plataforma (figs. 3.1b, 3.9).

3.1.7 MÓDULO DE RECORDS

Los Records constituyen el eje central de la organización y gestión de los datos en BIOMED; son **documentos tipados en formato JSON** (https://www.json.org/) que se almacenan en una base de datos no relacional de MongoDB (https://www.mongodb.com/) y que permiten almacenar información de manera estructurada y validada, ya sea por ejemplo para un análisis NGS o de otros tipos. Por ejemplo, se utilizan para guardar estadísticas de la ejecución de las máquinas de secuenciación, información de muestras utilizadas, librerías de secuenciación, o resultados de un análisis NGS. Son documentos tipados porque la creación de un Record está ligada a la validación de sus parámetros mediante la utilización de **plantillas o *templates***, otro tipo de objetos en

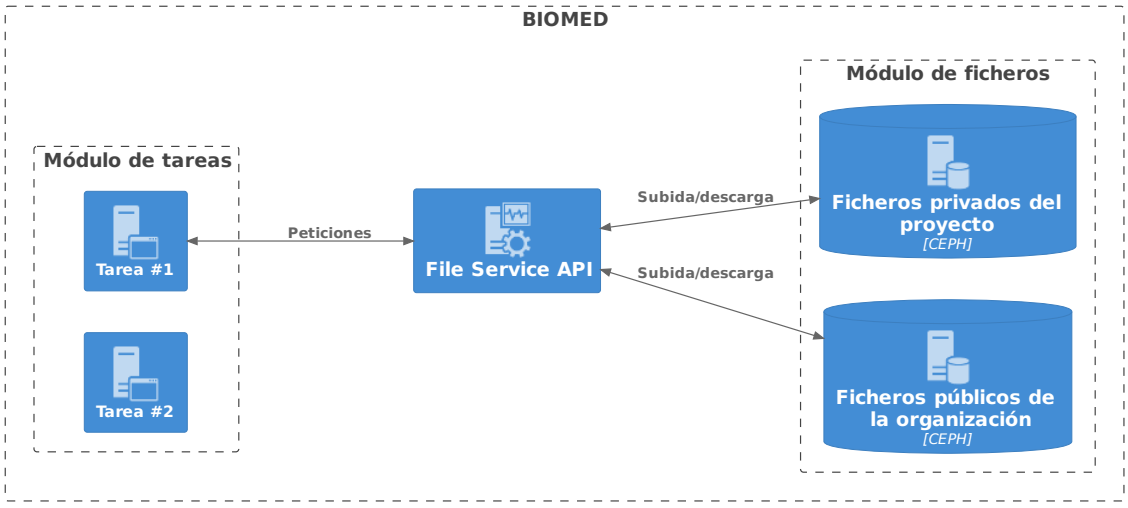


FIGURA 3.9: *Funcionamiento del módulo de ficheros y su integración con el módulo de tareas a través de la File Service API. Una tarea en ejecución realiza peticiones contra la API para poder subir o descargar ficheros en la división pública o privada del CEPH, el sistema de almacenamiento de BIOMED.*

formato JSON también almacenados en MongoDB. En cada plantilla se define un tipo de Record concreto (por ejemplo *Librería de secuenciación*) y se le configuran los campos obligatorios y el tipo de datos en cada campo (campo numérico, campo de texto libre o con opciones determinadas, etc.), para que al crear un Record de tipo *Librería*, se valide que se han rellenado los campos requeridos. Esta validación se realiza a través de la API de Records (sec. 3.1.2), que consulta la plantilla correspondiente y comprueba que la información del Record creado es correcta.

Con respecto a la experiencia de usuario en la plataforma, el módulo de Records se comunica mediante esta API a la base de datos, y permite visualizar los datos obtenidos mediante `ag-Grid` (https://www.ag-grid.com/), un *software* que implementa una cuadrícula personalizable; el usuario puede ver así todos los Records de una plantilla concreta en forma de listado, reordenar sus columnas y filtrar los campos con gran flexibilidad (fig. 3.10).

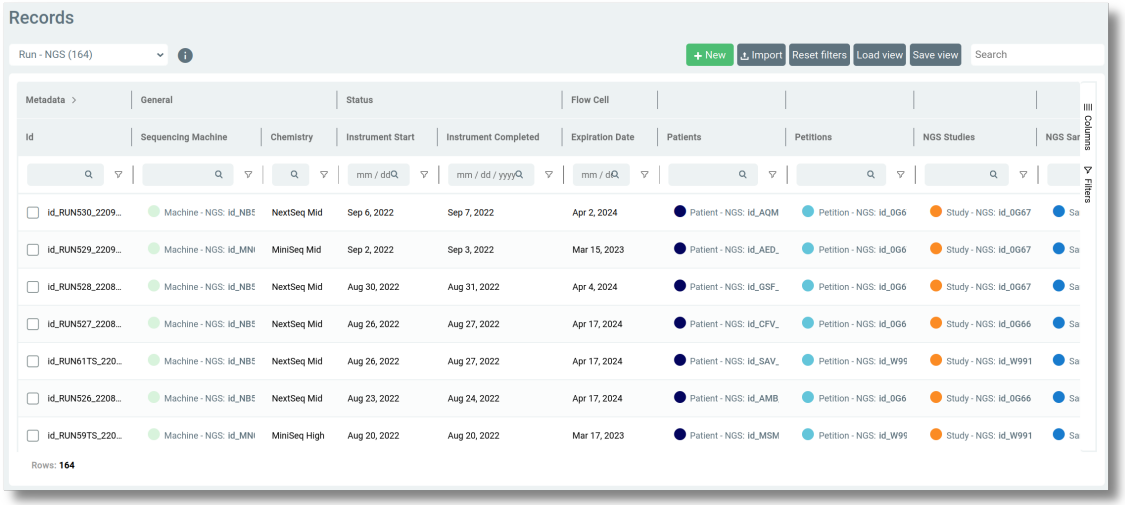


FIGURA 3.10: *Visualización de todos los Records para la plantilla 'Run - NGS' en BIOMED.*

3.1.8 MONITORIZACIÓN DE RECURSOS Y EVALUACIÓN

Todos los recursos utilizados por las diferentes aplicaciones que conforman el *pipeline* bioinformático de la empresa y el VPMS están monitorizados para analizar posibles errores que aparezcan en la ejecución de tareas, así como para gestionar todas las alertas necesarias y obtener información del estado de los diferentes sistemas y APIs de la plataforma. Debido a la magnitud de la cantidad de datos generados para este proceso de monitorización, el equipo de desarrollo de la empresa ha diseñado un sistema de indexación de datos que los almacena en `ElasticSearch` (https://www.elastic.co/elasticsearch/) para luego consultarlos y visualizarlos mediante `Kibana` (https://www.elastic.co/kibana/). Además, han diseñado un sistema de monitorización de servicios mediante *Application Performance Monitoring* (APM) en el cual se puede consultar el estado de cada API de la plataforma en tiempo real, así como el consumo de recursos de cada uno de sus componentes (https://www.elastic.co/observability/application-performance-monitoring).

Finalmente, existe también un sistema que centraliza la recepción de comentarios, sugerencias y errores por parte de los usuarios de la plataforma, basado en `Jira` (https://www.atlassian.com/software/jira) y metodología *Agile* (http://agilemanifesto.org/);

este sistema les proporciona a los usuarios herramientas web para reportar información al equipo de desarrollo. También disponen de atención por correo electrónico y teléfono en caso de que no les sea posible la primera opción.

3.2 HERRAMIENTA DE GENERACIÓN DE INFORMES

La RGT también utiliza la infraestructura de BIOMED descrita en la sección anterior, así como los diferentes módulos que la componen (secs. 3.1.2, 3.1.5, 3.1.6, 3.1.7, 3.1.8). Adicionalmente, se ha diseñado con los elementos descritos en la sección a continuación para el procesamiento de los datos de entrada, la firma digital de documentos, la generación de informes en formato PDF, y el reporte de resultados en forma de aplicación web.

3.2.1 DATOS BIOQUÍMICOS

Una de las fuentes utilizadas en la RGT como datos de entrada son los análisis bioquímicos realizados en laboratorios externos a la empresa mediante analíticas de sangre. Estos datos se almacenan en un fichero en el formato HL7 (https://www.hl7.org/fhir/), el formato estándar por excelencia para el intercambio de datos de salud, que usamos mayoritariamente en su versión 2.5 (fch. 3.2); para procesarlos, usamos la librería `HL7apy` (Meloni *et al.*, 2015).

FICHERO 3.2: *Contenido inicial de un HL7 versión 2.5, mostrando la analítica para la glucosa.*

```
1 MSH|^~\&|LIMS1|LAB1|BIOMED|MOG|20220319122327||ORU^R01|T111|P|2.5|||AL||||||SFTP
2 PID|1|peti123|46999999A|T111|SELLARÈS^RUBIO^ARNAU||19901212000000|M|||||^^^||8^SPANISH
3 PV1|1|O|||||||||||||||||2218|||||||||||||||||||||||||202202080000000||||||
4 ORC|RE|||T111|CM||^^^^^||20220216095300|^^||^^^^^|||202202170000000|||||||||
5 OBR|10000|T111||GLU^GLUCOSA||20220208074600|20220208175400|202202081754000||||^^|||||||||F
6 OBX|10000|NM|B0000^GLUCOSA||106,000|^mg/dL|74,000-106,000 INSIDE|||M|||202202081754000||^|
```

3.2.2 DATOS DE GENOTIPADO

La segunda fuente de datos utilizada en la RGT corresponde a datos genéticos obtenidos de *microarrays* de genotipado de SNPs, que se almacenan en ficheros en el formato *Comma-Separated Values* o valores separados por comas (CSV) (fch. 3.3); para procesarlos, usamos la librería `pandas` (https://pandas.pydata.org).

FICHERO 3.3: *Contenido inicial de un CSV de genotipado de SNPs.*

```
# Export Date : 03/22/2022 13:04:13 CET,,,,
# Study Name : MOGv1_p20p21_MOG2_p6,,,,
# Experiment Type : Real-time,,,,
# Instrument Type : QuantStudio 12K Flex Real-Time PCR System,,,,
# Software Version Number : 1.3,,,,
# Creation Date : 11/04/2021 11:10:23 CET,,,,
# Created By : GUEST,,,,
# Last Modified Date : 03/22/2022 11:50:35 CET,,,,
# Last Modified By : GUEST,,,,
Sample ID,Plate Barcode,Assay Name or ID,Allele 1 Call,Allele 2 Call
MUESTRA1234,XXX09,GSTT1_d,C,C
MUESTRA1234,XXX09,GSTT1,C,C
MUESTRA1234,XXX09,rs1042713,A,A
MUESTRA1234,XXX09,rs1042714,C,C
NTC,XXX11,GSTT1_d,,
NTC,XXX11,GSTT1,,
NTC,XXX11,rs1042713,,
NTC,XXX11,rs1042714,,
```

3.2.3 SIGNATURIT

El proceso de firma digital que se realiza para los usuarios que encargan un estudio de salud personalizada *Made of Genes* (MoG) se lleva a cabo a través de Signaturit (https://www.signaturit.com/), un servicio de firmas electrónicas que permite enviar por correo electrónico el consentimiento informado y la petición para el laboratorio para que el usuario los firme; permite además seguir el proceso de firma a tiempo real y descargar un documento acreditativo de trazabilidad de todo el proceso.

3.2.4 REPORTLAB

La generación de informes de resultados se realiza de manera automatizada a partir de unos datos de entrada, específicos para cada análisis, mediante una aplicación de BIOMED que contiene la librería Reportlab (https://www.reportlab.com/), concretamente, usando su metodología *Platypus* (https://docs.reportlab.com/reportlab/u

serguide/ch5_platypus/) que permite organizar todos los elementos a mostrar en el informe en formato PDF de manera sencilla y programática, creando así documentos personalizados hasta el más mínimo detalle respecto al contenido del informe.

3.2.5 Genomcore FRONTDESK y aplicación web

Además de BIOMED, Genomcore cuenta también con la plataforma FRONTDESK (sec. 1.6), que permite a los diferentes proveedores y colaboradores de la empresa la creación y personalización de aplicaciones orientadas al usuario final. Para la marca MoG de la empresa, se ha usado esta plataforma para crear y publicar una aplicación web, además de sus respectivas aplicaciones móviles para `Android` y `iOS`, que permiten la consulta y gestión de los resultados de un análisis de salud personalizada, centrándose en una experiencia de usuario lo más sencilla y directa posible (https://app.madeofgenes.com). Esta aplicación web se integra con las diferentes APIs de BIOMED (sec. 3.1.2) para mostrar toda esa información en una visualización más detallada que en la plataforma, orientada al usuario, y con un control de acceso mediante correo electrónico y contraseña.

CAPÍTULO 4

Resultados

Este capítulo presenta en dos secciones diferentes los sistemas automatizados desarrollados durante este proyecto: el VPMS es una herramienta de generación y selección de paneles virtuales de genes para optimizar los análisis NGS a las características específicas de cada paciente; la RGT es una herramienta de generación de informes de resultados a partir de un análisis, ya sean NGS u otras tipologías de análisis desarrollados dentro de la empresa como, por ejemplo, informes de salud personalizada.

4.1 Sistema de gestión de paneles virtuales de genes

El *Virtual Panel Management System* o sistema de gestión de paneles virtuales (VPMS) se ha desarrollado para proveer una solución automatizada de generación de paneles virtuales que permitan optimizar e incrementar la coste-eficiencia de los análisis NGS dirigidos a regiones genómicas específicas. De esta manera, el VPMS se ha integrado al *pipeline* bioinformático existente en la empresa (secs. 4.1.1, 4.1.3) en forma de aplicaciones que generan paneles virtuales a partir de un listado de genes objetivo

(sec. 4.1.4), y permiten seleccionarlos mediante identificadores que integran el módulo de tareas de la plataforma con el de Records para obtener de manera automatizada la información necesaria para realizar un análisis (sec. 4.1.5).

4.1.1 Pipeline bioinformático de análisis NGS

La empresa dispone de un *pipeline* propio de análisis NGS para variantes germinales, desarrollado dentro de la plataforma BIOMED, que está formado actualmente por los siguientes *workflows* (fig. 4.1):

- ***NGS LIMS Importer***: contiene solo una aplicación (con el mismo nombre) que importa, de manera periódica, la información almacenada en un *Laboratory Information Management System* o sistema de gestión de información de laboratorio (LIMS) sobre los pacientes y las peticiones en forma de Records a BIOMED (figs. 4.2a, 4.3).
- ***NGS Run Processing***: está constituido por dos aplicaciones (fig. 4.2b):
  - *BCL to FASTQ*: se encarga de demultiplexar los datos en bruto obtenidos de la máquina de secuenciación, que están en formato BCL, a ficheros FASTQ para cada muestra demultiplexada. Para ello, requiere de un fichero con la relación de muestras secuenciadas, subido a BIOMED de manera manual por un analista (fig. 4.1).
  - *NGS Records Importer*: aplicación que genera en BIOMED todos los Records relacionados con el experimento NGS específico de la ejecución (fig. 4.3), además de ejecutar automáticamente el *workflow* de detección de variantes.
- ***NGS Germline Variant Calling***: es el *workflow* encargado de la detección de variantes germinales, y está compuesto por cuatro aplicaciones (figs. 3.5, 4.2c):

  - *Mapping (Step A)*: realiza el alineamiento de las lecturas de secuenciación en formato FASTQ contra un genoma de referencia, generando así un fichero BAM.
  - *Germline Variant Calling (Step B)*: realiza la identificación de variantes germinales en las regiones de interés a partir de un fichero BAM de entrada, generando como fichero de salida un VCF anotado y un *genomic variant call format* (gVCF) sin anotar.
  - *Quality Metrics (Step B-2)*: produce estadísticas de cobertura de la secuenciación, del alineamiento y de lecturas duplicadas en formato Excel.
  - *Germline Variant Classification (Step C)*: realiza el filtraje y priorización de variantes del VCF generado en el paso B, y sube las variantes relevantes al módulo de BIOMED (sec. 4.1.6).

- ***NGS Study Report Generation***: genera los informes clínicos de resultados, que se envían de vuelta al LIMS una vez están validados. Es un *workflow* lineal formado por cuatro aplicaciones (fig. 4.2d):

  - *NGS Process Study*: analiza los Records de un análisis NGS generados hasta el momento para decidir si ya se puede proceder con la generación del informe de resultados.
  - *NGS Report Generator*: genera un informe de resultados, en formato borrador o definitivo dependiendo de lo que indiquen los parámetros de entrada.
  - *Update Study*: enlaza el informe generado con su Record Estudio (fig. 4.3).
  - *LIMS Report Uploader*: se encarga de enviar el informe de resultados validado a través del LIMS correspondiente (sec. 4.2.3).

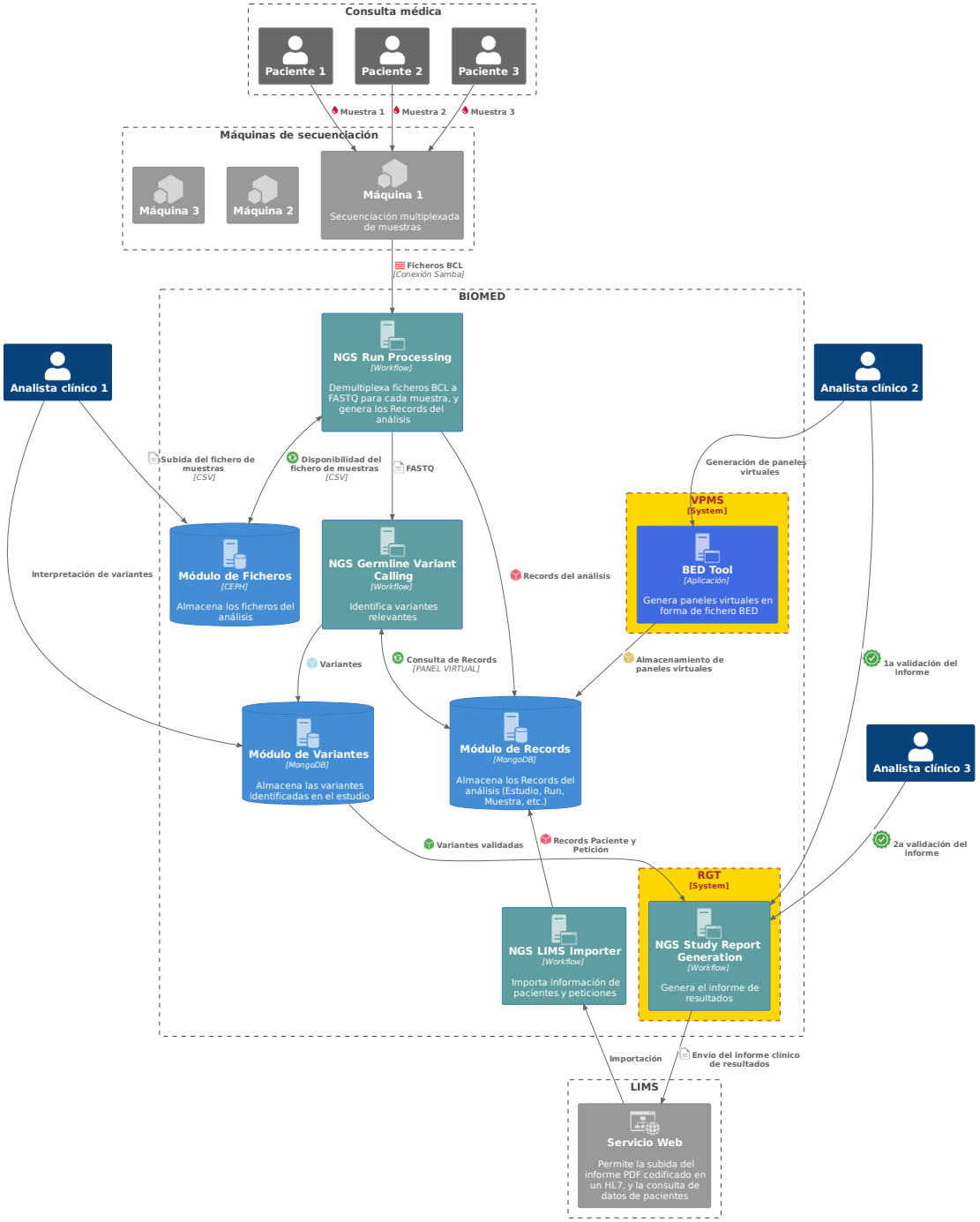
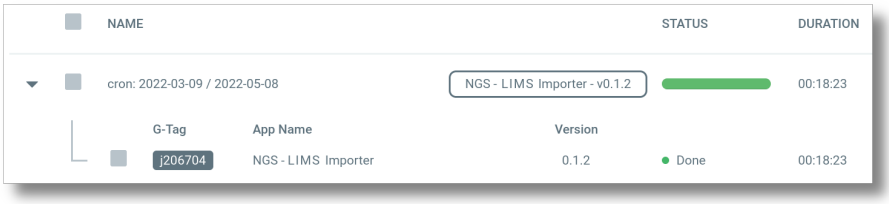
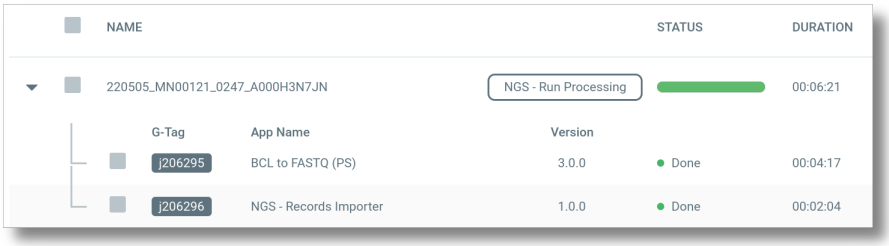


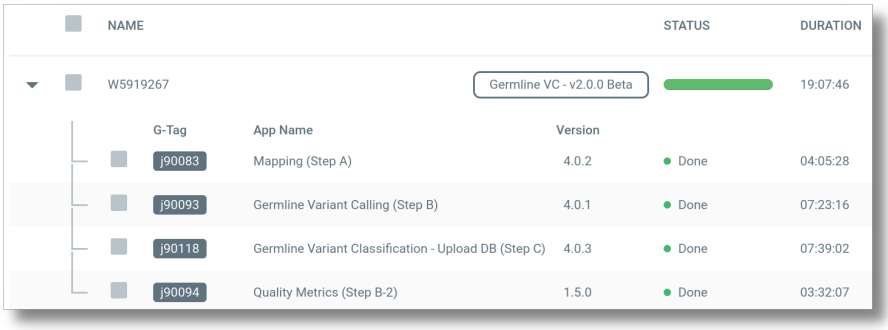
FIGURA 4.1: *Flujo de datos para un análisis NGS completo; en amarillo se destacan las herramientas desarrolladas como objetivo principal en esta tesis. Los datos en bruto se obtienen de las máquinas de secuenciación del laboratorio, conectadas a BIOMED mediante un directorio Samba; estos a su vez se transforman mediante la demultiplexación a ficheros FASTQ separados por muestras, utilizados en los workflows de la plataforma (junto con los paneles virtuales generados con el **VPMS**) para identificar las variantes relevantes, que serán interpretadas y validadas por analistas clínicos externos. Una vez generado el informe de resultados con la **RGT** y validado por analistas, se enviará al laboratorio a través de su LIMS.*



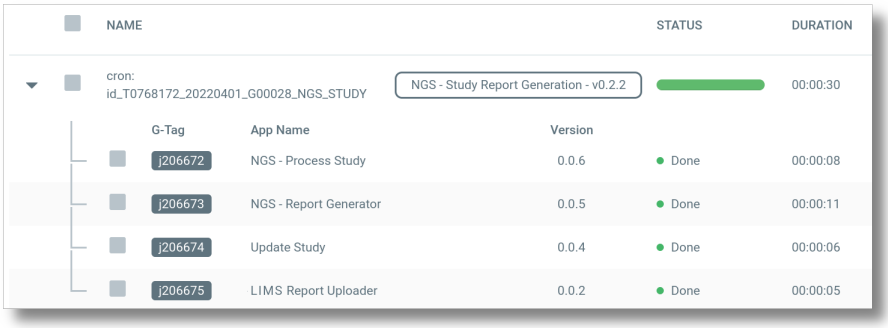
(A) *Workflow de importación de pacientes y peticiones del LIMS de un laboratorio.*



(B) *Workflow de procesamiento de los datos en bruto obtenidos de la máquina de secuenciación, formado por dos aplicaciones, una que transforma ficheros BCL a FASTQ y otra que genera Records.*



(C) *Workflow de identificación de variantes germinales, formado por aplicaciones que se encargan de realizar el análisis secundario: alineamiento con la referencia, identificación y clasificación de variantes, y control de calidad.*



(D) *Workflow de generación del informe de resultados, formado por aplicaciones que evalúan el Record Estudio, generan el informe PDF, lo enlazan al Estudio, y lo envían al LIMS correspondiente.*

FIGURA 4.2: *Workflows de BIOMED para un análisis NGS y la generación de resultados.*

Las aplicaciones que conforman cada *workflow* han sido creadas con una combinación de diferentes lenguajes de programación:

- `Python` (Rossum y Boer, 1991) para el desarrollo de las librerías necesarias para la comunicación con las APIs de BIOMED (sec. 3.1.2, tbl. 4.1).
- `R` (Ihaka y Gentleman, 1996; R Core Team, 2020) y ***GNU Bourne-Again Shell* (Bash)** (https://www.gnu.org/software/bash/) para las librerías específicas de procesamiento de datos NGS (tbl. 4.2), accesibles desde el sistema operativo Linux, y avaladas por la comunidad científica (Moorthie, Hall y Wright, 2013; Pabinger *et al.*, 2014; Ritchie y Flicek, 2014; Oliver, Hart y Klee, 2015).
- ***Common Workflow Language* (CWL)** para diseñar y organizar los diferentes pasos de cada aplicación que conforma un análisis NGS usando un lenguaje de código abierto y estándar (Amstutz *et al.*, 2016).

TABLA 4.1: *Librerías de Python creadas para las APIs de BIOMED.*

Herramienta	Descripción
auth-client	Interactúa con la API de autenticación.
biomed-client	Interactúa con múltiples APIs; la de Records, la de Variantes, la de correos electrónicos, la de tareas, y la de subida y descarga de ficheros.
biovalues-client	Interactúa con la API de *Biovalues* (otro tipo concreto de Record).
file-client	Interactúa con los diferentes sistemas de ficheros de los colaboradores de BIOMED para recuperar los archivos necesarios para los análisis, ya sean en un directorio remoto usando el *File Transfer Protocol* o protocolo de transferencia de archivos (FTP), el *Secure File Transfer Protocol* o protocolo seguro de transferencia de archivos (SFTP), el *Amazon Simple Storage Service* (S3) de AWS, o como servicio web.
insights-client	Interactúa con la API de *Insights* (un tipo concreto de Record).
mogops-client	Interactúa con la API de MoG *Operations*, una API interna (no disponible para usuarios externos de BIOMED) diseñada para facilitar operaciones de recuperación de datos necesarios para distintas aplicaciones de nuestros productos y estudios MoG.
variants-client	Interactúa con la API de variantes genéticas.

TABLA 4.2: *Herramientas usadas para el análisis NGS.*

Herramienta (Versión)	Descripción	Fuente
bcl2fastq (2.19.1)	Conversión de ficheros BCL (multiplexados) a ficheros FASTQ (demultiplexados), y generación de métricas de calidad.	https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html
trimmomatic (0.38)	Eliminación de adaptadores (usados en las máquinas de secuenciación Illumina) de los ficheros FASTQ.	http://www.usadellab.org/cms/?page=trimmomatic
fqtools (2.0)	Recuento de número de *reads* en los ficheros FASTQ.	https://github.com/alastair-droop/fqtools
bwa (0.7.17)	Alineamiento de los *reads* de los ficheros FASTQ contra el genoma de referencia.	https://github.com/lh3/bwa
picard (2.18.14)	Identificación de *reads* duplicados en ficheros BAM.	http://broadinstitute.github.io/picard/
samtools (1.9)	Manipulación de ficheros BAM.	https://github.com/samtools/samtools
bedtools (2.27.1)	Manipulación de ficheros BED y cálculo de cobertura genómica de ficheros BAM.	http://bedtools.readthedocs.org/
gatk4 (4.1.0.0)	Detección de SNPs e indels a partir de ficheros BAM usando *HaplotypeCaller*, recalibración de puntuaciones de calidad de cada base usando *BaseRecalibrator* y *ApplyBQSR*, filtraje de variantes de interés usando *SelectVariants* y *VariantFiltration*.	https://github.com/broadinstitute/gatk
tabix (0.2.6)	Indexación de ficheros BED y VCF.	https://github.com/samtools/htslib
bcftools (1.9)	Manipulación y compresión de ficheros VCF.	https://github.com/samtools/bcftools
vep (104)	Anotación de variantes identificadas en ficheros VCF.	http://www.ensembl.org/info/docs/tools/vep/index.html

Toda la información relevante para el análisis, desde los metadatos y librerías de la secuenciación ejecutada y la muestra a analizar, hasta el estudio realizado y el informe de resultados, se almacena en Records estructurados y relacionados entre sí. Estos Records tienen diferentes orígenes (fig. 4.3):

- Una parte procede de la importación de datos clínicos a través del LIMS de cada laboratorio colaborador (fig. 4.2a).
- Otros se generan a partir de los metadatos obtenidos de las máquinas de secuenciación (fig. 4.2b).
- Finalmente, hay Records que se tienen que importar manualmente a partir de la información en los catálogos de pruebas diagnósticas de cada laboratorio, las librerías que usan y los analistas que trabajan allí.

De esta manera, se garantiza la trazabilidad de los análisis, ya que toda esta información queda almacenada en la plataforma y a disponibilidad de los usuarios autorizados para su consulta posterior en caso de revisión o validación, o en caso de valorar si se tiene que repetir el análisis.

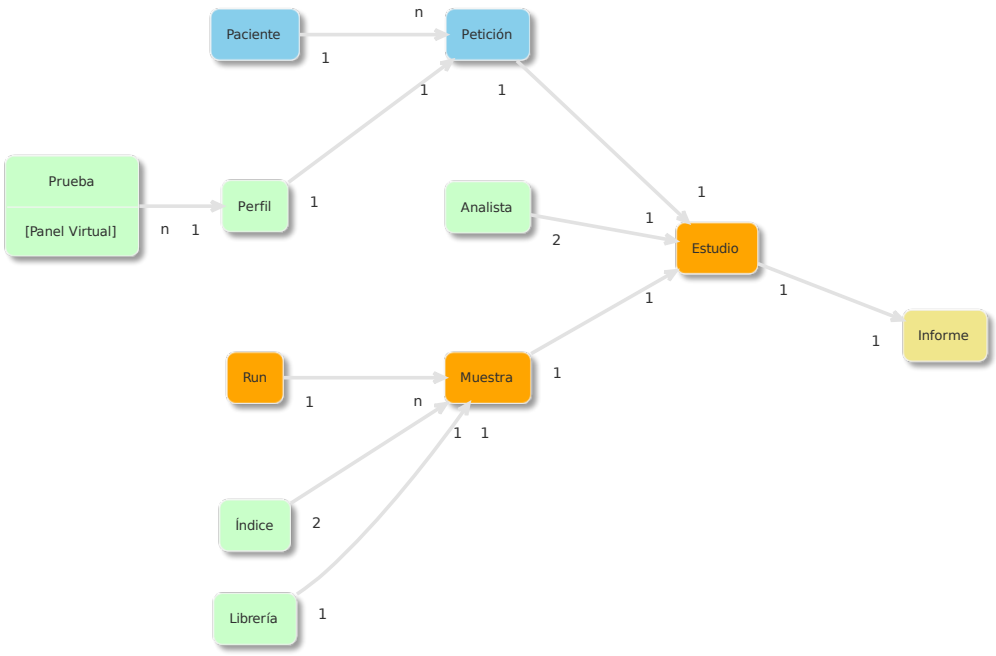


FIGURA 4.3: *Organización de la base de datos de Records para un análisis NGS, dependiendo si el origen de los datos es a través de una importación manual (color verde), del workflow de importación automatizada de Records a partir de datos del LIMS (azul), del workflow 'NGS Run Processing' (naranja) o del workflow 'NGS Study Report Generation' (amarillo). El Record Perfil puede contener múltiples Pruebas, de manera que en un Estudio con un solo informe de resultados se pueden mostrar a la vez resultados de diferentes técnicas: NGS, Sanger, y MLPA.*

4.1.2 Importación de datos

Los datos de entrada del VPMS son los ficheros en formato BCL generados por las máquinas de secuenciación de proveedores externos, que contienen los datos multiplexados. La transferencia de los ficheros desde las máquinas hasta nuestra plataforma de análisis se realiza mediante un directorio compartido con el protocolo Samba; tanto las máquinas de secuenciación como nuestra plataforma BIOMED tienen acceso a ese directorio, de manera que todos los datos generados por éstas se van guardando allí, y una vez han terminado, un proceso interno de la empresa detecta el estado completado de la máquina y sube los ficheros a la plataforma (fig. 4.1).

4.1.3 Implementación

Actualmente, el VPMS permite la creación de paneles virtuales de genes en forma de Records, enlazados a un fichero BED generado mediante una aplicación de la plataforma (sec. 4.1.4), que se utilizan posteriormente como datos de entrada de un análisis para determinar las *Regions Of Interest* o regiones de interés (ROIs), las regiones genómicas que se desea estudiar, con la finalidad de reducir el número de variantes identificadas para su interpretación clínica y posterior reporte al paciente por parte del profesional clínico involucrado, que puede ser externo o no a la empresa (fig. 4.1). La funcionalidad adicional que aporta el VPMS está integrada en el *pipeline* bioinformático de análisis NGS que existe en BIOMED, detallado al inicio de este capítulo (sec. 4.1.1).

4.1.4 Creación de paneles virtuales

La creación de los paneles virtuales de genes se realiza por parte de los analistas clínicos mediante la ejecución de una aplicación de BIOMED diseñada con ese fin, la *BED Tool*, que requiere de cuatro parámetros de entrada (fig. 4.4):

- Una lista de transcritos génicos en nomenclatura *Reference Sequence* (RefSeq); opcionalmente, si no se conoce el nombre del transcrito se puede proporcionar también el nombre del gen, y la aplicación se encarga de reportar las *coding sequences* o regiones codificantes de un gen (CDS) de todos los transcritos asociados a ese gen (fch. 4.1).
- El número de bases a usar para el *padding*, es decir, el número de bases adicionales incluidas en ambos extremos de cada CDS.
- El directorio de BIOMED donde almacenar el fichero BED generado.
- El fichero de correspondencia de nombres de transcritos con nombres de genes usando la referencia `hg19`, obtenido a partir de la base de datos de anotación genómica `GRCh37` (sec. 1.2.3.2), accesible de manera pública (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/).

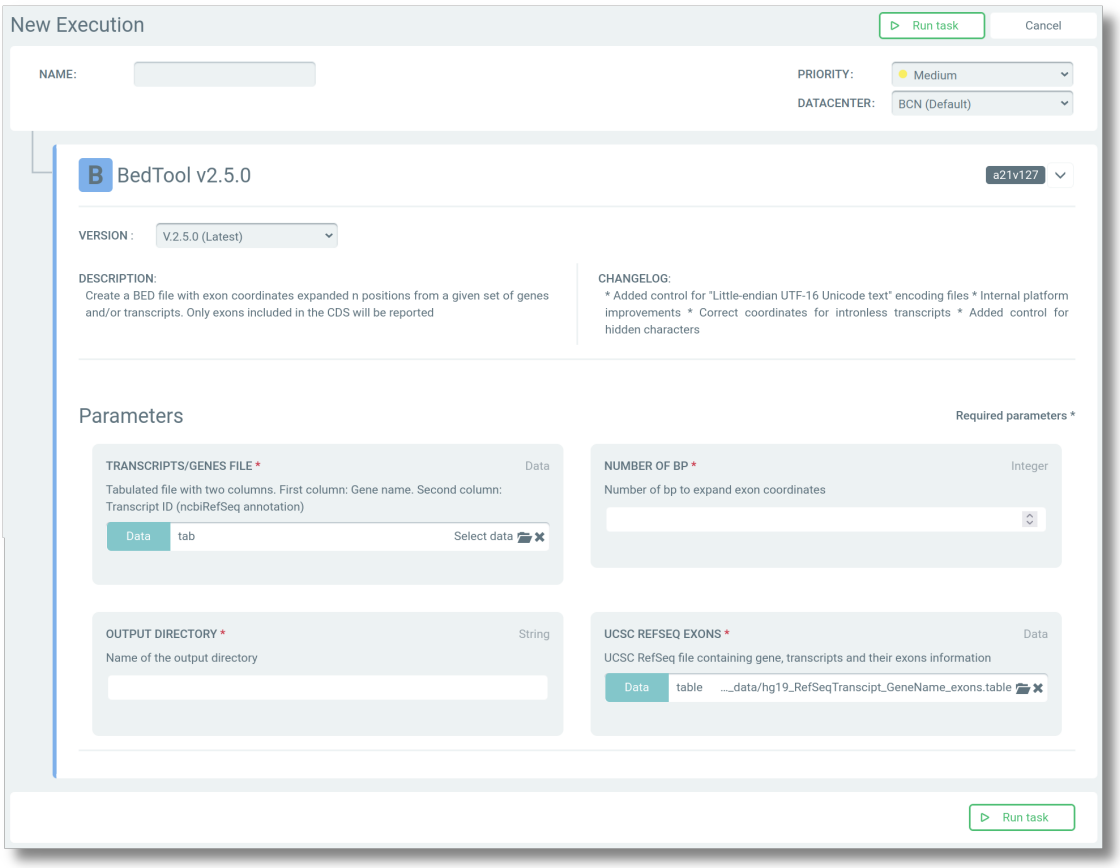


FIGURA 4.4: *La BED Tool es la aplicación de generación de paneles virtuales en formato BED.*

La aplicación se encarga de generar ficheros BED que contienen las CDS, es decir, las coordenadas exónicas de los genes candidatos (fch. 4.2), y lo sube al módulo de ficheros de la plataforma; internamente, se usa el lenguaje de programación `Awk` (Aho, Kernighan y Weinberger, 1979) para procesar los datos de entrada, corregirlos en caso de errores de nomenclatura, y ordenar las coordenadas resultantes para generar el fichero de salida.

FICHERO 4.1: *Fichero de entrada de la BED Tool.*

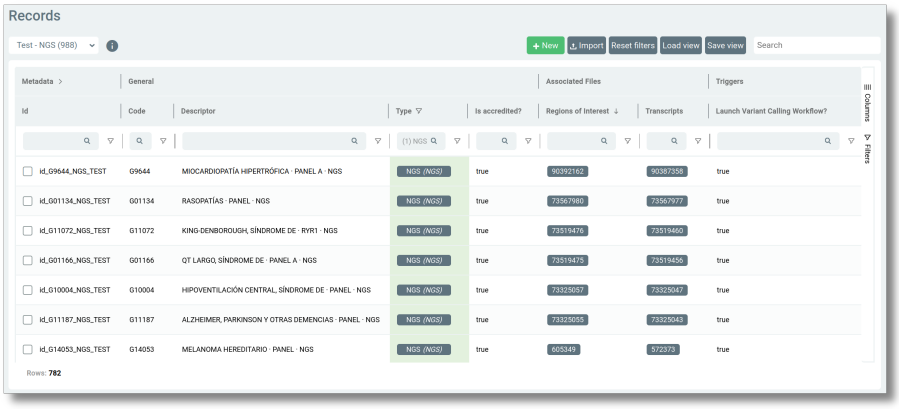
```
MLH1    NM_000249.3
MSH2    NM_000251.2
MSH6    NM_000179.2
PMS2    NM_000535.6
RET     NM_020975.5
APC     NM_000038.5
ATM     NM_000051.3
BMPR1A  NM_004329.2
BRCA1   NM_007294.3
BRCA2   NM_000059.3
CDKN2A  NM_000077.4
KRAS    NM_004985.4
MEN1    NM_130799.2
PALB2   NM_024675.3
```

FICHERO 4.2: *Fichero de salida de la BED Tool, con información adicional en la cabecera.*

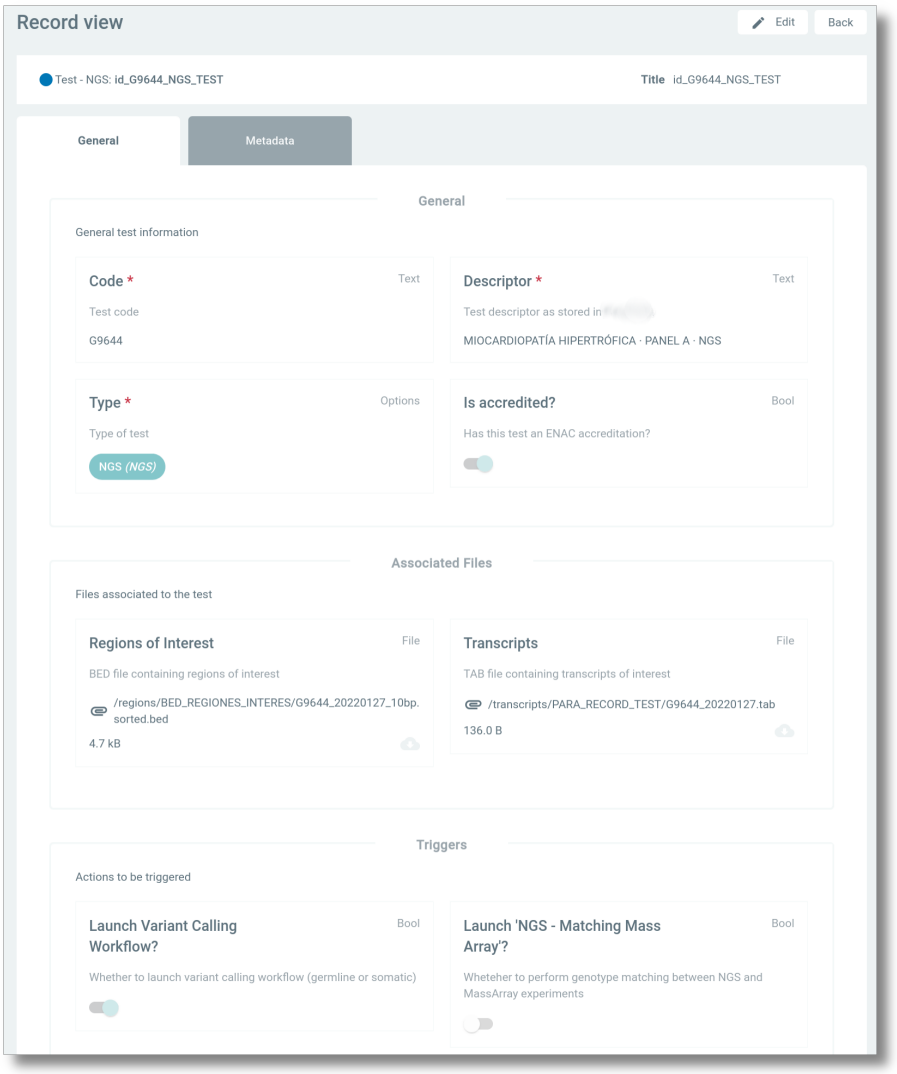
```
#2022-08-24 08:03:24
#BED file created with BedTool (Version:v2.5.0)
#Number of bp added: 21
10  43572685    43572800    RET,NM_020975.4
# [...]
10  43623538    43623738    RET,NM_020975.4
10  88635754    88635863    BMPR1A,NM_004329.2
# [...]
10  88683329    88683497    BMPR1A,NM_004329.2
11  64571784    64572309    MEN1,NM_130799.2
# [...]
11  64577115    64577602    MEN1,NM_130799.2
11  108098330   108098444   ATM,NM_000051.3
# [etc.]
```

El fichero BED generado se enlaza entonces a un Record de tipo Prueba, concretamente en un campo definido para las ROIs (figs. 4.5a, 4.5b), que se usará posteriormente en los análisis NGS como dato de entrada. Los Records Prueba y Perfil se crean o actualizan de manera manual a partir de los catálogos de pruebas diagnósticas de cada laboratorio (fig. 4.3). Las Pruebas en BIOMED pueden ser de tres tipos diferentes: NGS, Sanger o *Multiplex Ligation-dependent Probe Amplification* (MLPA).

Algunos ejemplos de los paneles virtuales generados con esta aplicación se muestran en la fig. 4.6 mediante el *Integrative Genomics Viewer* (IGV), mostrando así los diferentes tamaños que puede tomar un estudio genómico realizado dentro de BIOMED.



(A) *Listado de Pruebas, filtradas por las de tipo NGS.*



(B) *Detalle de un Record de tipo Prueba.*

FIGURA 4.5: *Record Prueba en BIOMED.*

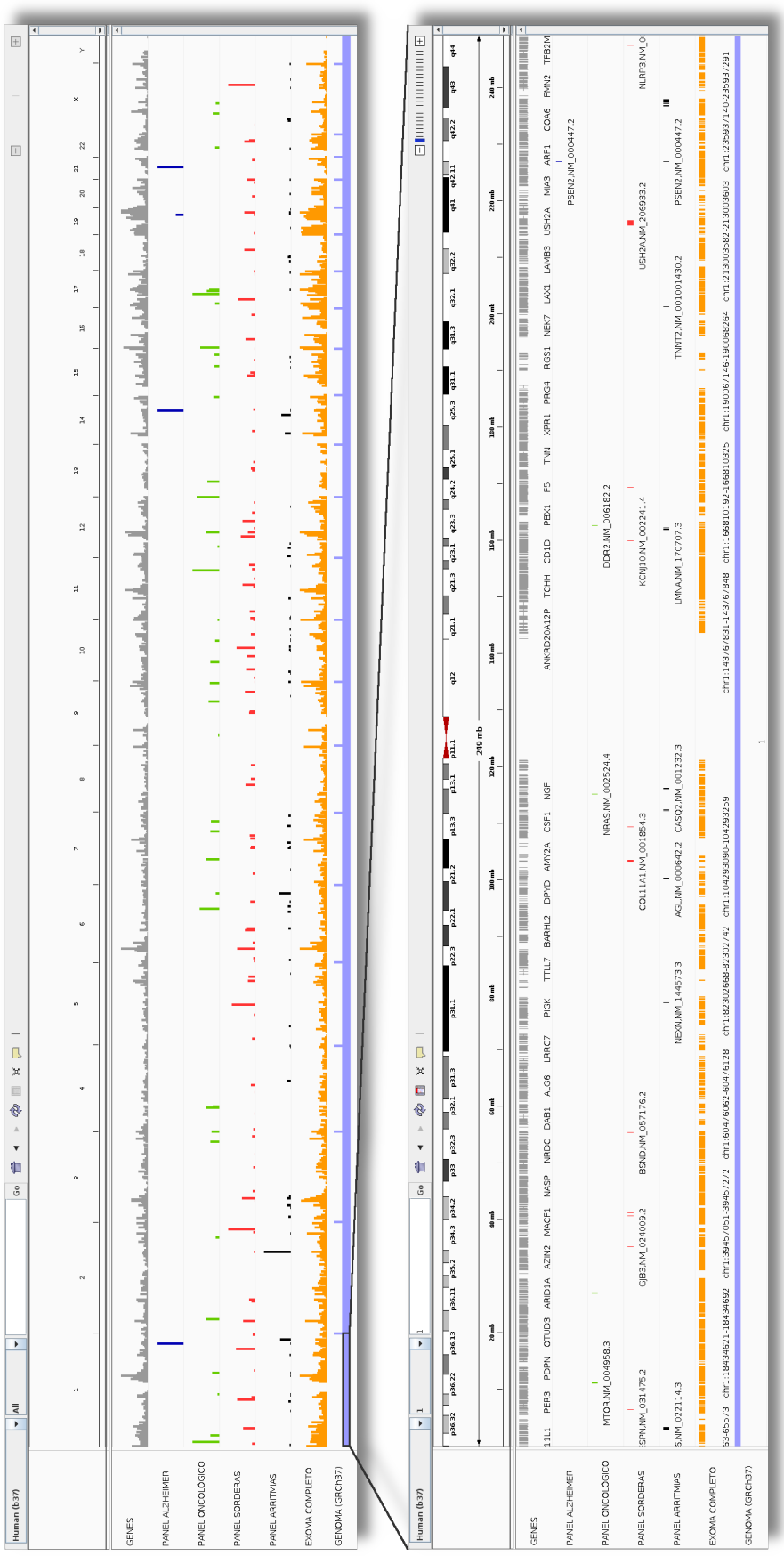


FIGURA 4.6: *Ficheros BED generados con la aplicación BED Tool y visualizados en el IGV. Imagen superior: visualización a lo largo de todo el genoma de diferentes ficheros BED; en azul, un panel de Alzheimer; en verde, un panel oncológico; en rojo, un panel de sordera; en negro, un panel de arritmias; en naranja, un exoma completo. Imagen inferior: la misma visualización, ampliada al cromosoma 1, para ver el detalle de los transcritos y regiones incluidas en cada fichero BED. En ambas imágenes, la primera pista de color gris corresponde a los nombres de los genes, y la última pista de color violeta muestra el fichero BED representativo de todo el genoma humano en la versión GRCh37.*

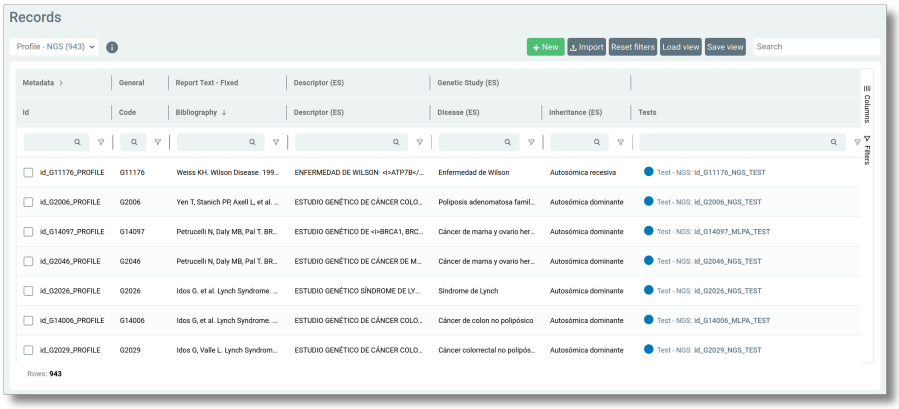
4.1.5 Selección de paneles virtuales

La selección de un panel virtual se realiza de manera automatizada en el *pipeline* bioinformático de análisis NGS de BIOMED; concretamente, se utiliza como uno de los datos de entrada en el *workflow* de identificación de variantes, el *NGS Germline Variant Calling* (figs. 4.1, 4.2c). Utiliza el fichero de muestras almacenado en BIOMED (sec. 4.1.1, fig. 4.1), que contiene los identificadores que deben coincidir con el que está especificado en el campo Código de un Record Perfil (figs. 4.7a, 4.7b), para obtener el fichero BED enlazado en el Record Prueba de tipo NGS correspondiente (figs. 4.5a, 4.5b). De esta manera, se realiza el enlace automatizado entre el análisis NGS en curso y el panel virtual generado con el VPMS. El alcance genómico del *workflow* de identificación de variantes se determina de esta forma en función del Record Librería para las regiones de cobertura del análisis, y del Record Perfil para las ROIs en las que llevar a cabo la detección de variantes (fig. 4.3).

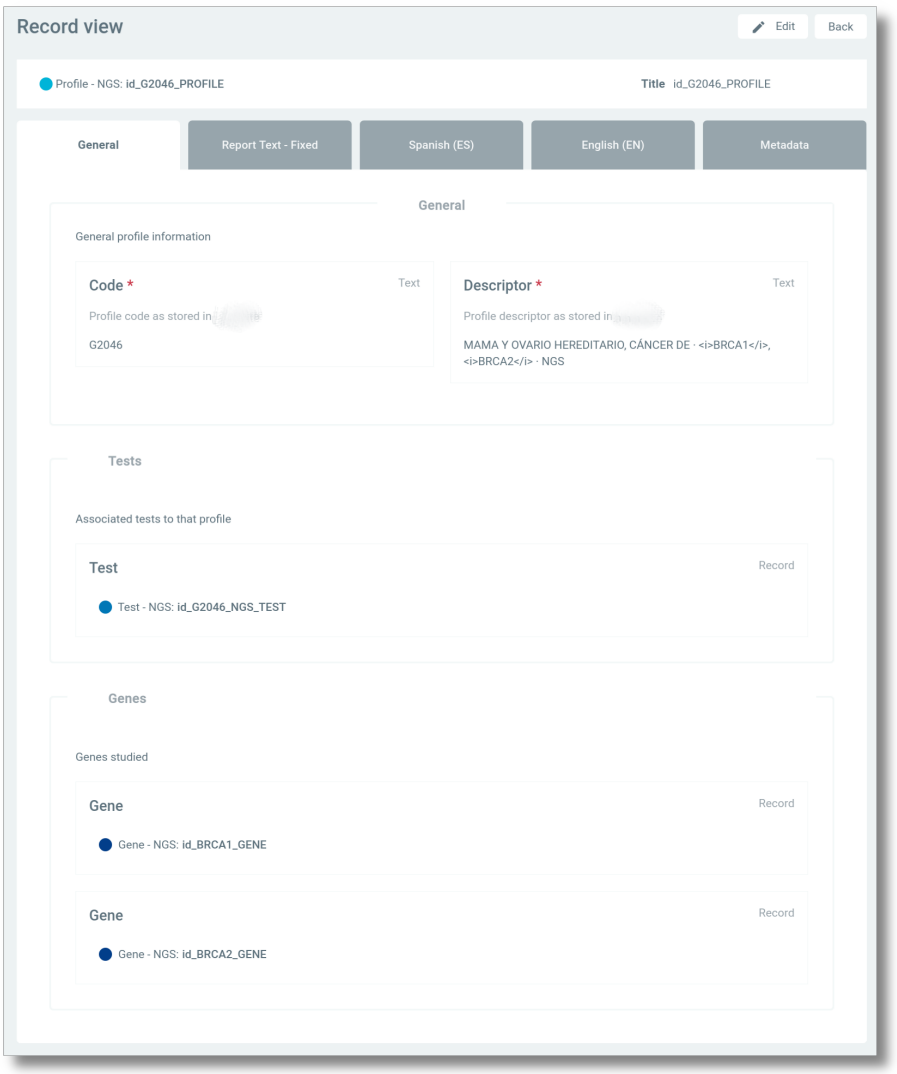
4.1.6 Almacenamiento de variantes

Las variantes identificadas en el *pipeline* bioinformático se suben a BIOMED a través del **módulo de variantes**, que usa la misma infraestructura que el módulo de Records (sec. 3.1.7), con la única diferencia que las variantes no son Records basados en una plantilla, sino que se almacenan directamente en una base de datos de MongoDB. Este módulo permite almacenar las variantes identificadas en cualquier tipo de análisis que genere variantes (*microarrays* de genotipado, NGS, etc.) para su posterior consulta y anotación por parte de analistas clínicos con acceso a BIOMED. Como en el caso de los ficheros y los Records, la comunicación entre la base de datos y la plataforma también se realiza con una API específica; su gestión y visualización en BIOMED funciona de igual manera que con los Records genéricos (fig. 4.8).

En el caso de los análisis NGS, para la subida masiva de todas las variantes que se pueden identificar en un experimento, el *workflow NGS Germline Variant Calling*



(A) *Listado de Perfiles usados por un laboratorio.*



(B) *Detalle de un Record de tipo Perfil.*

FIGURA 4.7: *Record Perfil en BIOMED.*

(fig. 4.2c) se encarga internamente de transformar los datos del fichero VCF generado a un fichero en formato *JSON Lines* (https://jsonlines.org/), que facilita su lectura y subida contra la API de variantes en *streaming*.

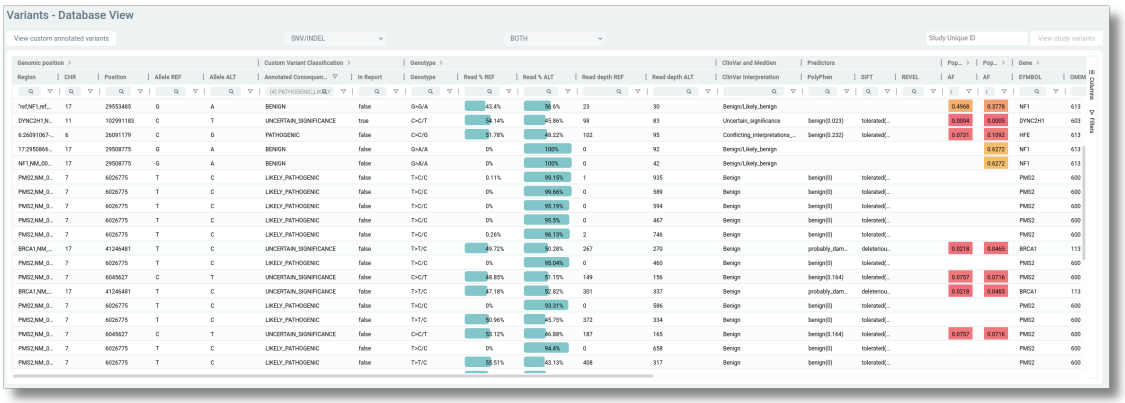
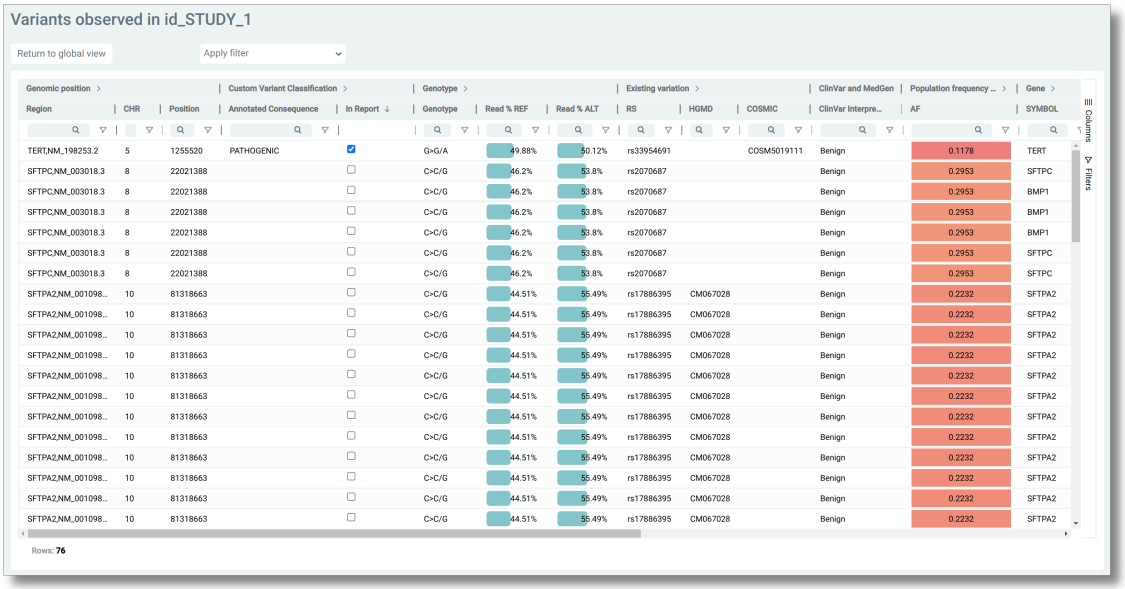


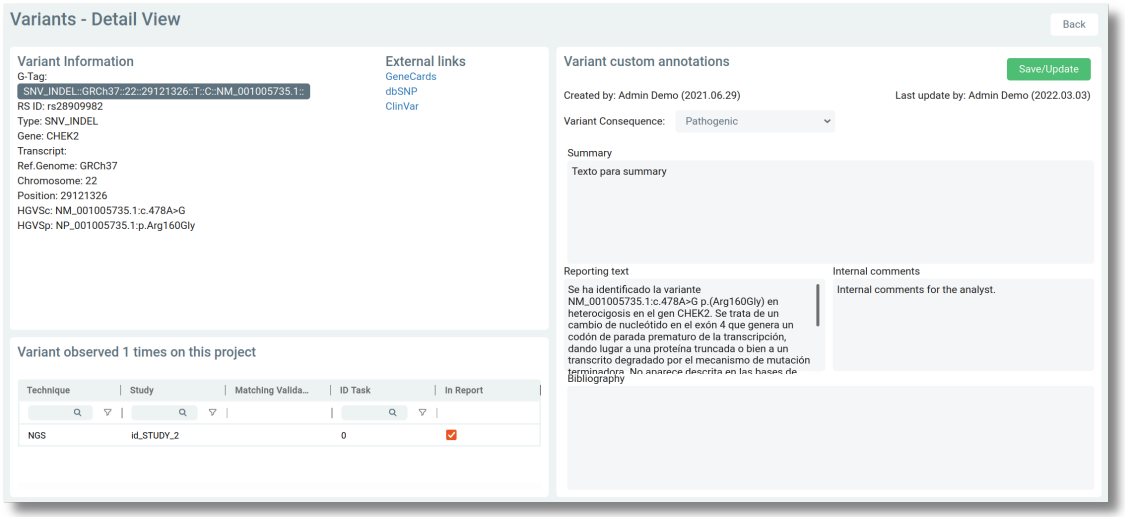
FIGURA 4.8: *Módulo de variantes de BIOMED, en el que se visualizan todas las variantes almacenadas para un proyecto concreto.*

4.1.7 INTERPRETACIÓN DE RESULTADOS

Una vez finalizado el *workflow* de identificación de variantes, el proceso se habrá encargado de almacenar las variantes relevantes, que se pueden visualizar en forma de listado, filtradas por el estudio específico (fig. 4.9a). BIOMED proporciona también una vista en detalle de cada una de ellas, para poder consultar información adicional de bases de datos públicas como dbSNP y ClinVar, y realizar anotaciones manuales personalizadas (fig. 4.9b) que se almacenarán en la **base de datos interna de variantes** (sec. 4.1.6). Este módulo es el componente central de la plataforma que permite a los analistas decidir el impacto sobre la salud del paciente de cada variante identificada, y seleccionarla o descartarla para su posterior incorporación en el informe clínico de resultados (sec. 4.2.3). La metodología de validación clínica de cada variante la decide el laboratorio externo encargado del análisis basándose en la frecuencia de variantes observadas, siendo posible combinar el uso de secuenciación Sanger con NGS para validarlas en un solo informe final de resultados.



(A) *Visualización de las variantes identificadas para un estudio concreto.*



(B) *Detalle de una variante concreta, con anotaciones personalizadas y editables por analistas clínicos con acceso a la plataforma.*

FIGURA 4.9: *Visualización y edición de variantes en BIOMED.*

4.1.8 Reanálisis de datos

Tanto la integración del VPMS como el almacenamiento a largo plazo de ficheros y datos relevantes para un análisis llevado a cabo en BIOMED (secs. 3.1.6, 3.1.7) facilitan el reanálisis de datos NGS, ya sea por la aparición de nueva evidencia científica que pueda ayudar a resolver diagnósticos incompletos, como por la actualización del *pipeline* bioinformático utilizado. Actualmente, el reanálisis de datos en la plataforma se produce a partir de los ficheros BAM, con un *workflow* diseñado para ese fin (fig. 4.10); estos ficheros se mantienen almacenados en la plataforma y son fácilmente identificables a partir de los Records que se generaron durante el análisis inicial (fig. 4.3).

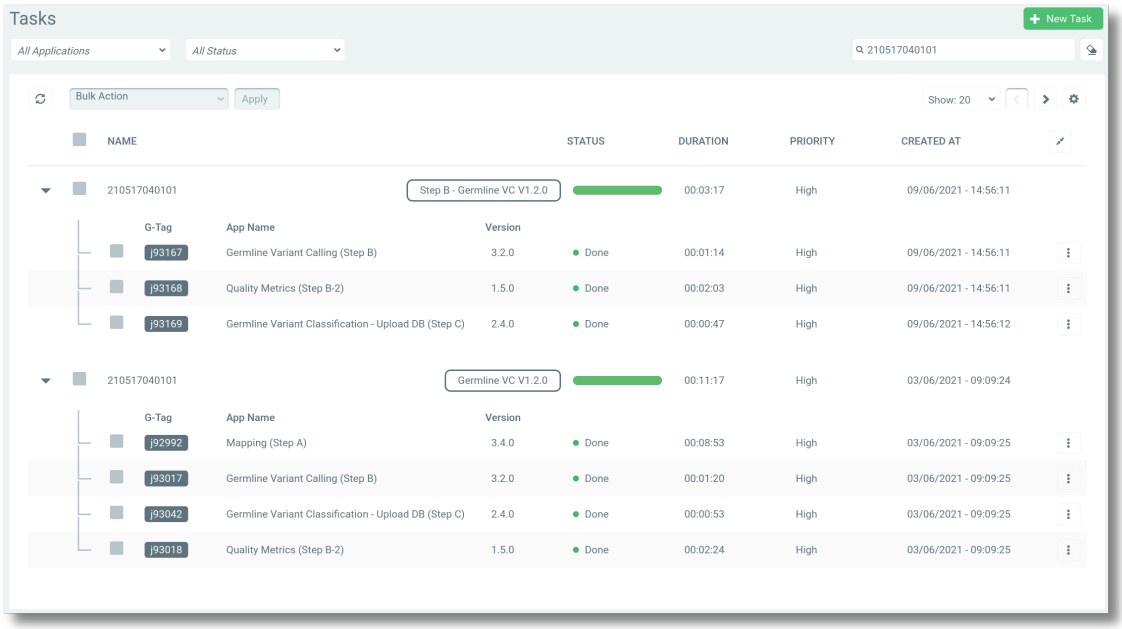


Figura 4.10: *Reanálisis de datos NGS en BIOMED a partir del fichero BAM; en la parte inferior se puede observar el workflow ejecutado inicialmente, que consta de cuatro aplicaciones y dura 11 minutos en total, y en la parte superior el workflow de reanálisis, que consta solo de tres aplicaciones porque ya no necesita realizar el alineamiento inicial contra el genoma de referencia, y reduce su ejecución a 3 minutos.*

4.2 Herramienta de generación de informes

La *Report Generation Tool* o herramienta de generación de informes (RGT) es una herramienta automatizada para generar informes de resultados a partir de los datos de un análisis, ya sean NGS u otras tipologías de análisis realizados dentro de la empresa (sec. 4.2.1). La tipología de informes generados con la RGT conforman dos categorías: los informes clínicos (sec. 4.2.3), focalizados en análisis NGS, y los informes de salud personalizada (sec. 4.2.4), que se generan con datos de diferentes orígenes y están más orientados a la prevención. En ambos casos, la RGT constituye una aplicación de BIOMED que genera informes de resultados en formato PDF a partir de datos dinámicos de entrada usando la librería Reportlab (sec. 3.2.4). Como se verá más adelante, se trata de un producto listo para entornos de producción y validado; la hoja de ruta actual de Genomcore y de cara al futuro es desarrollar un sistema aún más flexible, adaptado al uso incremental de los teléfonos móviles inteligentes como herramienta para visualizar los resultados de un estudio de una manera más dinámica que la que ofrece un fichero estático en formato PDF (sec. 4.2.5).

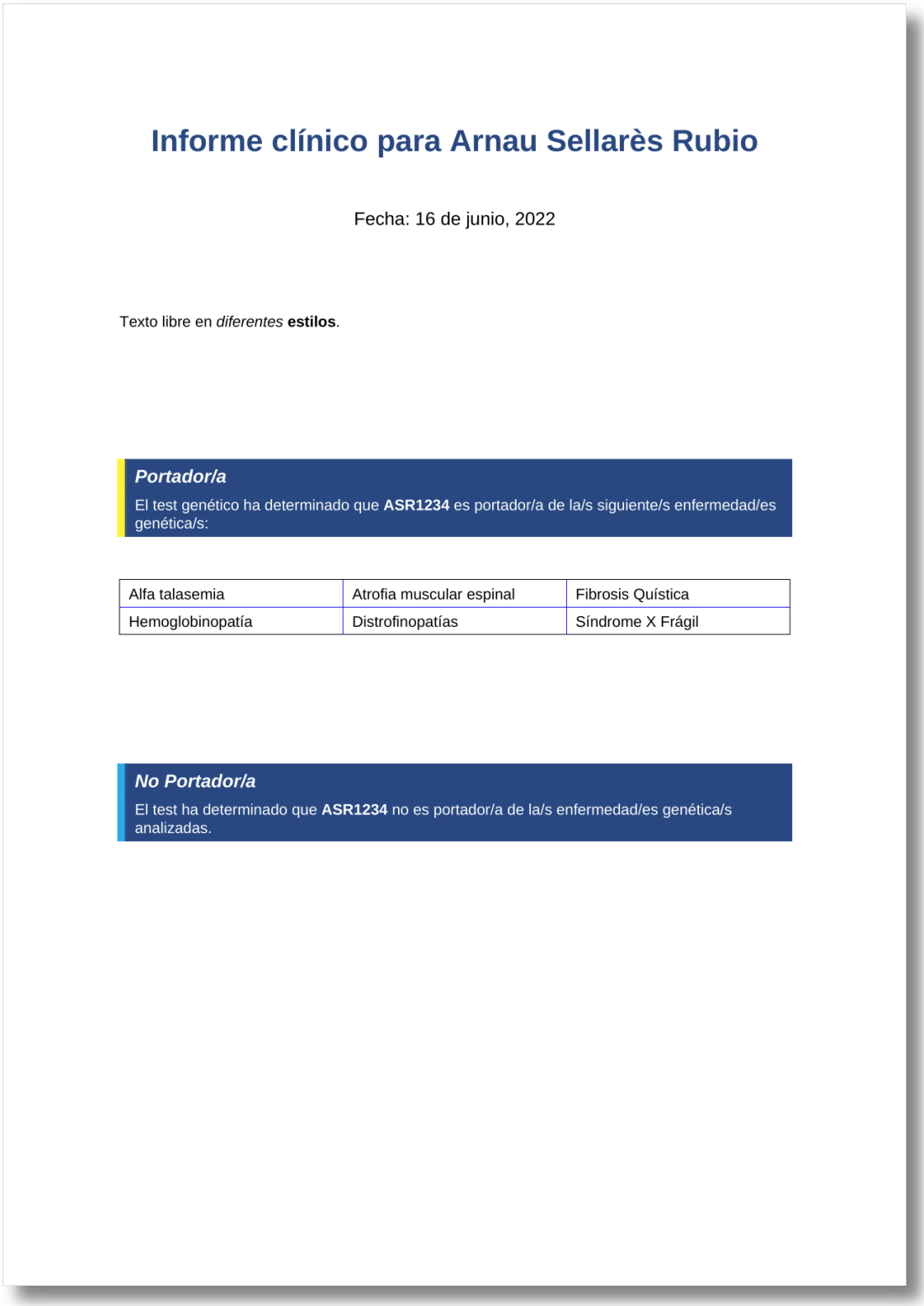
4.2.1 Implementación

La RGT se ha diseñado como una aplicación que, a partir de unos datos de entrada variables para cada ejecución, genera informes en formato PDF. Para lograrlo se ha usado Reportlab (sec. 3.2.4) como librería base para diseñar en `Python` el programa que permite la generación automatizada de los informes con mucha flexibilidad; se pueden añadir números de página automáticos, insertar ficheros de imagen, formatear el texto, crear tablas, realizar diseños en dos columnas, etc. Para ello, la metodología *Platypus* le permite ir añadiendo elementos del informe a una lista en forma de historia (*story*) (fch. 4.3), para finalmente generar el informe correspondiente, que se sube entonces mediante la librería `biomed-client` (tbl. 4.1) al módulo de ficheros de BIOMED, donde puede ser visualizado por los usuarios (fig. 4.11).

FICHERO 4.3: *Programa de ejemplo para generar un informe a partir de unos datos de entrada.*

```
# Filename: run.py
import argparse; import locale; from datetime import datetime as dt
from reportlab.lib import colors, units
from reportlab.lib.styles import getSampleStyleSheet, ParagraphStyle as PStyle
from reportlab.platypus import SimpleDocTemplate, Spacer, Table, TableStyle
from reportlab.platypus import Paragraph as P
doc = SimpleDocTemplate("sample-report.pdf")
# Parse input arguments
parser = argparse.ArgumentParser(allow_abbrev=False)
parser.add_argument("-n", "--name", type=str, help="Patient name and surnames")
parser.add_argument("-p", "--patient-id", type=str, help="Patient's laboratory ID")
parser.add_argument("-d", "--date", type=str, help="Date of the report")
args = parser.parse_args()
locale.setlocale(locale.LC_ALL, "")  # set language to Spanish
formatted_date = dt.strptime(args.date, "%Y-%m-%d").strftime("%d de %B, %Y")
# Define fixed panel of diseases, colors and styles
diseases = ["Alfa talasemia", "Atrofia muscular espinal", "Fibrosis Quística",
            "Hemoglobinopatía", "Distrofinopatías", "Síndrome X Frágil"]
white = colors.CMYKColor(0, 0, 0, 0)
dblue = colors.CMYKColor(.9, .6, 0, .3)
lblue = colors.CMYKColor(.86, .08, 0, 0)
yellow = colors.CMYKColor(0, 0, 1, 0)
S = getSampleStyleSheet()
S.add(PStyle("title_1", parent=S["Heading1"], fontSize=20, alignment=1, textColor=dblue))
S.add(PStyle(name="white", parent=S["Normal"], fontSize=10, textColor=white))
S.add(PStyle(name="date", parent=S["Normal"], fontSize=12, alignment=1))
S.add(PStyle(name="title_2", parent=S["Heading3"], fontSize=12, textColor=white))
# Define tables to use
def list_to_table_matrix(a_list, n):
    table = [a_list[i:i + n] for i in range(0, len(a_list), n)]
    matrix_table = [list(map(lambda x: P(x, S["Normal"]), row)) for row in table]
    return matrix_table
class DiseasesTable(Table):
    def __init__(self, *args, **kwargs):
        super().__init__(*args, **kwargs)
        table_style = TableStyle()
        table_style.add("VALIGN", (0, 0), (-1, -1), "MIDDLE")
        table_style.add("INNERGRID", (0, 0), (-1, -1), 0.25, colors.blue)
        table_style.add("BOX", (0, 0), (-1, -1), 0.5, colors.black)
        self.setStyle(table_style)
```

```
class ResultsTable(Table):
    def __init__(self, data, label_color=yellow, *args, **kwargs):
        super().__init__(data, *args, **kwargs)
        table_style = TableStyle()
        table_style.add("VALIGN", (0, 0), (-1, -1), "MIDDLE")
        table_style.add("SPAN", (0, 0), (0, -1))
        table_style.add("BACKGROUND", (0, 0), (0, -1), label_color)
        table_style.add("OUTLINE", (0, 0), (0, -1), 1, label_color)
        table_style.add("OUTLINE", (1, 0), (-1, -1), 1, dblue)
        table_style.add("INNERGRID", (1, 0), (-1, -1), 1, dblue)
        table_style.add("BACKGROUND", (1, 0), (-1, -1), dblue)
        table_style.add("TEXTCOLOR", (1, 0), (-1, -1), white)
        self.setStyle(table_style)
# Build all the required elements of the report
table_data_1 = [
    ["", P("Portador/a", S["title_2"])],
    ["", P(f"El test genético ha determinado que <b>{args.patient_id}</b> es "
           "portador/a de la/s siguiente/s enfermedad/es genética/s:",
           S["white"])]]
table_data = list_to_table_matrix(diseases, n=3)
table_diseases = DiseasesTable(table_data, hAlign="LEFT")
table_data_2 = [["", P("No Portador/a", S["title_2"])],
                ["", P(f"El test ha determinado que <b>{args.patient_id}</b> no es "
                       "portador/a de la/s enfermedad/es genética/s analizadas.",
                       S["white"])]]
positive = ResultsTable(table_data_1, colWidths=[5, doc.width - 15])
negative = ResultsTable(table_data_2, colWidths=[5, doc.width - 15], label_color=lblue)
# Create the Platypus story and generate the document
story = []
story.append(P(f"Informe clínico para {args.name}", S["title_1"]))
story.append(Spacer(1, 1*units.cm))
story.append(P(f"Fecha: {formatted_date}", S["date"]))
story.append(Spacer(1, 2*units.cm))
story.append(P("Texto libre en <i>diferentes</i> <b>estilos</b>.", S["Normal"]))
story.append(Spacer(1, 3*units.cm))
story.append(positive)
story.append(Spacer(1, 1*units.cm))
story.append(table_diseases)
story.append(Spacer(0, 3*units.cm))
story.append(negative)
doc.build(story)
```



Informe clínico para Arnau Sellarès Rubio

Fecha: 16 de junio, 2022

Texto libre en *diferentes* **estilos**.

Portador/a

El test genético ha determinado que **ASR1234** es portador/a de la/s siguiente/s enfermedad/es genética/s:

Alfa talasemia	Atrofia muscular espinal	Fibrosis Quística
Hemoglobinopatía	Distrofinopatías	Síndrome X Frágil

No Portador/a

El test ha determinado que **ASR1234** no es portador/a de la/s enfermedad/es genética/s analizadas.

FIGURA 4.11: *Ejemplo de un informe generado con Reportlab a partir del comando* `python run.py -n 'Arnau Sellarès Rubio' -p ASR1234 -d 2022-06-16` *usando el programa en el fch. 4.3.*

4.2.2 Importación de datos

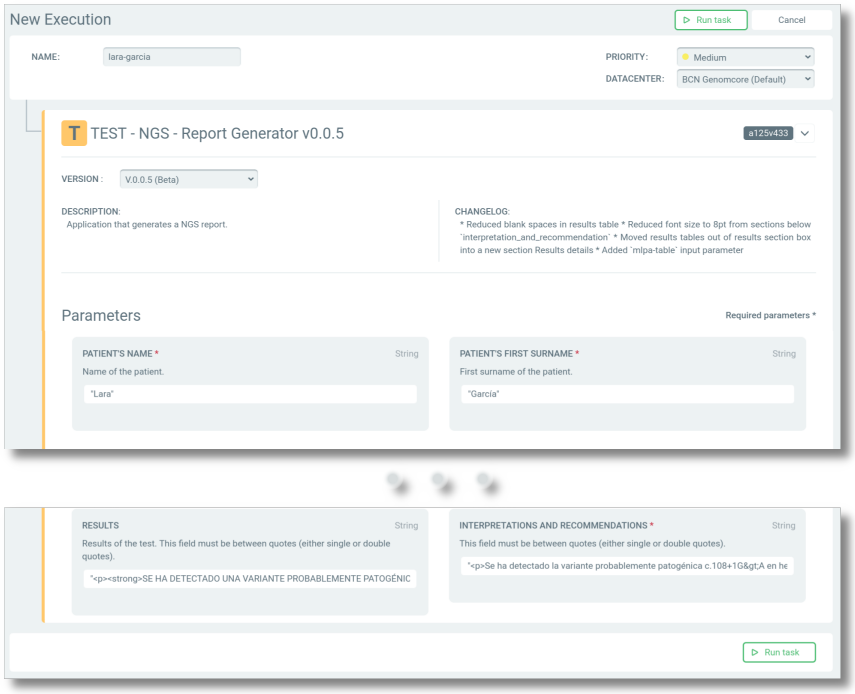
Adicionalmente a los datos NGS obtenidos del *pipeline* bioinformático de la empresa (sec. 4.1.1), la RGT puede generar también informes de resultados a partir de datos bioquímicos, obtenidos mediante analíticas de sangre, y genéticos, obtenidos con la metodología más tradicional de *microarrays* de SNPs. Los datos bioquímicos nos llegan de los laboratorios colaboradores a través de sus LIMS en forma de ficheros HL7 (sec. 3.2.1); por otro lado, los datos genéticos nos llegan en forma de ficheros CSV por correo electrónico (sec. 3.2.2), y nuestro equipo de MoG se encarga de validarlos y subirlos a BIOMED.

4.2.3 Informes de genómica clínica

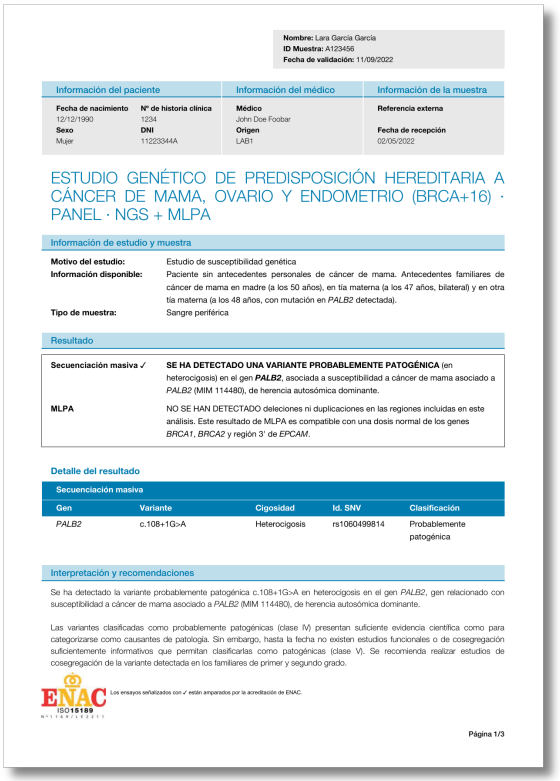
Los informes clínicos generados con la RGT son los que se usan para reportar los resultados del *pipeline* bioinformático de análisis NGS, explicado en detalle al inicio de este capítulo (fig. 4.1); a esta tipología de informes, por su naturaleza, se les llama también **informes de variantes** (fig. 4.12). Una vez el proceso de interpretación y validación de las variantes identificadas en un análisis finaliza (sec. 4.1.7), los mismos analistas inician el procedimiento para generar el informe de variantes (fig. 4.2d), que pasa también por su proceso de validación. Esta validación se realiza en dos fases por diferentes analistas clínicos:

- La primera validación del informe se encarga de rellenar la mayoría de textos del informe a partir de la información almacenada en los Records y las variantes reportadas, generando de esta manera un informe en forma de borrador.
- La segunda validación, llevada a cabo por otro analista que se encarga de revisar el borrador, genera con su visto bueno el informe final de resultados (fig. 4.12b).

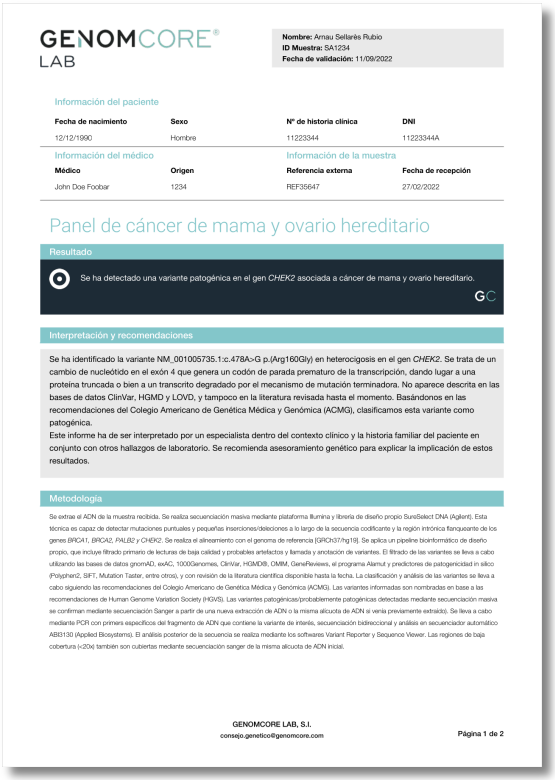
El texto que se incluye en el informe no está redactado manualmente en cada documento, sino que procede de los Records almacenados en la plataforma; el Record



(A) *Aplicación de generación de informes de variantes y datos de entrada; la imagen se ha recortado para mejorar la visualización debido a la extensa cantidad de parámetros de entrada.*



(B) *Informe final de variantes para un laboratorio; los logos del informe se han eliminado por motivos de confidencialidad.*



(C) *Informe de variantes de Genomcore.*

FIGURA 4.12: *Informes de variantes de laboratorio y de Genomcore.*

Perfil (figs. 4.7a, 4.7b), por ejemplo, contiene los textos estáticos que se muestran para un resultado negativo, así como las limitaciones de la técnica, la metodología, y la bibliografía relevante del análisis realizado. Un informe con resultados positivos comparte el texto de limitaciones y metodología, pero su texto sobre los resultados puede proceder o bien del Perfil, si es un resultado positivo común entre diferentes estudios, o del Record Estudio, si es un texto específico solo para ese análisis (fig. 4.3). Además, el informe puede contener resultados externos del laboratorio, obtenidos por otras metodologías diagnósticas como por ejemplo secuenciación Sanger o MLPA, para validar algunas de las variantes reportadas en el informe. Para estos casos, existen también los Records Prueba de tipo Sanger o MLPA (figs. 4.5a, 4.5b), cuyos campos de texto se utilizan en la generación del informe de resultados para incluir textos referentes a las metodologías adicionales. De esta forma se reporta en un solo informe de resultados diferentes técnicas usadas para el análisis, usando los textos combinados de los diferentes tipos de Pruebas.

Finalmente, cuando el informe se encuentra validado y completado, el mismo *workfow NGS Study Report Generation* que ha generado la versión final lo envía a través del LIMS de cada laboratorio (figs. 4.1, 4.2d). El envío del informe se realiza a través de un fichero HL7, que contiene el informe PDF de resultados codificado en `base64` dentro del mismo (Josefsson, 2006); una vez es recibido por el LIMS, el laboratorio al que pertenece se encarga de gestionar ya desde su sistema quien recibe esos informes. Este proceso de transferencia se monitoriza además a través del Record Estudio, que contiene un campo de tiempo de entrega, que es el rango de tiempo entre que se realizó la secuenciación y el envío del informe final de resultados.

El modelo del informe se puede adaptar a los requerimientos de diseño de cada laboratorio que lo solicite; además, disponemos de un modelo genérico para Genomcore (fig. 4.12c), así como de los informes de diagnóstico preconcepcional, que analizan el riesgo en un embarazo de que los padres sean portadores de enfermedades hereditarias a partir de los resultados de un laboratorio colaborador de la empresa

(fig. 4.13). En este último caso, el análisis preconcepcional se realiza con paneles de 6, 16 o 176 genes, y la aplicación es capaz de generar informes en cuatro idiomas distintos (español, inglés, francés e italiano).

FIGURA 4.13: *Informes de diagnóstico preconcepcional para un laboratorio colaborador, cuyos logos y datos privados han sido eliminados por motivos de confidencialidad; tampoco se muestran todas las páginas. Imagen superior: informe con resultados positivos para un panel de 6 genes. Imagen inferior: informe con resultados negativos para un panel de 16 genes.*

4.2.4 Informes de salud personalizada

4.2.4.1 Primera fase de desarrollo

Otra aplicación de la RGT ha sido la generación de informes de salud personalizada, ofrecidos en la empresa bajo la marca *Made of Genes* (MoG) a los usuarios que los adquieren a través de la plataforma web (sec. 1.6). Estos informes integran diferentes fuentes de datos: información genética (cómo eres), datos bioquímicos y analíticas de sangre (cómo estás) e información sobre los hábitos de vida del usuario (cómo te cuidas), para ofrecer recomendaciones de salud personalizada a partir de la evidencia científica actual (fig. 4.15). El proceso de análisis para este tipo de estudios es el siguiente (fig. 4.14):

- La **compra de un estudio MoG** implica:
  - Por una parte, la recepción de un correo electrónico para firmar digitalmente el consentimiento informado (sec. 3.2.3).
  - Por otra, la extracción de muestras del usuario (sec. 4.2.2); esa extracción se lleva a cabo en uno de los centros de la red de laboratorios colaboradores, y puede implicar tanto la extracción de muestras para su análisis bioquímico, como para el genotipado de SNPs, actualmente llevado a cabo mediante *microarrays* genómicos. La programación de la cita con el laboratorio está automatizada una vez el usuario firma el consentimiento informado, que debe rellenar con sus datos personales y llevarlo al laboratorio el día de la extracción.
- La **detección de nuevos documentos firmados** está integrada en BIOMED, de manera que hay un proceso que monitoriza nuevas firmas y actualiza la información correspondiente en la plataforma para su correcta trazabilidad.
- El laboratorio se encarga de **enviar los resultados del análisis** a través de un fichero HL7 cuando se trata de datos bioquímicos, o de un fichero CSV si es

un genotipado; los datos contenidos en ambos se combinan en un solo fichero en formato JSON, que se almacena en BIOMED.

- Este **fichero JSON procesado** se usa como datos de entrada del *workflow* destinado a generar informes de salud personalizada, formado por dos aplicaciones (fig. 4.14):
  - ***MoG Evaluator***: se encarga de validar que todos los datos disponibles son correctos para la generación del informe de salud personalizada. Tiene como dato de entrada el fichero JSON procesado, y genera como datos de salida un nuevo fichero JSON con toda la información estandarizada y necesaria para poder generar el informe.
  - ***MoG Renderer***: esta aplicación se encarga de identificar el tipo de informe a generar a partir del estudio MoG encargado, genera el informe en formato PDF a partir de la información contenida en el JSON estandarizado, el fichero de salida de la primera aplicación, y lo sube a BIOMED (fig. 4.15).

Una vez subido, el informe pasa por un proceso de revisión manual por parte de los ***Health Coach*** antes de ser entregado al usuario que lo encargó; un *Health Coach* es la persona experta en salud dentro de MoG, encargada de interpretar los resultados del estudio y ajustar las recomendaciones de salud al estilo de vida y objetivos del usuario.

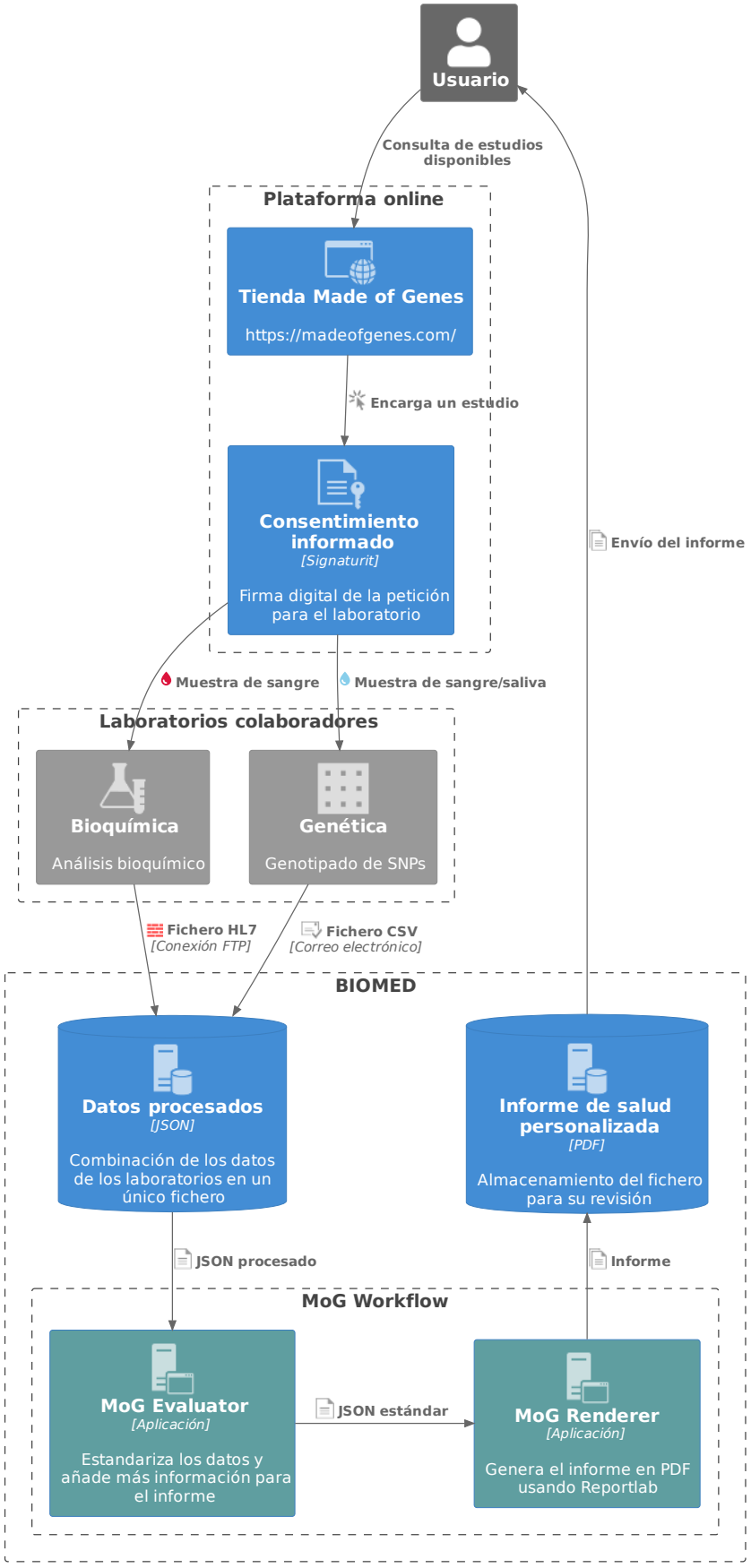
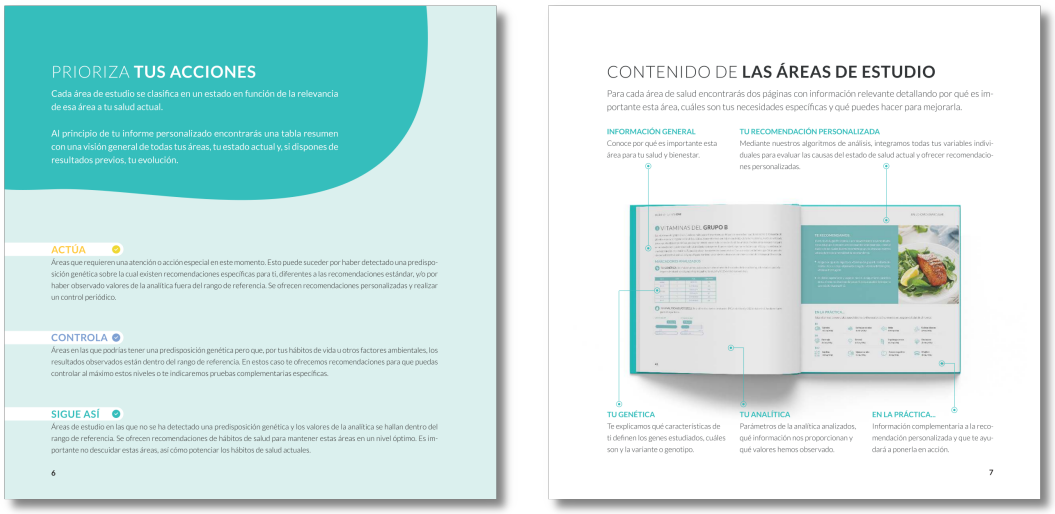
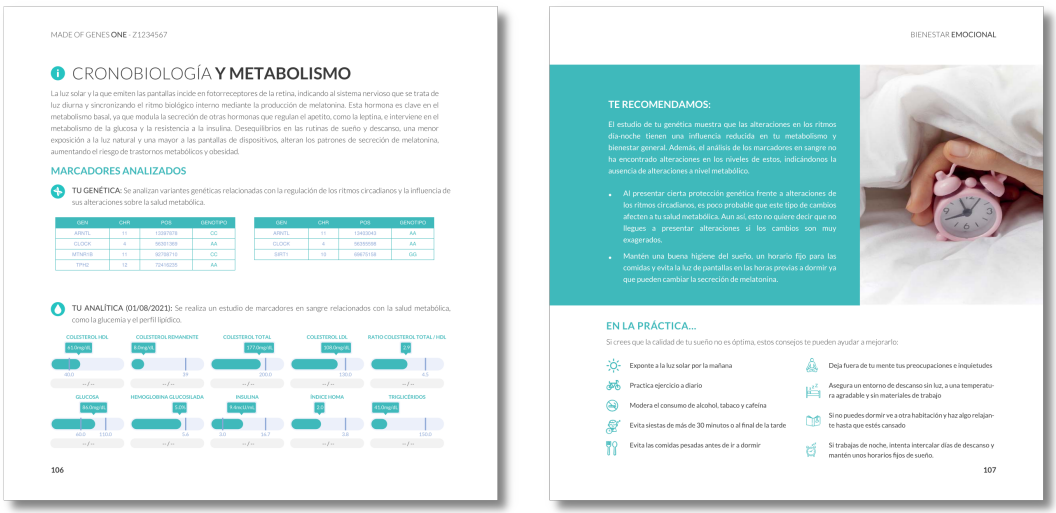


FIGURA 4.14: *Flujo de datos para el análisis de los estudios MoG.*



(A) *Explicación de las acciones y las áreas de estudio.*

(B) *Tabla resumen de las acciones recomendadas para cada área de estudio analizada.*



(C) *Resultados de una área de estudio específica.*

FIGURA 4.15: *Informe Made of Genes ONE de salud personalizada.*

4.2.4.2 Segunda fase de desarrollo

La automatización del procesamiento de los ficheros HL7 y CSV recibidos de los laboratorios se implementó más tarde en una segunda fase; en ésta, una vez se reciben en BIOMED los ficheros correspondientes, se activan una serie de *workflows* automatizados que procesan los datos (fig. 4.16). Hemos desarrollado una API que centraliza los diferentes procesos automatizados de los que disponemos en la plataforma, la `Cron Python API`, de manera que cuando detecta la subida de nuevos ficheros HL7 o CSV en proyectos determinados (organizados por laboratorios), se encarga de ejecutar un *workflow* para procesarlos. Este *workflow* concreto recibe el nombre de **MoGIA**, e incluye, como en el caso de los análisis NGS, una transformación de los datos externos de entrada a Records en BIOMED (fig. 4.17): cada resultado de bioquímica procesado del fichero HL7 se transforma en un Record de tipo *Biovalue* (sec. 4.2.5.1), y cada SNP individual del fichero CSV en un Record de tipo Variante (sec. 4.1.6). Los Records adicionales que se generan durante el proceso (Paciente, Orden, Estudio, etc.) se utilizan como en el caso de los análisis NGS para asegurar la trazabilidad de todos los datos del estudio. Esta organización en *Biovalues* y Variantes nos permite posteriormente agilizar el proceso de evaluación del estudio y hacerlo de una manera mucho más modular.

El *workflow* de MoGIA consta de 3 aplicaciones (fig. 4.16):

- ***Importer***: esta aplicación se encarga de importar los ficheros HL7 y CSV, localizados en los proyectos de cada laboratorio, a un proyecto particular de MoG, que es el que usamos para todos los análisis de salud personalizada. Esta división de proyectos nos permite procesar datos de distintos laboratorios sin que éstos puedan acceder a datos que no sean de su propio laboratorio.
- ***Parser***: su función principal es la de transformar los datos contenidos en los ficheros de entrada a sus correspondientes Records en BIOMED. En el caso de un HL7, la aplicación genera Records de tipo *Biovalue*; en el caso de un CSV,

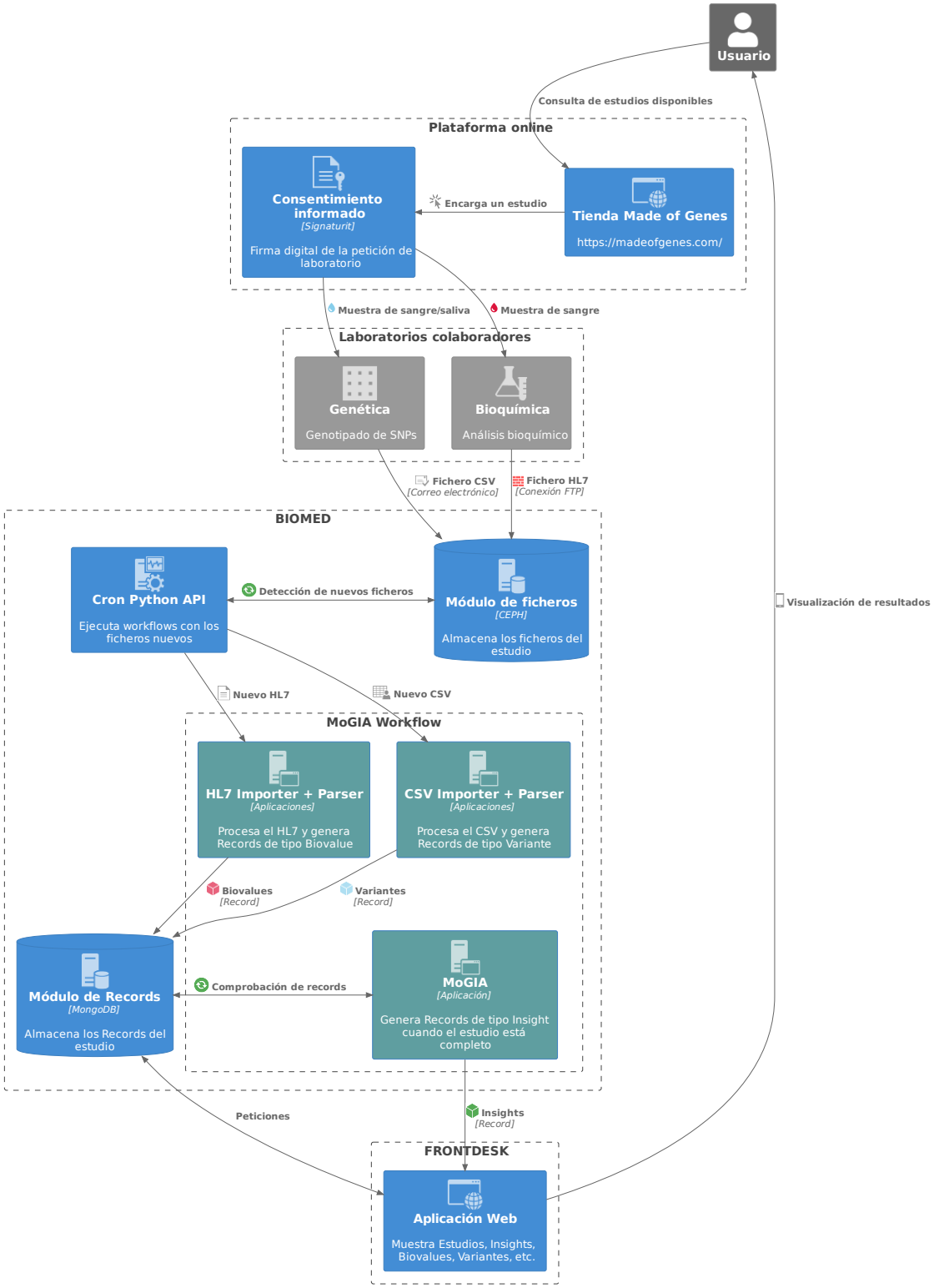


FIGURA 4.16: *Segunda fase del flujo de datos para el análisis de los estudios MoG.*

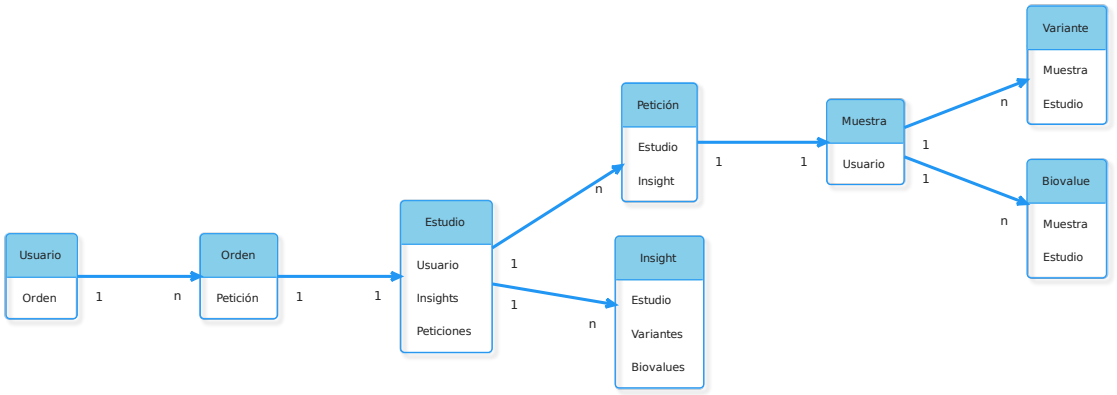


FIGURA 4.17: *Relaciones entre los Records generados para los estudios MoG; debajo de cada Record, se esquematizan los enlaces a otros Records para su correcta trazabilidad.*

genera Records de tipo Variante. En ambos casos, se genera también un Record de tipo Muestra por motivos de trazabilidad.

- ***MoGIA***: se encarga de generar Records de tipo *Insight* a partir de los *Biovalues* y Variantes almacenados para un estudio concreto.

Los *Insights* son los Records que más importancia tienen en esta segunda fase de automatización y migración desde un sistema que genera informes estáticos en PDF a otro que muestra los resultados de un estudio MoG de manera más dinámica (ver sec. 4.2.5.2 más adelante); esta nueva organización de la información ha permitido el **desarrollo de una aplicación web** orientada a los usuarios de los estudios MoG, donde la visualización de los resultados se realiza de forma más flexible (sec. 4.2.5).

4.2.5 APLICACIÓN WEB

La implementación de los *Insights* como entidades en los que reflejar los datos y las recomendaciones de salud para cada área de un estudio MoG se ha integrado al desarrollo de una aplicación web orientada al usuario que solicita un estudio (sec. 3.2.5). Su funcionamiento es el siguiente: cuando un usuario encarga un estudio MoG, además de recibir por correo electrónico el consentimiento informado, recibe ahora otro correo para activar su usuario en la aplicación web. Una vez activado, el acceso a la

aplicación le permite ir siguiendo el estado de su estudio, desde la extracción de muestras en el laboratorio y su análisis en BIOMED, hasta la generación de los Records de tipo *Insight*, que tras ser revisados por los *Health Coach*, enviarán al usuario una notificación automática conforme su estudio se encuentra validado y finalizado.

4.2.5.1 Biovalues

El módulo de *Biovalues* es un componente de BIOMED que permite almacenar los datos de un análisis bioquímico de laboratorio de manera individualizada y flexible. Un *Biovalue* es un tipo de plantilla de Records (sec. 3.1.7) especializada, que además dispone de una API propia para su integración con otros componentes de la infraestructura de la empresa. Los Records de tipo *Biovalue* almacenan información bioquímica, como por ejemplo, el valor de la analítica, el rango normal, la fecha en que se analizó, la tarea de BIOMED con la que se generaron, etc. (fig. 4.18).

4.2.5.2 Insights

Para la segunda fase de la RGT, se ha diseñado también el módulo de *Insights*. Un ***Insight*** es una entidad individual en BIOMED, un tipo de Record específico que se utiliza para guardar la información sobre recomendaciones personalizadas de salud basándose en los *Biovalues* y variantes disponibles para un usuario, enmarcadas dentro de áreas de salud, por ejemplo: niveles de vitamina D, presión arterial, o efectividad de las dietas bajas en grasas (fig. 4.19). Este módulo es muy similar al de *Biovalues* (sec. 4.2.5.1), ya que también requieren de una plantilla de Records especializada y una API propia.

4.2.5.3 Implementación

En esta aplicación web, la presentación de los *Insights*, *Biovalues* y otros datos de salud se realiza de manera mucho más visual que en BIOMED (fig. 4.20): el usuario puede consultar fácilmente mediante un menú lateral todos los datos asociados a su usuario

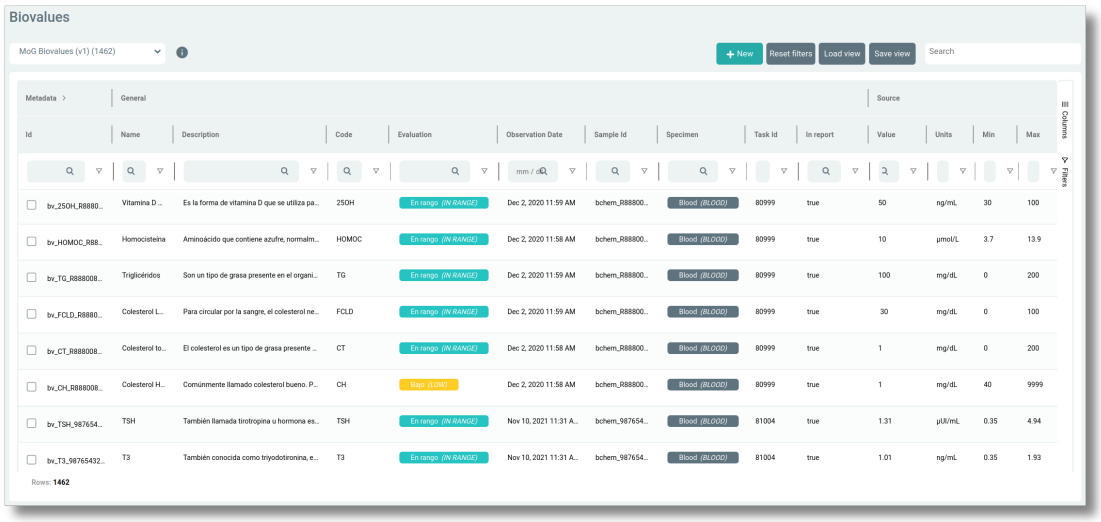


FIGURA 4.18: *Módulo de Biovalues de BIOMED.*

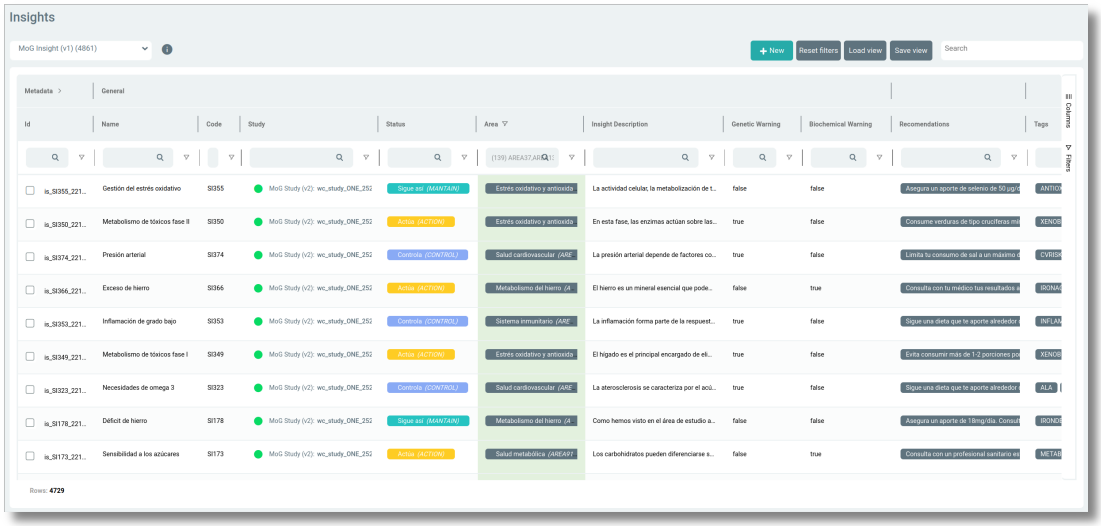
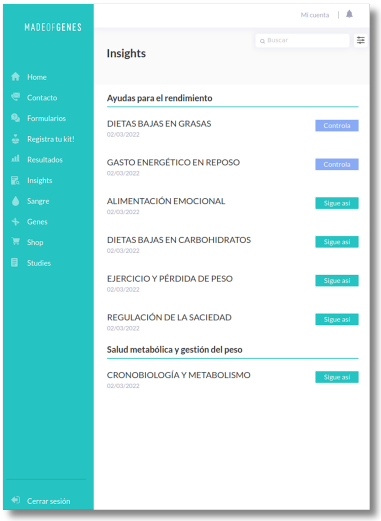


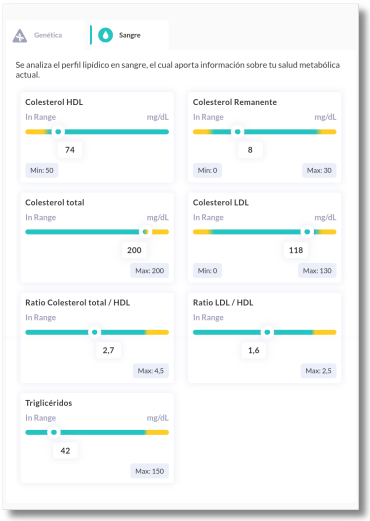
FIGURA 4.19: *Módulo de Insights de BIOMED.*

(fig. 4.20a); al navegar a un *Insight* específico obtendrá información más detallada de éste, además del listado de variantes genéticas y marcadores bioquímicos asociados a él (figs. 4.20b, 4.20c). La aplicación también permite visualizar los valores históricos de elementos que se han reanalizado en estudios posteriores (fig. 4.20d), así como todos los estudios realizados y los informes asociados (fig. 4.20e). El asesoramiento de los resultados se realiza igualmente por parte de los *Health Coach* de MoG, en este caso ya no con un informe estático en PDF, sino con una aplicación dinámica, fácilmente consultable desde el móvil o un navegador web cualquiera.

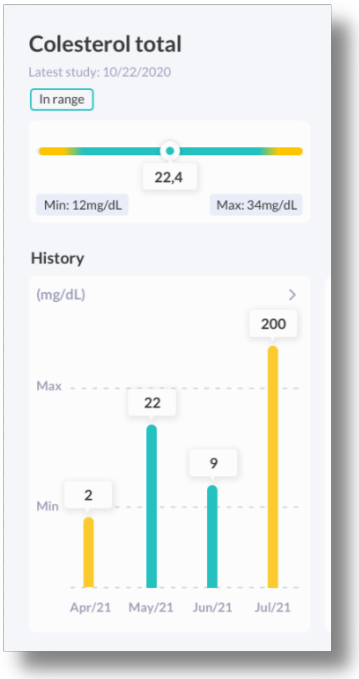


(A) *Listado de Insights, junto con las recomendaciones sugeridas para cada uno. En el menú lateral se pueden observar los distintos módulos de visualización de datos del usuario.*

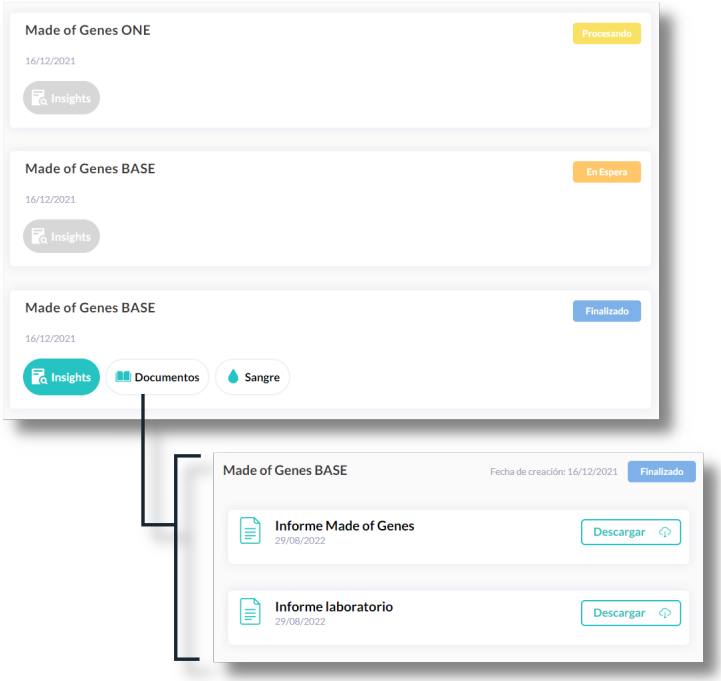
(B) *Detalle de un Insight, sección de genética.*



(C) *Detalle de un Insight, sección de bioquímica.*



(D) *Visualización del histórico para un Biovalue.*



(E) *Visualización de Estudios, su estado y los informes asociados.*

FIGURA 4.20: *Visualización de resultados e informes en la aplicación web de MoG.*

CAPÍTULO 5

Discusión

En este capítulo se evalúan las herramientas desarrolladas en esta tesis doctoral, y se discute el objetivo planteado al inicio de la tesis de validar en un entorno clínico el VPMS y la RGT. Finalmente, se discuten las problemáticas encontradas y las mejoras que se podrían aplicar en un futuro.

5.1 Sistema de gestión de paneles virtuales de genes

El VPMS desarrollado en este proyecto permite incrementar la automatización de los análisis NGS llevados a cabo en la empresa, además de brindar un sistema de fácil acceso y ejecución para el reanálisis de datos con diferentes preguntas clínicas. Esto permite realizar estudios dirigidos para evitar la detección de VUS, UFs, y SFs problemáticos, ya que un número demasiado elevado de variantes identificadas durante un análisis puede dificultar y retrasar el proceso de interpretación y validación de los resultados clínicamente relevantes que se reportan al paciente. En la fig. 5.1 se puede observar como la cantidad de variantes incrementa al ampliar las ROIs, pasando de identificar solo cuatro variantes en los exones del gen `BRCA2` para el panel, a identificar

más variantes en los exones de BRCA2 y ZAR1L en la WES, a identificar en la WGS muchísimas más variantes tanto en intrones como exones de ambos genes, además de en las regiones intergénicas.

FIGURA 5.1: *Amplitud de tres análisis NGS diferentes, visualizando en el IGV los ficheros VCF generados por el pipeline bioinformático, y aumentado a una región del cromosoma 13 humano. En los carriles superiores se pueden observar, respectivamente, las variantes identificadas en un análisis para un panel de genes (4 en esta región, 4 en el cromosoma 13, y 21 en total), para WES (13 en esta región, 2,358 en el cromosoma 13, y 140,570 en total) y para WGS (144 en esta región, 179,065 en el cromosoma 13, y 5,011,327 en total); la combinación del azul y rojo indica si la variante en esa posición es homocigótica o heterocigótica. Los cuatro carriles inferiores determinan el nombre estándar de los genes en esa región, además de las ROIs utilizadas en los tres tipos de análisis, definidas en el fichero BED del panel virtual.*

Adicionalmente, el VPMS resulta coste-eficiente para las pruebas de WES y WGS realizadas, debido a que el continuo reanálisis de los datos posibilita aumentar el rendimiento diagnóstico de la prueba, facilitando así la implementación de la genómica en el campo del diagnóstico genético, y reduciendo potencialmente el tiempo de respuesta respecto al enfoque tradicional basado en metodología Sanger.

5.1.1 REANÁLISIS

El reanálisis de datos y/o la publicación de nueva evidencia científica puede llevar a una reclasificación de las variantes previamente identificadas, lo que puede suponer una actualización de su efecto biológico y su patogenicidad en el paciente o pacientes en los que se identificaron. En otras ocasiones, la actualización de las versiones de las

herramientas utilizadas en el *pipeline* bioinformático de la empresa puede suponer una mejora de su sensibilidad y especificidad para detectar variantes, dando fruto a nuevos diagnósticos a partir del reanálisis de los datos de secuenciación generados en estudios anteriores; en este caso, el reanálisis se puede producir en diferentes puntos del proceso bioinformático, dependiendo del motivo por el que se realiza (sec. 1.4.3.2).

La implementación actual del reanálisis resulta trivial a ojos del analista clínico, que vuelve a ejecutar con los mismos datos NGS el *workflow* de identificación de variantes sin la aplicación inicial de alineamiento con el genoma de referencia (fig. 4.10). Esta modularidad permite reagrupar las aplicaciones en *workflows* más reducidos, ignorando la ejecución de procesos o aplicaciones que no sean necesarios para el reanálisis, reduciendo así el tiempo de ejecución. No obstante, actualmente esta flexibilidad en el reanálisis no está todavía integrada en procesos automatizados de la plataforma, por lo que cualquier nueva evidencia publicada o nueva versión de *software* debe ser manualmente revisada para decidir qué pacientes o estudios tendrían que ser reanalizados; por el momento, esta decisión recae en los analistas que realizan sus estudios en la plataforma. En consecuencia, una continuidad de este presente proyecto sería el desarrollo de procesos automatizados que identifiquen las variantes afectadas por publicaciones de nueva evidencia científica, y notifiquen a los actores involucrados la posibilidad de realizar un reanálisis, con la finalidad de optimizar el diagnóstico clínico a partir de los datos de secuenciación generados en estudios anteriores; la actualización de los *pipelines* bioinformáticos de la empresa también serían motivo de procesos de reanálisis de la base de datos de variantes existente en la plataforma, ya que el refinamiento y mejora de los algoritmos de alineamiento genómico, identificación, priorización y clasificación de variantes de cada programa utilizado podría incrementar la especificidad y sensibilidad del análisis por si solo. No obstante, actualmente aún es necesaria una definición e implementación concreta de los protocolos y recomendaciones a seguir por parte de los laboratorios y las instituciones que realizan los análisis bioinformáticos, ya que sus implicaciones morales y éticas son relevantes tanto para

el paciente afectado como para sus familiares directos (Richards *et al.*, 2015; Rehder *et al.*, 2021).

5.1.1.1 Beneficios y limitaciones del reanálisis

La reevaluación de una variante a causa de nuevos descubrimientos y/o pruebas puede conducir a la reclasificación de esa variante. Para los pacientes que esperan un diagnóstico molecular, una VUS que ha sido reclasificada como patogénica o probablemente patogénica puede resultar transformadora y proporcionar beneficios tales como un pronóstico informado, gestión adecuada de su salud e información sobre el riesgo de recurrencia. Sin embargo, la reevaluación de una variante también puede conducir a una degradación de su clasificación, lo que da como resultado que la variante que se categorizó en el pasado como patogénica sea reclasificada como VUS, probablemente benigna, o benigna (Xiang *et al.*, 2020); la degradación en la clasificación de una variante puede ser particularmente devastador para los pacientes a los que previamente se les proporcionó un diagnóstico molecular, ya que tomaron decisiones médicas importantes basándose en variantes patogénicas o probablemente patogénicas previas. Este potencial de reclasificación para alterar un diagnóstico molecular establecido tiene ramificaciones significativas para los sistemas de salud, ya que demuestra que incluso los pacientes que han recibido un diagnóstico molecular pueden beneficiarse de un reanálisis rutinario, no solo aquellos que esperan un diagnóstico. No obstante, el reanálisis rutinario de todos los conjuntos de datos genómicos recopilados anteriormente aumentará exponencialmente las cargas de trabajo (Robertson *et al.*, 2022).

El reanálisis y la reevaluación de los datos genéticos puede implicar costes significativos, ya que los laboratorios pueden llegar a dedicar **entre 20 y 40 horas de trabajo** realizado por expertos para producir un informe clínico inicial (Wenger *et al.*, 2017); múltiples grupos han expresado su preocupación por la carga de trabajo asociada con el reanálisis y han destacado la necesidad de un sistema para compensar el coste económico para los sistemas sanitarios (Machini *et al.*, 2019). Aunque el reanálisis pueda

suponer un medio para garantizar que a los pacientes no se les receten tratamientos ineficaces, sin un cambio significativo en el modelo de reembolso existente, el reanálisis rutinario será insostenible en los sistemas de salud pública actuales. No obstante, existen estudios que describen el desarrollo de sistemas parcialmente automatizados que reducen esta carga de trabajo en más de un 90% (Baker *et al.*, 2019), aunque esta automatización parcial indica que la mano de obra experta sigue siendo esencial para el reanálisis y requiere de una compensación adecuada (Robertson *et al.*, 2022).

La generación de un informe de genómica clínica normalmente requiere del procesamiento de grandes volúmenes de datos; para que el reanálisis se convierta en una práctica rutinaria para todos los pacientes con datos genómicos disponibles, se deben implementar sistemas para el almacenamiento, procesamiento y uso clínico de estos datos. Cada etapa del proceso de reanálisis tiene diferentes requisitos y desafíos para los datos tratados; por ejemplo, la reidentificación de variantes requiere del almacenamiento y procesamiento de hasta cientos de *gigabytes* de datos genómicos sin procesar (en bruto) por paciente. Algunos estudios han planteado la problemática de que el tamaño y la complejidad de estos datos pueden hacer que el reanálisis universal sea prohibitivamente costoso, que estos ficheros deberían eliminarse directamente y que la resecuenciación debería realizarse con futuras tecnologías más económicas (Costain *et al.*, 2018). En contra de este argumento está el hecho de que el coste de la secuenciación solo ha disminuido aproximadamente un 50% desde mediados de 2015 (Wetterstrand, 2022) y sigue siendo significativamente más caro que el coste estimado de almacenamiento (Krumm y Hoffman, 2020); sin embargo, siguen quedando dudas sobre en quién recae la responsabilidad del coste de almacenamiento y recuperación de los datos y cómo se sufragarán dichos costes.

Por el contrario, es posible que los procesos de reanotación y repriorización de variantes solo requieran del archivo VCF del análisis original. El tamaño comparativamente pequeño de este archivo permite superar muchos de los obstáculos en el almacenamiento de ficheros de secuenciación; sin embargo, no permite aprovechar

ninguno de los beneficios derivados de la realineación y la reidentificación. Aunque los ficheros VCF pueden ser más manejables, este formato fue diseñado para entornos de investigación (Danecek *et al.*, 2011); en consecuencia, existen limitaciones acerca de cómo se adaptará la variabilidad asociada a los ficheros VCF dentro de los sistemas de salud modernos e interoperables. Actualmente se están desarrollando recursos para cerrar esta brecha (Dolin *et al.*, 2021); sin embargo, para que el reanálisis de los datos genómicos se convierta en algo rutinario, debe realizarse una síntesis de la bioinformática con la informática de la salud.

Actualmente, no existen dudas acerca del potencial que el reanálisis periódico de los datos genómicos existentes tiene para transformar la atención médica de los pacientes; los diagnósticos adicionales logrados pueden alterar la trayectoria clínica de un paciente y conducirle a mejores tratamientos, y mejor gestión y asesoramiento. Algunos de los desafíos para la implementación de reanálisis clínicos rutinarios se superarán a través de políticas concretas, mientras que otros se resolverán mediante la mejora de los sistemas de genómica clínica y los avances en salud digital. Lo más importante es seguir trabajando para ayudar a todos los pacientes que siguen sin un diagnóstico genético concluyente, a pesar de las mejores prácticas en las pruebas genéticas actuales.

5.1.2 Anotación de variantes

El proceso de anotación (sec. 1.3.1.3) constituye un paso muy importante del *pipeline* bioinformático que puede someterse a una mayor estandarización para aumentar la consistencia de los análisis NGS entre instituciones y laboratorios. Uno de los principales desafíos para la implementación generalizada de las capacidades predictivas de la NGS es la falta de bases de datos de anotaciones genómicas completas y consistentes que sean de acceso público. Actualmente, existen diferentes bases de datos: la *consensus coding sequence* (CCDS) (Pruitt *et al.*, 2009), RefSeq (Pruitt, Tatusova y

Maglott, 2007), *Known Genes* (Hsu *et al.*, 2006), GENCODE (Harrow *et al.*, 2006), ENSEMBL (Hubbard *et al.*, 2009), HGMD (Stenson *et al.*, 2008), y OMIM (Amberger *et al.*, 2015), entre las más importantes. Cada una de ellas tiene sus ventajas e inconvenientes, por ejemplo, errores en la curación de los datos, o la existencia de datos desactualizados; además, el uso inconsistente de estas bases de datos en la literatura dificulta la replicación de los descubrimientos de investigación (Dewey *et al.*, 2012).

Como este proceso se basa en la recopilación de información almacenada en bases de datos estáticas en constante evolución, es necesario implementar políticas para realizar un versionado de estos ficheros descargados, así como actualizaciones periódicas para descargar versiones más nuevas. En este sentido, además de varias aplicaciones comerciales diseñadas para facilitar la anotación e interpretación de los datos de secuenciación (Oliver, Hart y Klee, 2015), y el uso de herramientas tradicionales para encontrar similitudes en otras especies (Stein, 2001), ANNOVAR para integrar datos entre diferentes categorías de anotaciones (Wang, Li y Hakonarson, 2010), o el *Ensembl Variant Effect Predictor* (VEP) (McLaren *et al.*, 2016), se han desarrollado diversas estrategias automatizadas de anotación de genes que codifican para proteínas (Brent, 2008), así como de extracción de datos fenotípicos a partir de notas clínicas o EHRs; investigaciones preliminares sugieren que utilizar algoritmos de procesamiento de lenguaje natural y otros métodos de automatización incrementan significativamente la tasa diagnóstica y la eficiencia del análisis a la hora de priorizar variantes reportables en análisis genómicos, aunque su sensibilidad y especificidad siguen siendo insuficientes para sustituir el papel de un analista humano (Thuriot *et al.*, 2018; Clark *et al.*, 2019; Austin-Tse *et al.*, 2022). Con el fin de optimizar este proceso, el uso de herramientas que capturen información detallada en un formato estándar y estructurado es vital a la hora de escalar y automatizar el análisis y reporte en genómica clínica.

Otra de las limitaciones del proceso de anotación es el de conseguir anotar variantes no codificantes para proteína, ya que históricamente los estudios genómicos

han ignorado más del 99% de la variación genómica humana; en este sentido, están apareciendo herramientas que integran capacidades predictivas con iniciativas como el proyecto *Encyclopedia of DNA elements* (ENCODE) para analizar más allá de las regiones codificantes (Dunham *et al.*, 2012; Ritchie *et al.*, 2014).

5.1.3 Priorización de variantes basada en términos fenotípicos

En las pruebas tradicionales enfocadas a una enfermedad, el número de variantes identificadas suele ser lo suficientemente pequeño como para permitir la evaluación individual de todas las variantes en cada muestra; sin embargo, en la WES se identifican decenas de miles de variantes y en WGS hasta varios millones, lo que hace imposible esta aproximación. Aplicar el potencial de los datos predictivos genómicos obtenidos de la NGS es una de las tareas más difíciles en la interpretación de datos. Dado que se pueden identificar potencialmente millones de variantes a partir de un solo análisis, el proceso de priorización de variantes es importante para reducir la lista de variantes candidatas a interpretación. Estos pasos deben valorar el equilibrio entre maximizar la sensibilidad y minimizar la cantidad de VUS, reduciendo las variantes identificadas para la revisión de los analistas, y garantizando que no se excluyan variantes patogénicas durante el proceso.

En los análisis basados en el fenotipo para el filtraje y priorización de variantes, los laboratorios pueden beneficiarse enormemente estableciendo políticas para automatizar la selección de genes de interés para un fenotipo específico, utilizando plataformas que se basan en datos estructurados de fenotipos de pacientes para priorizar variantes encontradas en esos genes (Firth *et al.*, 2009; Wang *et al.*, 2019), con controles de calidad adicionales para garantizar que las listas de genes seleccionadas sean curadas y revisadas adecuadamente por expertos en la materia. Además, es recomendable que dichas listas sean verificadas y actualizadas periódicamente; el ACMG recomienda que las listas de genes de los paneles sean examinadas cada 6 meses para determinar

si los datos más recientes sugieren la adición o eliminación de genes de dichas listas (Rehder *et al.*, 2021).

En este sentido, otro aspecto mejorable para el VPMS vendría dado por el diseño de estrategias automatizadas de creación de paneles virtuales mediante la selección de genes candidatos a partir de términos fenotípicos estándar definidos por la HPO (Saklatvala, Dand y Simpson, 2018; Wang *et al.*, 2019). La *BED Tool*, la aplicación diseñada durante este proyecto para la generación de paneles virtuales (sec. 4.1.4), permite mucha flexibilidad a la hora de definir las regiones para un panel virtual específico; como la inclusión de ROIs en el fichero BED de salida depende de la introducción en la lista de entrada de un transcrito concreto, la aplicación es capaz de diseñar paneles virtuales con tantas combinaciones como transcritos humanos existan, resultando así muy personalizable a las necesidades determinadas de cada análisis NGS que se desee realizar. Por otro lado, es importante que la información del EHR del paciente sea lo más precisa posible, para que tal sistema sea capaz de recuperarla y realizar una correcta y ajustada selección de genes candidatos para un análisis dirigido. La automatización incremental del VPMS facilitará de esta manera un incremento del rendimiento diagnóstico a partir de la mejora de la información fenotípica estructurada a la que se pueda acceder desde BIOMED.

5.1.4 Interpretación de variantes

Actualmente, la interpretación de un análisis NGS representa un desafío para analistas y laboratorios que desean beneficiarse del potencial diagnóstico de la genómica clínica. Este reto implica que, por ejemplo, en ClinVar haya un 17% de variantes interpretadas por más de un analista con discrepancias en la clasificación de la variante (Rehm *et al.*, 2015), o que por ejemplo a fecha de 2014 hubiera discrepancias entre la base de datos de transcritos RefSeq y el genoma de referencia GRCh37, que afectaban a 5308 transcritos en un total de 3039 genes (Oliver, Hart y Klee, 2015). Para facilitar

esta labor, se están desarrollando herramientas automatizadas de interpretación de variantes que reduzcan la brecha entre los resultados reportados por el laboratorio y su significado clínico (Chunn *et al.*, 2020; Chin *et al.*, 2022).

Para optimizar el proceso de interpretación de variantes, un sistema ideal se dedicaría a presentar al analista toda la evidencia pertinente, literatura científica e información sobre la variante específica y los genes afectados, reduciendo así la necesidad de consultar recursos externos, permitiéndole resumir rápidamente todos los datos y considerar la relevancia potencial con respecto a la indicación clínica principal del paciente para la prueba. El sistema podría ayudar a facilitar este proceso al permitir que los analistas introduzcan notas y comentarios para ayudar en la revisión final del caso; mantener toda esta información en una base de datos interna de variantes mejoraría así la optimización del análisis cuando se detecten las mismas variantes en futuros análisis realizados en otros individuos. El módulo de variantes desarrollado durante este proyecto en BIOMED facilita esta labor interpretativa para el analista (sec. 4.1.6), ya que permite la introducción de comentarios personalizados para una variante analizada específica, además de enlaces externos para consultas adicionales (fig. 4.9b).

En un futuro, disponer de bases de datos con datos genómicos de grado clínico y de calidad será crucial para avanzar en nuestra comprensión de la importancia clínica de la variación genética y para una interpretación más exacta y reproducible de los resultados NGS (sec. 1.3.3); los esfuerzos actuales a nivel nacional e internacional para crear bases de datos genómicas exhaustivas incluyen ClinGen (Rehm *et al.*, 2015) y el *Human Variome Project* o proyecto del varioma humano (HVP) (Burn y Watson, 2016).

5.1.5 Importación automatizada de Records

Actualmente, la creación de todos los Records de BIOMED necesarios para el correcto funcionamiento del *pipeline* bioinformático requiere aún de la intervención manual del equipo de la empresa para crear, por ejemplo, nuevos Records de Librerías o Perfiles. Aunque la mayoría de esos Records no cambian muy a menudo y su información es utilizada por múltiples análisis, el VPMS podría mejorarse también desarrollando nuevas automatizaciones que faciliten esa importación a partir de los catálogos de pruebas de los laboratorios. Sin embargo, la falta de estandarización y estructuración de los datos en cada institución dificulta la implementación de dichas mejoras, por lo que es preciso mejorar la interoperabilidad entre sistemas, por ejemplo, fomentando la gestión de la información en bases de datos estándar, y estableciendo procesos automatizados de consulta e intercambio de información entre las bases de datos de diferentes instituciones. En este sentido, queda mucho camino por recorrer, debido a que dicha automatización depende del grado de colaboración entre proveedores de salud.

5.1.6 Diferencias entre Variantes y Records

El almacenamiento de variantes generadas en análisis NGS en la plataforma ha pasado por diferentes fases de desarrollo debido a la magnitud de los datos gestionados. Inicialmente, se utilizó una base de datos de `ElasticSearch` para almacenar directamente las variantes desde el *workflow* de detección de variantes, para luego recuperarlas desde BIOMED mediante una API especializada para ello. Actualmente, las variantes se almacenan en una base de datos de MongoDB, como sucede con el módulo de Records, siendo la única diferencia entre ambos que, en el caso de las variantes, no existe una plantilla JSON que sirva de intermediaria de validación de campos. Al ser la nomenclatura de las variantes mucho más estandarizada y estática que la diseñada para los Records, no se prevé una reducción en el rendimiento de este módulo, que

valida directamente las variantes almacenadas en el momento de creación mediante su API especializada.

5.1.7 Limitaciones de CWL

El uso de CWL como lenguaje estándar para diseñar los diferentes *workflows* que componen el *pipeline* bioinformático de la empresa ha sufrido también una evolución desde su implementación inicial. Aunque se trate de un lenguaje común para diseñar *workflows* de procesamiento de datos NGS (Strozzi *et al.*, 2019; Korhonen *et al.*, 2019), nuestra implementación inicial basada en diseñar una única aplicación que incluía todas las etapas del análisis bioinformático resultó problemática al procesar datos de WES y WGS, por una limitación del propio CWL de no eliminar ficheros intermedios entre diferentes pasos del *workflow* (https://github.com/common-workflow-language/cwltool/issues/892). Este hecho provocó problemas en la asignación de recursos computacionales en los nodos de la empresa, porque un solo *workflow* consumía mucho espacio de disco, al estar almacenando de manera incremental todos los ficheros intermedios del procesamiento. La implementación actual del *pipeline* bioinformático de la empresa tiene una aproximación más modular debido al desarrollo de aplicaciones que se pueden conectar entre ellas mediante los *binders* (sec. 3.1.3.2), y la asignación de recursos para cada ejecución resulta mucho más estable y escalable a procesos con datos genómicos de intensa computación, ya que SLURM es capaz de gestionar mejor los recursos disponibles en los diferentes nodos de computación cuanto más modulares sean las aplicaciones esperando a ejecutarse.

5.2 Herramienta de generación de informes

La RGT es una herramienta que, actualmente, permite generar en el entorno de producción de BIOMED informes de una amplia variedad de tipologías, desde informes clínicos de variantes hasta informes de salud personalizada orientados a la prevención. Se trata además de un sistema escalable debido a la automatización de la mayoría de los procesos y *workflows* que la componen. Sin embargo, el sistema se puede seguir mejorando; una de las propuestas para realizar en etapas futuras es la de centralizar su funcionalidad en una sola API que, a partir de unos datos biomédicos de entrada, sea capaz de generar informes clínicos o de salud personalizada en formato PDF y/o *Biovalues*, Variantes e *Insights* dependiendo de cada implementación particular, de manera que el proceso sea más estandarizado y su monitorización y ejecución no se vea dividida entre varias aplicaciones y *workflows*. Adicionalmente, se quiere mejorar también: el proceso de validación de los datos que nos llegan de los laboratorios (sec. 4.2.2), debido a que no siempre cumplen con el formato estándar HL7 y eso genera problemas en su procesamiento, al generar errores de ejecución por campos inválidos por su tipología; y el proceso de estandarización de éstos, ya que el uso de identificadores internos por parte de los laboratorios provoca también que puedan existir colisiones con los identificadores de los marcadores bioquímicos de otros laboratorios. En este sentido, la integración en la RGT de la nomenclatura *Logical Observation Identifier Names and Codes* (LOINC) eliminaría estos problemas, por lo que se considera otra de las potenciales mejoras a implementar en el futuro inmediato (sec. 5.4.4).

5.2.1 DocuSign

El proceso de firma digital encargado actualmente a Signaturit (sec. 3.2.3) cumple con su funcionalidad de facilitar la firma del consentimiento informado para el análisis de laboratorio cómodamente de manera remota. Sin embargo, su API no dispone de una funcionalidad para crear y editar formularios de consentimientos informados de

manera programática, trabajo que recae así al equipo de Genomcore, que actualmente se realiza de manera manual, tanto para subir el documento PDF que se va a usar como plantilla para el consentimiento informado, como para crear los campos rellenables encima del PDF para generar el formulario.

En este sentido, estamos investigando una migración del sistema de firma digital a DocuSign (https://www.docusign.com/), que sí que dispone de una API completa para la generación y edición de formularios (https://developers.docusign.com/docs/esign-rest-api/esign101/concepts/templates/). Esto nos permitirá desarrollar una aplicación de BIOMED que permita generar formularios de consentimientos informados de manera automatizada para todos los laboratorios colaboradores que lo soliciten.

5.2.2 Reporte de resultados

El uso de Reportlab como base para la creación de informes de resultados en PDF ha sido acertada, ya que cubre todos los casos de uso encontrados hasta el momento, y permite una gran flexibilidad para implementar nuevas características. Paulatinamente a su desarrollo, la aplicación inicial se ha ido dividiendo en aplicaciones de funcionalidad más reducida para aumentar la modularidad de la RGT. Actualmente, existen diferentes *workflows* en los que la aplicación final es la encargada de generar el documento PDF a partir de unos datos de entrada estandarizados (secs. 4.2.3, 4.2.4.1); en uno de ellos, la funcionalidad está migrando hacia un sistema de representación de los resultados en una aplicación web, que permite la visualización de resultados de una manera cómoda desde cualquier dispositivo móvil o de sobremesa con acceso a un navegador (sec. 4.2.4.2). La consulta de datos de salud en esta aproximación resulta más accesible que la que se almacena en un informe PDF impreso. Hasta el momento, la recepción por parte de los usuarios de la aplicación web ha sido positiva, y se espera seguir monitorizando su adopción para analizar las preferencias de visualización del usuario o paciente de un estudio.

Actualmente, personalizar la inclusión o exclusión de resultados en los informes para cada paciente es un reto que puede superarse mediante la automatización de procesos que permitan a los médicos especificar la pregunta clínica principal durante el paso inicial de solicitud del análisis. Otros aspectos del informe de variantes están bajo desarrollo, como por ejemplo la introducción de la relevancia clínica de las variantes en la nomenclatura LOINC (https://loinc.org/53037-8/) y en el estándar HL7 (http://www.hl7.org/fhir/uv/genomics-reporting/history.html).

5.2.3 Herramientas alternativas

En la fase inicial de diseño y desarrollo de la RGT, se consideraron diferentes herramientas alternativas a Reportlab para la generación automatizada de informes de resultados. Una de ellas fue `pdftk` (https://www.pdflabs.com/tools/pdftk-the-pdf-toolkit/), que permite rellenar la información de un formulario PDF de manera programática. El motivo de su descarte fue sobre todo por la imposibilidad de usarlo para generar toda la estructura del informe desde cero mediante código, como sí es posible con Reportlab, pero también porque ocasionaba errores de visualización con letras que contenían caracteres especiales como tildes.

Otra de las herramientas consideradas fue *Jupyter Notebooks* (https://jupyter.org/), una potente herramienta de visualización dinámica de resultados en unos ficheros especializados para ello, los *notebooks*; sin embargo, resulta complicada de utilizar para personas que no tienen un perfil técnico enfocado a la programación, hecho que la descartó también para nuestro caso de uso, en la que queríamos implementar una RGT que fuera sencilla y clara de utilizar.

5.3 Validación clínica del VPMS y la RGT

El objetivo planteado inicialmente de validar en un entorno clínico el VPMS y la RGT se diseñó pensando en dos fases de desarrollo:

- En la primera fase, la intención era diseñar una prueba piloto de un producto mínimo viable en la que se usaran datos de variantes genéticas previamente validadas por los colaboradores, para integrar las herramientas en su rutina clínica y observar si los datos generados eran consistentes con la validación, y evaluar la experiencia de los analistas clínicos respecto a la mejora de la visualización y reporte de resultados.
- En la segunda fase, se deseaba probar el mínimo producto viable validado en un conjunto de muestras nuevas sin variantes validadas previamente, para analizar el impacto del VPMS y la RGT en el rendimiento diagnóstico de las pruebas genéticas ejecutadas.

Desafortunadamente, el desarrollo de esta prueba piloto, en colaboración con el servicio de medicina genética y genómica de un hospital para integrar ambas herramientas en su rutina clínica, no se ha podido llevar a cabo; la aparición de la pandemia provocada por la COVID-19 a principios de 2020 obligó a los sistemas de salud nacionales y autonómicos a dar máxima prioridad a la realización de pruebas diagnósticas para esta enfermedad durante meses, hecho que ha imposibilitado la formalización de un convenio colaborativo debido, en parte, a la dificultad logística para conseguir el consentimiento informado de pacientes con patologías clínicas.

No obstante, los productos desarrollados se encuentran funcionando actualmente en producción en la plataforma de la empresa, en la cual los analistas externos validan clínicamente los resultados NGS obtenidos a partir del VPMS y la RGT (figs. 5.2, 5.3). Con la paulatina descongestión de los sistemas de salud, se espera poder presentar más adelante una prueba piloto de integración de los productos en servicios hospitalarios,

para que evalúen y validen su funcionalidad y utilidad respecto a la optimización de recursos bioinformáticos e interpretación y reporte de variantes clínicamente relevantes. Esta imposibilidad de realizar uno de los tres objetivos planteados inicialmente ha provocado una replanificación del trabajo del doctorando, centrándose más en el diseño de una segunda fase de desarrollo para la RGT y los *workflows* implicados para implementar un sistema que permita almacenar los datos de una manera más dinámica en una aplicación web, en forma de Records en vez de informes estáticos en formato PDF (sec. 4.2.4.2).

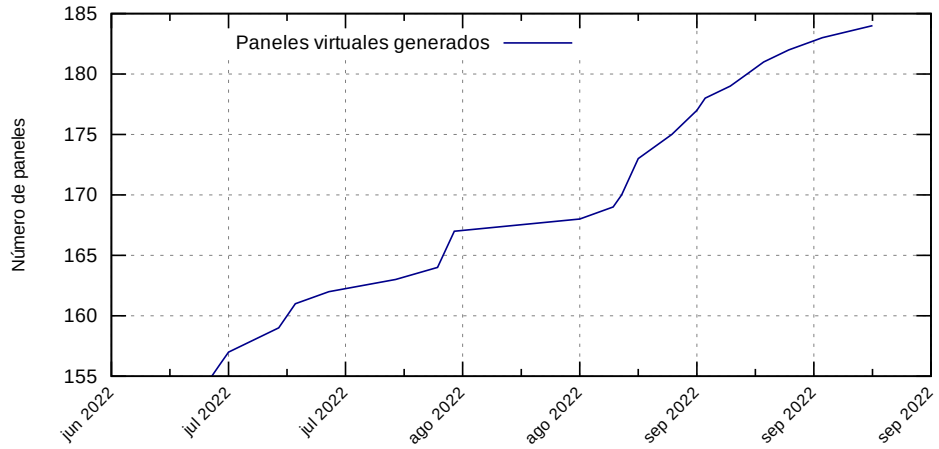


FIGURA 5.2: *Paneles virtuales generados con la herramienta BED Tool del VPMS para un laboratorio colaborador desde julio a septiembre de 2022, representados por la fecha de creación del Record Prueba.*

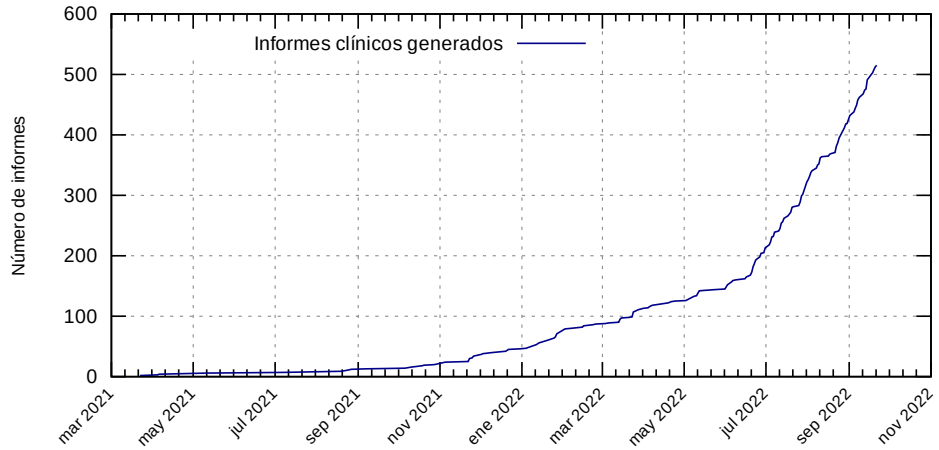


FIGURA 5.3: *Informes clínicos de resultados generados con la RGT en el pipeline bioinformático de análisis NGS de la empresa desde marzo de 2021 hasta septiembre de 2022, para un laboratorio colaborador, representados por la fecha de creación del informe validado.*

5.4 Implementación de las herramientas

5.4.1 Flujo de desarrollo

Los *workflows* y aplicaciones desarrollados en este proyecto disponen actualmente de herramientas para realizar su despliegue en BIOMED de manera automatizada. Este sistema de *Continuous Integration and Continuous Delivery* o integración y entrega continuas (CI/CD) ha ido evolucionando con el tiempo; inicialmente, cada aplicación disponía de su repositorio de código en Bitbucket (sec. 3.1.4), y las nuevas versiones desarrolladas se desplegaban directamente en el entorno de producción para ser comprobadas por nuestro equipo antes de ser visibles para el usuario final. Este despliegue se realizaba de manera manual con un *script* programado en Bash, que construía la imagen Docker correspondiente y la subía al AWS ECR privado de la empresa, lista para ser descargada y ejecutada en los nodos de computación. Sin embargo, no se trataba de un proceso óptimo, ya que el despliegue de la aplicación dependía de las condiciones del sistema operativo en el ordenador del desarrollador que ejecutaba el *script*, en el que las versiones del *software* instalado (por ejemplo, de Docker) podía variar, dando lugar a imágenes diferentes dependiendo de quién realizara el despliegue. La implementación actual permite que el sistema de desarrollo de nuevas aplicaciones y versiones sea mucho más estable y consistente, utilizando entornos homogéneos para generar las imágenes Docker, a la vez que permite recuperar versiones anteriores en caso de errores inesperados en el despliegue. Adicionalmente, el CI/CD implementado permite una mejor trazabilidad de las versiones desplegadas con respecto a los estándares internacionales `ISO`.

5.4.2 Categorización de versiones

La categorización de versiones en *Deprecated*, *Beta* y *Latest* (sec. 3.1.3.1) ha mejorado también el sistema de despliegue y fases de pruebas para las versiones pendientes de

validación y en desarrollo. La implementación anterior, que no disponía de dichas categorías, no permitía un sistema tan flexible para que los usuarios pudieran probar las nuevas versiones desarrolladas; de la misma manera, no había un criterio establecido de qué versiones anteriores seguían siendo funcionales, de forma que el usuario no tenía una forma sencilla de saber si podía seguir utilizando una versión antigua para reanalizar tareas específicas, o si ésta había dejado de funcionar. Actualmente, la categoría *Beta* permite al usuario de la plataforma probar la nueva versión, evaluar si cumple con sus requisitos y, solo entonces, solicitar modificación a la categoría *Latest* para determinar la nueva versión estable de la aplicación correspondiente. De la misma manera, la categoría *Deprecated* indica al usuario que esa versión sigue funcionando, pero que en un futuro es posible que pierda su funcionalidad porque ya no está actualizada e integrada a los cambios continuos que van sufriendo las APIs de BIOMED, momento en el cual el administrador de su organización puede desactivarla desde el panel de control. Esta solución es satisfactoria para todos los actores involucrados en el desarrollo, despliegue y ejecución de aplicaciones.

5.4.3 Base de datos de Records

Inicialmente, la base de datos implementada para el módulo de Records (sec. 3.1.7) se basó en PostgreSQL, una base de datos relacional. Sin embargo, la continua monitorización de sus recursos y la integración con los distintos *workflows* de análisis de la empresa demostraron que una base de datos relacional no era óptima para trabajar con Records, a causa del esfuerzo computacional adicional que supone el hecho de mapear los objetos y campos de cada Record a las tablas de una base de datos, definidas en su esquema relacional; en esas condiciones, las peticiones contra la API de Records se volvían lentas, y en algunos casos devolvían errores al intentar filtrar por valores de campos específicos de la plantilla, si existían demasiados Records en la base de datos. Por este motivo se migró todo el sistema a MongoDB, una base de

datos no relacional, que para nuestro caso de uso resulta en una implementación más flexible, y que además permite realizar consultas muy específicas sobre parámetros concretos dentro de un Record, una funcionalidad muy útil para el VPMS. Tener la base de datos de Records en MongoDB permite actualmente disponer de un sistema que sí es escalable a las necesidades de la plataforma y que de momento ha soportado correctamente los centenares de peticiones por segundo que recibe desde las distintas aplicaciones y *workflows* ejecutándose en BIOMED.

El diseño y desarrollo de una batería de plantillas de Records que cubran todos los casos de uso para los análisis realizados en la plataforma ha sido también crucial a la hora de implementar flujos de trabajo que puedan hacer uso de todo el potencial de la plataforma en la gestión y obtención de datos de entrada, y el almacenamiento de datos de salida, para cualquier tipo de análisis soportado actualmente. El uso de plantillas refuerza la validación de campos en el momento de crear un Record nuevo, tanto por parte de las aplicaciones como también por parte del usuario de la plataforma, que los puede editar desde el módulo de Records. Además, la inclusión de parámetros nuevos en las plantillas no implica la migración de los Records generados hasta ese momento, ya que, al ser una base de datos no relacional, la estructura de cada Record existente sigue siendo válida, y lo único que se actualiza es su interacción con la API para ver si sigue cumpliendo con los requisitos de la plantilla; sin embargo, la modificación de la tipología de un campo existente en la plantilla sí que implica tener que recurrir a operaciones en la base de datos para asegurar una consistencia entre los Records antiguos y los nuevos para, por ejemplo, evitar que un mismo campo pueda ser numérico en un Record antiguo y de texto en Records nuevos. Disponer de plantillas de Records en formato JSON también permite su versionado en la plataforma y la gestión de cambios a través de `git` y Bitbucket, de manera que una plantilla puede ir evolucionando con el tiempo acorde a las nuevas especificaciones diseñadas sin perder información contextual de desarrollos anteriores.

5.4.4 Estandarización e intercambio de datos

El proceso de análisis y reanálisis de datos NGS lleva consigo una serie de promesas sobre el futuro que depara al campo de la genómica clínica; un sistema en el que la información generada por diferentes actores e instituciones sea compartida a nivel global facilitaría mucho el trabajo de interpretación de las variantes detectadas, especialmente de aquellas con un origen en enfermedades raras, ya que mejoraría mucho el tiempo de respuesta clínico y, por ende, implicaría también una mejora en la salud del paciente o pacientes involucrados, al reducir la odisea diagnóstica (sec. 1.4.1).

Para que este escenario sea posible, la información biomédica que circula entre sistemas (federados o centralizados) debe ser estandarizada a una nomenclatura y estructura que sea interoperable e interpretable por cualquier institución que quiera hacer uso de ella. Actualmente, este sistema universal no existe, aunque como se ha observado en la introducción, sí que existen diversos esfuerzos colectivos y públicos avanzando con ese objetivo (sec. 1.5.3). El uso de ficheros que permitan el intercambio de datos clínicos de manera estándar, como es el caso del formato HL7, facilita mucho esta colaboración; es un formato aún en constante evolución, implementado ya actualmente en entornos clínicos. No obstante, estos ficheros llevan consigo datos de salud en los que la nomenclatura y escala utilizada es aún específica de cada laboratorio, es decir, que no usan identificadores universales únicos; este hecho puede generar problemas cuando se desea realizar una integración de datos de diferentes laboratorios en una única plataforma, como ha sido el caso del presente proyecto.

La adopción de la nomenclatura LOINC reduciría estos obstáculos de integración de los datos obtenidos de diferentes laboratorios (Huff *et al.*, 1998; Bodenreider, Cornet y Vreeman, 2018); en este sentido, se está desarrollando en la empresa una API destinada a la gestión del sistema LOINC, que normalice los valores analíticos mediante un código estandarizado que permita la integración e interoperabilidad de los datos de diferentes proveedores. Esta API se ha diseñado con el motor `Flask`, imple-

mentado en `Python` (https://palletsprojects.com/p/flask/), para desplegar un servicio interno que permita a los *workflows* de BIOMED realizar peticiones para conseguir códigos LOINC a partir de los códigos internos de cada laboratorio colaborador, y realizar posteriormente la conversión de valores biomédicos para su estandarización; todo esto se almacenará en la base de datos de Records (por ejemplo, en la de *Biovalues*) para centralizar la información. Para que sea posible, los laboratorios colaboradores deberán reportar el código interno para cada valor de bioquímica analizado en el campo correspondiente del fichero HL7; actualmente, esto no siempre es así, hecho que dificulta su implementación.

Un sistema de intercambio de datos estandarizados mejoraría mucho las perspectivas de diagnóstico genético, ya que promovería el desarrollo de una extensa red de conocimientos globalizada y accesible. Herramientas como ClinVar o *PanelApp* ya van dirigidas con esa finalidad (secs. 1.3.1.3, 1.4.4); las bases de datos clínicas, especialmente las públicas, deben recibir e incorporar los nuevos descubrimientos reportados de manera constante, y los laboratorios deben tener políticas que promuevan el intercambio de datos con estas plataformas. El intercambio de datos genotípicos y fenotípicos fomentará así la recopilación de una masa crítica de pruebas clínicas que respalden las supuestas variantes de una enfermedad.

CAPÍTULO 6

Conclusiones

A continuación se detallan las conclusiones del presente proyecto de tesis doctoral:

1. El análisis genómico conlleva el procesamiento bioinformático de grandes volúmenes de datos que necesitan estar debidamente estructurados para poder implementar procesos automatizados que faciliten su interpretación posterior por parte del profesional clínico.
2. La integración de los datos involucrados en todo el proceso, desde la extracción de muestras en el laboratorio hasta la interpretación de los datos y el informe de resultados, es vital para agilizar la entrega de resultados a médicos y pacientes y poner así fin a la odisea diagnóstica.
3. La interoperabilidad de los datos generados es importante para optimizar la transferencia y el procesamiento de datos entre instituciones.
4. El reanálisis de datos genómicos mediante el uso de paneles virtuales de genes permite a los laboratorios e instituciones involucradas ser más coste-eficientes, además de incrementar la tasa diagnóstica de las pruebas realizadas.
5. El VPMS y la RGT facilitan la automatización y la implementación de los análisis genómicos en la rutina clínica, y su traducción en un informe de resultados relevantes y accionables para la salud del paciente.

6. El VPMS y la RGT facilitan el reanálisis de datos en la plataforma BIOMED a muy bajo coste y de manera segura.
7. La RGT facilita la generación automatizada de resultados accionables a partir de datos ómicos, como los obtenidos del genoma y el metaboloma de una persona.
8. El uso de tecnologías móviles permite la visualización dinámica de los resultados en un formato mucho más accesible que el formato PDF, empoderando al usuario propietario de los datos.

Bibliografía

Aganezov, S. *et al.* (2022) «A complete reference genome improves analysis of human genetic variation», *Science*, 376(6588), p. eabl3533. doi:10.1126/science.abl3533.

Aho, A.V., Kernighan, B.W. y Weinberger, P.J. (1979) «Awk — a pattern scanning and processing language», 9, pp. 267-279. doi:10.1002/spe.4380090403.

Altman, D.G. y Bland, J.M. (1994) «Statistics Notes: Diagnostic tests 1: sensitivity and specificity», *BMJ*, 308(6943), p. 1552. doi:10.1136/bmj.308.6943.1552.

Amberger, J.S. *et al.* (2015) «OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders», *Nucleic Acids Research*, 43(D1), pp. D789-D798. doi:10.1093/nar/gku1205.

Amendola, L.M. *et al.* (2015) «Actionable exomic incidental findings in 6503 participants: challenges of variant classification», *Genome Research*, 25(3), pp. 305-315. doi:10.1101/gr.183483.114.

Amstutz, P. *et al.* (2016) «Common Workflow Language v1.0». Common Workflow Language working group. doi:10.6084/m9.figshare.3115156.v2.

Andermann, A. *et al.* (2008) «Revisiting Wilson and Jungner in the genomic age: a review of screening criteria over the past 40 years», *Bulletin of the World Health Organization*, 86(4), pp. 317-319. doi:10.2471/blt.07.050112.

Arteche-López, A. *et al.* (2021) «Sanger sequencing is no longer always necessary based on a single-center validation of 1109 NGS variants in 825 clinical exomes.», *Scientific reports*, 11, p. 5697. doi:10.1038/s41598-021-85182-w.

Ashley, E.A. *et al.* (2010) «Clinical assessment incorporating a personal genome», *The Lancet*, 375(9725), pp. 1525-1535. doi:10.1016/S0140-6736(10)60452-7.

Austin-Tse, C.A. *et al.* (2022) «Best practices for the interpretation and reporting of clinical whole genome sequencing», *npj Genomic Medicine*, 7(1), pp. 1-13. doi:10.1038/s41525-022-00295-z.

Avent, N.D. y Reid, M.E. (2000) «The Rh blood group system: a review», *Blood*, 95(2), pp. 375-387. doi:10.1182/blood.V95.2.375.

Baker, S.W. *et al.* (2019) «Automated Clinical Exome Reanalysis Reveals Novel Diagnoses», *The Journal of molecular diagnostics: JMD*, 21(1), pp. 38-48. doi:10.1016/j.jmoldx.2018.07.008.

Bamford, S. *et al.* (2004) «The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website», *British Journal of Cancer*, 91(2), pp. 355-358. doi:10.1038/sj.bjc.6601894.

Bauman, J.G.J. *et al.* (1980) «A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA», *Experimental Cell Research*, 128(2), pp. 485-490. doi:10.1016/0014-

4827(80)90087-7.

Bean, L. *et al.* (2020) «Diagnostic gene sequencing panels: from design to report - a technical standard of the American College of Medical Genetics and Genomics (ACMG)», *Genetics in Medicine*, 22(3), pp. 453-461. doi:10.1038/s41436-019-0666-z.

Beck, T.F., Mullikin, J.C. y NISC Comparative Sequencing Program (2016) «Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants», *Clinical Chemistry*, 62(4), pp. 647-654. doi:10.1373/clinchem.2015.249623.

Berger, B., Peng, J. y Singh, M. (2013) «Computational solutions for omics data», *Nature Reviews Genetics*, 14(5), pp. 333-346. doi:10.1038/nrg3433.

Birney, E., Vamathevan, J. y Goodhand, P. (2017) *Genomics in healthcare: GA4GH looks to 2022*. bioRxiv, p. 203554. doi:10.1101/203554.

Bodenreider, O., Cornet, R. y Vreeman, D.J. (2018) «Recent Developments in Clinical Terminologies — SNOMED CT, LOINC, and RxNorm», *Yearbook of Medical Informatics*, 27(01), pp. 129-139. doi:10.1055/s-0038-1667077.

Bowdin, S. *et al.* (2016) «Recommendations for the integration of genomics into clinical practice.», *Genetics in medicine : official journal of the American College of Medical Genetics*, 18, pp. 1075-1084. doi:10.1038/gim.2016.17.

Boycott, K. *et al.* (2015) «The clinical application of genome-wide sequencing for monogenic diseases in Canada: Position Statement of the Canadian College of Medical Geneticists.», *Journal of medical genetics*, 52, pp. 431-437. doi:10.1136/jmedgenet-2015-103144.

Brent, M.R. (2008) «Steady progress and recent breakthroughs in the accuracy of automated genome annotation», *Nature Reviews Genetics*, 9(1), pp. 62-73. doi:10.1038/nrg2220.

Burke, W. (2002) «Genetic Testing», *New England Journal of Medicine*, 347(23), pp. 1867-1875. doi:10.1056/NEJMoa012113.

Burn, J. y Watson, M. (2016) «The Human Variome Project», *Human Mutation*, 37(6), pp. 505-507. doi:10.1002/humu.22986.

Carey, D.J. *et al.* (2016) «The Geisinger MyCode community health initiative: an electronic health record–linked biobank for precision medicine research», *Genetics in Medicine*, 18(9), pp. 906-913. doi:10.1038/gim.2015.187.

Carter, C.O. *et al.* (1960) «Chromosome translocation as a cause of familial mongolism», *The Lancet*, 276(7152), pp. 678-680. doi:10.1016/S0140-6736(60)91749-9.

Chacon, S. y Straub, B. (2014) *Pro git*. Apress.

Chandler, N. *et al.* (2018) «Rapid prenatal diagnosis using targeted exome sequencing: a cohort study to assess feasibility and potential impact on prenatal counseling and pregnancy management», *Genetics in Medicine*, 20(11), pp. 1430-1437. doi:10.1038/gim.2018.30.

Check Hayden, E. (2009) «Genome sequencing: the third generation», *Nature* [Preprint]. doi:10.1038/news.2009.86.

Chen, R. *et al.* (2012) «Personal omics profiling reveals dynamic molecular and medical phenotypes.», *Cell*, 148, pp. 1293-1307. doi:10.1016/j.cell.2012.02.009.

Cheon, J.Y., Mozersky, J. y Cook-Deegan, R. (2014) «Variants of uncertain significan-

ce in BRCA: a harbinger of ethical and policy issues to come?», *Genome Medicine*, 6(12), p. 121. doi:10.1186/s13073-014-0121-3.

Chin, H.-L. *et al.* (2022) «The Clinical Variant Analysis Tool: Analyzing the evidence supporting reported genomic variation in clinical practice», *Genetics in Medicine*, 0(0). doi:10.1016/j.gim.2022.03.013.

Choi, M. *et al.* (2009) «Genetic diagnosis by whole exome capture and massively parallel DNA sequencing», *Proceedings of the National Academy of Sciences*, 106(45), pp. 19096-19101. doi:10.1073/pnas.0910672106.

Chunn, L.M. *et al.* (2020) «Mastermind: A Comprehensive Genomic Association Search Engine for Empirical Evidence Curation and Genetic Variant Interpretation», *Frontiers in Genetics*, 11. Disponible en: https://www.frontiersin.org/articles/10.3389/fgene.2020.577152 (Accedido: 25 de agosto de 2022).

Church, D.M. *et al.* (2011) «Modernizing Reference Genome Assemblies», *PLOS Biology*, 9(7), p. e1001091. doi:10.1371/journal.pbio.1001091.

Clark, M.M. *et al.* (2018) «Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases», *npj Genomic Medicine*, 3(1), pp. 1-10. doi:10.1038/s41525-018-0053-8.

Clark, M.M. *et al.* (2019) «Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation», *Science Translational Medicine*, 11. doi:10.1126/scitranslmed.aat6177.

ClinGen (2022) «The Low-Penetrance/Risk Allele Working Group». Disponible en: https://www.clinicalgenome.org/working-groups/low-penetrance-risk-allele-working-group/.

Costain, G. *et al.* (2018) «Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing», *European Journal of Human Genetics*, 26(5), pp. 740-744. doi:10.1038/s41431-018-0114-6.

Crick, F. (1979) «Split Genes and RNA Splicing», *Science*, 204(4390), pp. 264-271. doi:10.1126/science.373120.

Crick, F.H.C. *et al.* (1961) «General Nature of the Genetic Code for Proteins», *Nature*, 192(4809), pp. 1227-1232. doi:10.1038/1921227a0.

Cummings, B.B. *et al.* (2017) «Improving genetic diagnosis in Mendelian disease with transcriptome sequencing.», *Science translational medicine*, 9. doi:10.1126/scitranslmed.aal5209.

Danecek, P. *et al.* (2011) «The variant call format and VCFtools», *Bioinformatics*, 27(15), pp. 2156-2158. doi:10.1093/bioinformatics/btr330.

Deignan, J.L. *et al.* (2019) «Points to consider in the reevaluation and reanalysis of genomic test results: a statement of the American College of Medical Genetics and Genomics (ACMG)», *Genetics in Medicine*, 21(6), pp. 1267-1270. doi:10.1038/s41436-019-0478-1.

Den Dunnen, J.T. *et al.* (2016) «HGVS Recommendations for the Description of Sequence Variants: 2016 Update», *Human Mutation*, 37(6), pp. 564-569. doi:10.1002/humu.22981.

DePristo, M.A. *et al.* (2011) «A framework for variation discovery and genotyping using next-generation DNA sequencing data», *Nature Genetics*, 43(5), pp. 491-498.

doi:10.1038/ng.806.
Dewey, F.E. *et al.* (2012) «DNA Sequencing», *Circulation*, 125(7), pp. 931-944. doi:10.1161/CIRCULATIONAHA.110.972828.
Dimmock, D. *et al.* (2021) «Project Baby Bear: Rapid precision care incorporating rWGS in 5 California children's hospitals demonstrates improved clinical outcomes and reduced costs of care», *The American Journal of Human Genetics*, 108(7), pp. 1231-1238. doi:10.1016/j.ajhg.2021.05.008.
Dobzhansky, T. (1937) *Genetics and the origin of species*. New York: Columbia University Press (11).
Dolin, R.H. *et al.* (2021) «vcf2fhir: a utility to convert VCF files into HL7 FHIR format for genomics-EHR integration», *BMC Bioinformatics*, 22(1), p. 104. doi:10.1186/s12859-021-04039-1.
Driessen, V. (2010) «A successful Git branching model». Disponible en: http://nvie.com/posts/a-successful-git-branching-model/ (Accedido: 19 de julio de 2022).
Dunham, I. *et al.* (2012) «An integrated encyclopedia of DNA elements in the human genome», *Nature*, 489(7414), pp. 57-74. doi:10.1038/nature11247.
El, C.G. van *et al.* (2013) «Whole-genome sequencing in health care: recommendations of the European Society of Human Genetics.», *European journal of human genetics : EJHG*, 21, pp. 580-584. doi:10.1038/ejhg.2013.46.
Farnaes, L. *et al.* (2018) «Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization», *npj Genomic Medicine*, 3(1), pp. 1-8. doi:10.1038/s41525-018-0049-4.
FDA (2021) «Overview of IVD Regulation». Disponible en: https://www.fda.gov/medical-devices/ivd-regulatory-assistance/overview-ivd-regulation (Accedido: 23 de junio de 2022).
Ferguson-Smith, M.A. (2015) «History and evolution of cytogenetics», *Molecular Cytogenetics*, 8(1), p. 19. doi:10.1186/s13039-015-0125-8.
Feuk, L., Carson, A.R. y Scherer, S.W. (2006) «Structural variation in the human genome», *Nature Reviews Genetics*, 7(2), pp. 85-97. doi:10.1038/nrg1767.
Fiegler, H. *et al.* (2003) «Array painting: a method for the rapid analysis of aberrant chromosomes using DNA microarrays», *Journal of Medical Genetics*, 40(9), pp. 664-670. doi:10.1136/jmg.40.9.664.
Firth, H.V. *et al.* (2009) «DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources», *The American Journal of Human Genetics*, 84(4), pp. 524-533. doi:10.1016/j.ajhg.2009.03.010.
Ford, C.E. *et al.* (1959) «A sex chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome)», *The Lancet*, 273(7075), pp. 711-713. doi:10.1016/S0140-6736(59)91893-8.
Freed, A.S. *et al.* (2020) «The Impact of Rapid Exome Sequencing on Medical Management of Critically Ill Children», *The Journal of Pediatrics*, 226, pp. 202-212.e1. doi:10.1016/j.jpeds.2020.06.020.
Gayon, J. (2016) «From Mendel to epigenetics: History of genetics», *Comptes Rendus Biologies*, 339(7), pp. 225-230. doi:10.1016/j.crvi.2016.05.009.
Genome Reference Consortium (2022) «Human Genome Overview». Disponible en: https://www.ncbi.nlm.nih.gov/grc/human (Accedido: 29 de mayo de 2022).

Genomics England (2022) «What are additional findings?», *Genomics England*. Disponible en: https://www.genomicsengland.co.uk/initiatives/100000-genomes-project/additional-findings (Accedido: 11 de junio de 2022).

Ginsburg, G.S. (2019) «A Global Collaborative to Advance Genomic Medicine», *The American Journal of Human Genetics*, 104(3), pp. 407-409. doi:10.1016/j.ajhg.2019.02.010.

Glenn, T.C. (2011) «Field guide to next-generation DNA sequencers», *Molecular Ecology Resources*, 11(5), pp. 759-769. doi:10.1111/j.1755-0998.2011.03024.x.

Global Alliance for Genomics and Health (2016) «GENOMICS. A federated ecosystem for sharing genomic, clinical data.», *Science*, 352, pp. 1278-1280. doi:10.1126/science.aaf6162.

Global Alliance for Genomics and Health (2018) «Beacon API v1». Disponible en: https://www.ga4gh.org/genomic-data-toolkit/beacon-api/ (Accedido: 29 de mayo de 2022).

Global Health Observatory (2018a) «Barriers to implementing electronic health records - Reported data by country», *WHO*. World Health Organization. Disponible en: https://apps.who.int/gho/data/node.main.GOE0502 (Accedido: 18 de febrero de 2022).

Global Health Observatory (2018b) «Individual rights - Reported data by country», *WHO*. World Health Organization. Disponible en: https://apps.who.int/gho/data/node.main.GOE0602 (Accedido: 18 de febrero de 2022).

Global Health Observatory (2018c) «Policy or legislation - Reported data by country», *WHO*. World Health Organization. Disponible en: https://apps.who.int/gho/data/node.main.GOE0601 (Accedido: 18 de febrero de 2022).

Global Health Observatory (2019) «National plan and legislation - Reported data by country», *WHO*. World Health Organization. Disponible en: https://apps.who.int/gho/data/node.main.GOE0501 (Accedido: 18 de febrero de 2022).

Goldstein, S. (1971) «Somatic cell genetics.», *Canadian Medical Association Journal*, 105(7), pp. 738-741. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1931160/.

Goodwin, S., McPherson, J.D. y McCombie, W.R. (2016) «Coming of age: ten years of next-generation sequencing technologies.», *Nature reviews. Genetics*, 17, pp. 333-351. doi:10.1038/nrg.2016.49.

Green, E.D. y Guyer, M.S. (2011) «Charting a course for genomic medicine from base pairs to bedside», *Nature*, 470(7333), pp. 204-213. doi:10.1038/nature09764.

Green, R.C. *et al.* (2013) «ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing», *Genetics in Medicine*, 15, pp. 565-574. doi:10.1038/gim.2013.73.

Gubbels, C.S. *et al.* (2020) «Prospective, phenotype-driven selection of critically ill neonates for rapid exome sequencing is associated with high diagnostic yield», *Genetics in Medicine*, 22(4), pp. 736-744. doi:10.1038/s41436-019-0708-6.

Gusella, J.F. *et al.* (1983) «A polymorphic DNA marker genetically linked to Huntington's disease», *Nature*, 306(5940), pp. 234-238. doi:10.1038/306234a0.

Haer-Wigman, L. *et al.* (2019) «1 in 38 individuals at risk of a dominant medically actionable disease», *European Journal of Human Genetics*, 27(2), pp. 325-330.

doi:10.1038/s41431-018-0284-2.
Harper, P.S. (2007) «Paul Polani and the development of medical genetics», *Human Genetics*, 120(5), pp. 723-731. doi:10.1007/s00439-006-0271-5.
Harris, H. (1953) *An introduction to human biochemical genetics*. Cambridge: Cambridge University Press (37).
Harrow, J. *et al.* (2006) «GENCODE: producing a reference annotation for ENCODE», *Genome Biology*, 7(1), p. S4. doi:10.1186/gb-2006-7-s1-s4.
Hodges, E. *et al.* (2007) «Genome-wide in situ exon capture for selective resequencing», *Nature Genetics*, 39(12), pp. 1522-1527. doi:10.1038/ng.2007.42.
Holland, P.M. *et al.* (1991) «Detection of specific polymerase chain reaction product by utilizing the 5'—-3' exonuclease activity of Thermus aquaticus DNA polymerase.», *Proceedings of the National Academy of Sciences*, 88(16), pp. 7276-7280. doi:10.1073/pnas.88.16.7276.
Hoßfeld, U. *et al.* (2017) «150 years of Johann Gregor Mendel's "Versuche über Pflanzen-Hybriden".», *Molecular genetics and genomics : MGG*, 292, pp. 1-3. doi:10.1007/s00438-016-1254-4.
Hsu, F. *et al.* (2006) «The UCSC Known Genes», *Bioinformatics*, 22(9), pp. 1036-1046. doi:10.1093/bioinformatics/btl048.
Hubbard, T.J.P. *et al.* (2009) «Ensembl 2009», *Nucleic Acids Research*, 37(suppl_1), pp. D690-D697. doi:10.1093/nar/gkn828.
Huff, S.M. *et al.* (1998) «Development of the Logical Observation Identifier Names and Codes (LOINC) Vocabulary», *Journal of the American Medical Informatics Association*, 5(3), pp. 276-292. doi:10.1136/jamia.1998.0050276.
Ihaka, R. y Gentleman, R. (1996) «R: a Language for Data Analysis and Graphics», *Journal of Computational and Graphical Statistics*, 5(3), pp. 299-314. doi:10.1080/10618600.1996.10474713.
International Human Genome Sequencing Consortium (2004) «Finishing the euchromatic sequence of the human genome.», *Nature*, 431, pp. 931-945. doi:10.1038/nature03001.
Jacob, F. y Monod, J. (1961) «Genetic regulatory mechanisms in the synthesis of proteins», *Journal of Molecular Biology*, 3(3), pp. 318-356. doi:10.1016/S0022-2836(61)80072-7.
Jacob, H.J. *et al.* (2013) «Genomics in clinical practice: lessons from the front lines.», *Science translational medicine*, 5, p. 194cm5. doi:10.1126/scitranslmed.3006468.
Jacobs, P.A. y Strong, J.A. (1959) «A Case of Human Intersexuality Having a Possible XXY Sex-Determining Mechanism», *Nature*, 183(4657), pp. 302-303. doi:10.1038/183302a0.
Johannsen, W. (1914) «Elemente der exakten Erblichkeitslehre. Mit Grundzügen der biologischen Variationsstatistik», *Zeitschrift für induktive Abstammungs-und Vererbungslehre*, 11(1), pp. 200-200. doi:10.1007/BF01704312.
Johnston, J.J. *et al.* (2012) «Secondary Variants in Individuals Undergoing Exome Sequencing: Screening of 572 Individuals Identifies High-Penetrance Mutations in Cancer-Susceptibility Genes», *The American Journal of Human Genetics*, 91(1), pp. 97-108. doi:10.1016/j.ajhg.2012.05.021.
Josefsson, S. (2006) *The Base16, Base32, and Base64 Data Encodings*. RFC 4648.

doi:10.17487/RFC4648.
Kalia, S.S. *et al.* (2017) «Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics», *Genetics in Medicine*, 19(2, 2), pp. 249-255. doi:10.1038/gim.2016.190.
Kallioniemi, A. *et al.* (1992) «Comparative Genomic Hybridization for Molecular Cytogenetic Analysis of Solid Tumors», *Science*, 258(5083), pp. 818-821. doi:10.1126/science.1359641.
Kan, Y.W. y Dozy, A.M. (1978) «Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation.», *Proceedings of the National Academy of Sciences*, 75(11), pp. 5631-5635. doi:10.1073/pnas.75.11.5631.
Karczewski, K.J. *et al.* (2020) «The mutational constraint spectrum quantified from variation in 141,456 humans», *Nature*, 581(7809), pp. 434-443. doi:10.1038/s41586-020-2308-7.
Karczewski, K.J. y Snyder, M.P. (2018) «Integrative omics for health and disease», *Nature reviews. Genetics* [Preprint]. doi:10.1038/nrg.2018.4.
Katsonis, P. *et al.* (2014) «Single nucleotide variations: Biological impact and theoretical interpretation», *Protein Science*, 23(12), pp. 1650-1666. doi:10.1002/pro.2552.
Khera, A.V. *et al.* (2018) «Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations», *Nature Genetics*, 50(9), pp. 1219-1224. doi:10.1038/s41588-018-0183-z.
Korhonen, P.K. *et al.* (2019) «Common workflow language (CWL)-based software pipeline for de novo genome assembly from long- and short-read data», *GigaScience*, 8(4), p. giz014. doi:10.1093/gigascience/giz014.
Kosaki, R. *et al.* (2020) «Consecutive medical exome analysis at a tertiary center: Diagnostic and health-economic outcomes.», *American journal of medical genetics. Part A*, 182, pp. 1601-1607. doi:10.1002/ajmg.a.61589.
Kremer, L.S. *et al.* (2017) «Genetic diagnosis of Mendelian disorders via RNA sequencing.», *Nature communications*, 8, p. 15824. doi:10.1038/ncomms15824.
Krumm, N. y Hoffman, N. (2020) «Practical estimation of cloud storage costs for clinical genomic data», *Practical Laboratory Medicine*, 21, p. e00168. doi:10.1016/j.plabm.2020.e00168.
Kumar-Sinha, C. y Chinnaiyan, A.M. (2018) «Precision oncology in the age of integrative genomics.», *Nature biotechnology*, 36, pp. 46-60. doi:10.1038/nbt.4017.
Lander, E.S. *et al.* (2001) «Initial sequencing and analysis of the human genome.», *Nature*, 409, pp. 860-921. doi:10.1038/35057062.
Landrum, M.J. *et al.* (2014) «ClinVar: public archive of relationships among sequence variation and human phenotype», *Nucleic Acids Research*, 42(D1), pp. D980-D985. doi:10.1093/nar/gkt1113.
Landry, L.G. *et al.* (2018) «Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice», *Health Affairs*, 37(5), pp. 780-785. doi:10.1377/hlthaff.2017.1595.
Lappalainen, I. *et al.* (2013) «dbVar and DGVa: public archives for genomic structural variation», *Nucleic Acids Research*, 41(D1), pp. D936-D941.

doi:10.1093/nar/gks1213.

Lavelle, T.A. *et al.* (2022) «Cost-effectiveness of exome and genome sequencing for children with rare and undiagnosed conditions», *Genetics in Medicine*, 0(0). doi:10.1016/j.gim.2022.03.005.

Lazaridis, K.N. *et al.* (2016) «Outcome of Whole Exome Sequencing for Diagnostic Odyssey Cases of an Individualized Medicine Clinic: The Mayo Clinic Experience», *Mayo Clinic Proceedings*, 91(3), pp. 297-307. doi:10.1016/j.mayocp.2015.12.018.

Lejeune, J., Gautier, M. y Turpin, R. (1959) «Etude des chromosomes somatiques de neuf enfants mongoliens.», *Comptes Rend Acad Sci Paris*, 248(11), pp. 1721-2. Disponible en: https://www.lissa.fr/rep/articles/13639368 (Accedido: 15 de julio de 2022).

Levenson, D. (2016) «23andMe markets carrier screening service directly to consumers», *American Journal of Medical Genetics Part A*, 170(2), pp. 293-294. doi:10.1002/ajmg.a.37305.

Li, C. *et al.* (2020) «Cost-effectiveness of genome-wide sequencing for unexplained developmental disabilities and multiple congenital anomalies.», *Genetics in medicine : official journal of the American College of Medical Genetics* [Preprint]. doi:10.1038/s41436-020-01012-w.

Li, H. *et al.* (2009) «The Sequence Alignment/Map format and SAMtools», *Bioinformatics*, 25(16), pp. 2078-2079. doi:10.1093/bioinformatics/btp352.

Li, M.M. *et al.* (2017) «Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists», *The Journal of Molecular Diagnostics*, 19(1), pp. 4-23. doi:10.1016/j.jmoldx.2016.10.002.

Li, M.M. *et al.* (2022) «Clinical evaluation and etiologic diagnosis of hearing loss: A clinical practice resource of the American College of Medical Genetics and Genomics (ACMG)», *Genetics in Medicine*, 0(0). doi:10.1016/j.gim.2022.03.018.

Liu, L. *et al.* (2012) «Comparison of Next-Generation Sequencing Systems», *Journal of Biomedicine and Biotechnology*, 2012, p. e251364. doi:10.1155/2012/251364.

Lunke, S. *et al.* (2020) «Feasibility of Ultra-Rapid Exome Sequencing in Critically Ill Infants and Children With Suspected Monogenic Conditions in the Australian Public Health Care System», *JAMA*, 323(24), pp. 2503-2511. doi:10.1001/jama.2020.7671.

Machini, K. *et al.* (2019) «Analyzing and Reanalyzing the Genome: Findings from the MedSeq Project», *The American Journal of Human Genetics*, 105(1), pp. 177-188. doi:10.1016/j.ajhg.2019.05.017.

Maddox, B. (2003) «The double helix and the 'wronged heroine'», *Nature*, 421(6921), pp. 407-408. doi:10.1038/nature01399.

Maher, B. (2012) «ENCODE: The human encyclopaedia», *Nature*, 489(7414), pp. 46-48. doi:10.1038/489046a.

Major, E. *et al.* (2013) «HLA Typing from 1000 Genomes Whole Genome and Whole Exome Illumina Data», *PLOS ONE*, 8(11), p. e78410. doi:10.1371/journal.pone.0078410.

Mandelker, D. *et al.* (2014) «Comprehensive Diagnostic Testing for Stereocilin: An Approach for Analyzing Medically Important Genes with High Homology», *The Journal of Molecular Diagnostics*, 16(6), pp. 639-647. doi:10.1016/j.jmoldx.2014.06.003.

Mandelker, D. *et al.* (2016) «Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing», *Genetics in Medicine*, 18(12), pp. 1282-1289. doi:10.1038/gim.2016.58.

Manolio, T.A. *et al.* (2015) «Global implementation of genomic medicine: We are not alone», *Science Translational Medicine*, 7(290), pp. 290ps13-290ps13. doi:10.1126/scitranslmed.aab0194.

Margulies, M. *et al.* (2005) «Genome sequencing in microfabricated high-density picolitre reactors», *Nature*, 437(7057), pp. 376-380. doi:10.1038/nature03959.

Marshall, C.R. *et al.* (2020) «Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease», *npj Genomic Medicine*, 5(1), pp. 1-12. doi:10.1038/s41525-020-00154-9.

Martin, A.R. *et al.* (2019) «PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels», 51, pp. 1560-1565. doi:10.1038/s41588-019-0528-2.

Matthijs, G. *et al.* (2016) «Guidelines for diagnostic next-generation sequencing.», *European journal of human genetics : EJHG*, 24, pp. 2-5. doi:10.1038/ejhg.2015.226.

Mattocks, C.J. *et al.* (2010) «A standardized framework for the validation and verification of clinical molecular genetic tests», *European Journal of Human Genetics*, 18(12), pp. 1276-1288. doi:10.1038/ejhg.2010.101.

Maxam, A.M. y Gilbert, W. (1977) «A new method for sequencing DNA.», *Proc. of the National Academy of Sciences*, 74(2), pp. 560-564. doi:10.1073/pnas.74.2.560.

McCombie, W.R., McPherson, J.D. y Mardis, E.R. (2019) «Next-Generation Sequencing Technologies», *Cold Spring Harbor Perspectives in Medicine*, 9(11), p. a036798. doi:10.1101/cshperspect.a036798.

McKenna, A. *et al.* (2010) «The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data», *Genome Research*, 20(9), pp. 1297-1303. doi:10.1101/gr.107524.110.

McKusick, V.A. (1975) «The growth and development of human genetics as a clinical discipline», *American Journal of Human Genetics*, 27(3), pp. 261-273.

McKusick, V.A. (1993) «Medical Genetics: A 40-Year Perspective on the Evolution of a Medical Specialty From a Basic Science», *JAMA*, 270(19), pp. 2351-2356. doi:10.1001/jama.1993.03510190107035.

McKusick, V.A. (1998) *Mendelian Inheritance in Man: A catalog of human genes and genetic disorders*. 12th edn. Baltimore: Johns Hopkins University Press.

McKusick, V.A. y Ruddle, F.H. (1987) «A new discipline, a new name, a new journal», *Genomics*, 1(1), pp. 1-2. doi:10.1016/0888-7543(87)90098-X.

McLaren, W. *et al.* (2016) «The Ensembl Variant Effect Predictor», *Genome biology*, 17, p. 122. doi:10.1186/s13059-016-0974-4.

McPherson, E. (2006) «Genetic Diagnosis and Testing in Clinical Practice», *Clinical Medicine & Research*, 4(2), pp. 123-129. doi:10.3121/cmr.4.2.123.

Meloni, V. *et al.* (2015) «HL7apy: a Python library to parse, create and handle HL7

v2.x messages», *EJBI - European Journal for Biomedical Informatics*, 11(2), pp. en31-en40. doi:10.24105/ejbi.2015.11.2.6.

Mendel, G. (1866) «Versuche über Pflanzenhybride», *Verhandlungen desNaturforschenden Vereins in Brünn*, 4, pp. 3-47.

Metzker, M.L. (2010) «Sequencing technologies - the next generation.», *Nature reviews. Genetics*, 11, pp. 31-46. doi:10.1038/nrg2626.

Miller, D.T. *et al.* (2010) «Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies», *The American Journal of Human Genetics*, 86(5), pp. 749-764. doi:10.1016/j.ajhg.2010.04.006.

Miller, D.T. *et al.* (2021) «ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG).», *Genetics in Medicine* [Preprint]. doi:10.1038/s41436-021-01172-3.

Miller, D.T. *et al.* (2022) «ACMG SF v3.1 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG)», *Genetics in Medicine*, 24(7), pp. 1407-1414. doi:10.1016/j.gim.2022.04.006.

Mills, M.C. y Rahal, C. (2020) «The GWAS Diversity Monitor tracks diversity by disease in real time.», *Nature genetics*, 52, pp. 242-243. doi:10.1038/s41588-020-0580-y.

Moorthie, S., Hall, A. y Wright, C.F. (2013) «Informatics and clinical genome sequencing: opening the black box», *Genetics in Medicine*, 15(3), pp. 165-171. doi:10.1038/gim.2012.116.

Motulsky, A.G. y King, M.-C. (2016) «The Great Adventure of an American Human Geneticist.», *Annual review of genomics and human genetics*, 17, pp. 1-15. doi:10.1146/annurev-genom-083115-022528.

Mullis, K.B. y Faloona, F.A. (1987) «[21] Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction», en. Academic Press (Recombinant DNA Part F), pp. 335-350. doi:10.1016/0076-6879(87)55023-6.

Murphy, S.L. *et al.* (2021) *Mortality in the United States, 2020*. National Center for Health Statistics (U.S.). doi:10.15620/cdc:112079.

National Institute of Health (2016) «The Cost of Sequencing a Human Genome». Disponible en: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (Accedido: 29 de mayo de 2022).

Ng, S.B. *et al.* (2009) «Targeted capture and massively parallel sequencing of 12 human exomes.», *Nature*, 461, pp. 272-276. doi:10.1038/nature08250.

Ng, S.B. *et al.* (2010) «Exome sequencing identifies the cause of a mendelian disorder», *Nature Genetics*, 42(1), pp. 30-35. doi:10.1038/ng.499.

Nurk, S. *et al.* (2022) «The complete sequence of a human genome», *Science*, 376(6588), pp. 44-53. doi:10.1126/science.abj6987.

O'Rawe, J. *et al.* (2013) «Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing», *Genome Medicine*, 5(3), p. 28. doi:10.1186/gm432.

Olby, R. (2003) «Quiet debut for the double helix», *Nature*, 421(6921), pp. 402-405.

doi:10.1038/nature01397.
Oliver, G.R., Hart, S.N. y Klee, E.W. (2015) «Bioinformatics for Clinical Next Generation Sequencing», *Clinical Chemistry*, 61(1), pp. 124-135. doi:10.1373/clinchem.2014.224360.
Pabinger, S. *et al.* (2014) «A survey of tools for variant analysis of next-generation genome sequencing data», *Briefings in Bioinformatics*, 15(2), pp. 256-278. doi:10.1093/bib/bbs086.
Passarge, E. (2021) «Origins of human genetics. A personal perspective», *European Journal of Human Genetics*, 29(7), pp. 1038-1044. doi:10.1038/s41431-020-00785-7.
Payne, K. *et al.* (2018) «Cost-effectiveness analyses of genetic and genomic diagnostic tests.», *Nature reviews. Genetics* [Preprint]. doi:10.1038/nrg.2017.108.
Peplow, M. (2016) «The 100 000 Genomes Project», *BMJ*, 353, p. i1757. doi:10.1136/bmj.i1757.
Philippakis, A.A. *et al.* (2015) «The Matchmaker Exchange: a platform for rare disease gene discovery», *Human Mutation*, 36, pp. 915-921. doi:10.1002/humu.22858.
Piening, B. *et al.* (2018) «Integrative Personal Omics Profiles during Periods of Weight Gain and Loss.», *Cell systems*, 6, pp. 157-170. doi:10.1016/j.cels.2017.12.013.
Polani, P., Hunter, W.F. y Lennox, B. (1954) «Chromosomal sex in Turner's syndrome with coarctation of the aorta», *The Lancet*, 264(6829), pp. 120-121. doi:10.1016/S0140-6736(54)90100-2.
Pollack, J.R. *et al.* (1999) «Genome-wide analysis of DNA copy-number changes using cDNA microarrays», *Nature Genetics*, 23(1), pp. 41-46. doi:10.1038/12640.
Portin, P. y Wilkins, A. (2017) «The Evolving Definition of the Term "Gene"», *Genetics*, 205(4), pp. 1353-1364. doi:10.1534/genetics.116.196956.
Primrose, S.B. y Twyman, R.M. (2003) *Principles of genome analysis and genomics*. Blackwell. Disponible en: https://www.wiley.com/en-us/Principles of Genome Analysis and Genomics, 3rd Edition-p-9781405101202.
Pruitt, K.D. *et al.* (2009) «The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes», *Genome Research*, 19(7), pp. 1316-1323. doi:10.1101/gr.080531.108.
Pruitt, K.D., Tatusova, T. y Maglott, D.R. (2007) «NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins», *Nucleic Acids Research*, 35(suppl_1), pp. D61-D65. doi:10.1093/nar/gkl842.
Puck, T.T. y Kao, F. (1982) «Somatic Cell Genetics and Its Application to Medicine», *Ann. Rev. Genetics*, 16, pp. 225-271. doi:10.1146/annurev.ge.16.120182.001301.
Pujol, P. *et al.* (2018) «Guidelines for reporting secondary findings of genome sequencing in cancer genes: the SFMPP recommendations», *European Journal of Human Genetics*, 26(12), pp. 1732-1742. doi:10.1038/s41431-018-0224-1.
Quinlan, A.R. (2014) «BEDTools: The Swiss-Army Tool for Genome Feature Analysis», *Current Protocols in Bioinformatics*, 47(1), pp. 11.12.1-11.12.34. doi:10.1002/0471250953.bi1112s47.
R Core Team (2020) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Disponible en: https://www.R-

project.org/.
Rasmussen, L.V. *et al.* (2016) «Practical considerations for implementing genomic information resources», *Applied Clinical Informatics*, 07(03), pp. 870-882. doi:10.4338/ACI-2016-04-RA-0060.
Recchia, G. *et al.* (2020) «Creating genetic reports that are understood by nonspecialists: a case study», *Genetics in Medicine*, 22(2), pp. 353-361. doi:10.1038/s41436-019-0649-0.
Reese, M.G. *et al.* (2010) «A standard variation file format for human genome sequences», *Genome Biology*, 11(8), p. R88. doi:10.1186/gb-2010-11-8-r88.
Rehder, C. *et al.* (2021) «Next-generation sequencing for constitutional variants in the clinical laboratory, 2021 revision: a technical standard of the American College of Medical Genetics and Genomics (ACMG)», *Genetics in Medicine*, 23(8), pp. 1399-1415. doi:10.1038/s41436-021-01139-4.
Rehm, H.L. *et al.* (2013) «ACMG clinical laboratory standards for next-generation sequencing», *Genetics in Medicine*, 15(9), pp. 733-747. doi:10.1038/gim.2013.92.
Rehm, H.L. *et al.* (2015) «ClinGen - the Clinical Genome Resource.», *The New England journal of medicine*, 372, pp. 2235-2242. doi:10.1056/NEJMsr1406261.
Rehm, H.L. (2017) «Evolving health care through personal genomics.», *Nature reviews. Genetics*, 18, pp. 259-267. doi:10.1038/nrg.2016.162.
Richards, S. *et al.* (2015) «Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.», *Genetics in medicine : official journal of the American College of Medical Genetics*, 17, pp. 405-424. doi:10.1038/gim.2015.30.
Ritchie, G.R. y Flicek, P. (2014) «Computational approaches to interpreting genomic sequence variation.», *Genome medicine*, 6, p. 87. doi:10.1186/s13073-014-0087-1.
Ritchie, G.R.S. *et al.* (2014) «Functional annotation of noncoding sequence variants», *Nature Methods*, 11(3), pp. 294-296. doi:10.1038/nmeth.2832.
Roach, J.C. *et al.* (2010) «Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing», *Science*, 328(5978), pp. 636-639. doi:10.1126/science.1186802.
Roberts, J.S. *et al.* (2017) «Direct-to-Consumer Genetic Testing: User Motivations, Decision Making, and Perceived Utility of Results», *Public Health Genomics*, 20(1), pp. 36-45. doi:10.1159/000455006.
Robertson, A.J. *et al.* (2022) «Re-analysis of genomic data: An overview of the mechanisms and complexities of clinical adoption», *Genetics in Medicine*, 24(4), pp. 798-810. doi:10.1016/j.gim.2021.12.011.
Rodgers, G.P. y Collins, F.S. (2020) «Precision Nutrition - the Answer to "What to Eat to Stay Healthy".», *JAMA*, 324, pp. 735-736. doi:10.1001/jama.2020.13601.
Ronaghi, M. *et al.* (1996) «Real-Time DNA Sequencing Using Detection of Pyrophosphate Release», *Analytical Biochemistry*, 242(1), pp. 84-89. doi:10.1006/abio.1996.0432.
Rossum, G. van y Boer, J. de (1991) «Interactively testing remote servers using the Python programming language», *CWI Quarterly*, 4(4), pp. 283-304. Disponible

en: https://ir.cwi.nl/pub/18204 (Accedido: 19 de julio de 2022).
Roy, S. *et al.* (2018) «Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists», *The Journal of Molecular Diagnostics*, 20(1), pp. 4-27. doi:10.1016/j.jmoldx.2017.11.003.
Rubinstein, W.S. *et al.* (2013) «The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency», *Nucleic Acids Research*, 41(D1), pp. D925-D935. doi:10.1093/nar/gks1173.
Saiki, R.K. *et al.* (1986) «Analysis of enzymatically amplified β-globin and HLA-DQα DNA with allele-specific oligonucleotide probes», *Nature*, 324(6093), pp. 163-166. doi:10.1038/324163a0.
Saklatvala, J.R., Dand, N. y Simpson, M.A. (2018) «Text-mined phenotype annotation and vector-based similarity to improve identification of similar phenotypes and causative genes in monogenic disease patients», *Human Mutation*, 39(5), pp. 643-652. doi:10.1002/humu.23413.
Sanger, F., Nicklen, S. y Coulson, A.R. (1977) «DNA sequencing with chain-terminating inhibitors», *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp. 5463-5467. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC431765/ (Accedido: 25 de abril de 2022).
Saripalle, R., Runyan, C. y Russell, M. (2019) «Using HL7 FHIR to achieve interoperability in patient health record», *Journal of Biomedical Informatics*, 94, p. 103188. doi:10.1016/j.jbi.2019.103188.
Sawyer, S.L. *et al.* (2016) «Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care.», *Clinical genetics*, 89, pp. 275-284. doi:10.1111/cge.12654.
Schneider, V.A. *et al.* (2017) «Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly», *Genome Research*, 27(5), pp. 849-864. doi:10.1101/gr.213611.116.
Schön, U. *et al.* (2021) «HPO-driven virtual gene panel: a new efficient approach in molecular autopsy of sudden unexplained death», *BMC Medical Genomics*, 14(1), p. 94. doi:10.1186/s12920-021-00946-7.
Schork, N.J. (2015) «Personalized medicine: Time for one-person trials», *Nature*, 520(7549), pp. 609-611. doi:10.1038/520609a.
Sehn, J.K. (2015) «Chapter 9 - Insertions and Deletions (Indels)», en Kulkarni, S. y Pfeifer, J. (eds.). Boston: Academic Press, pp. 129-150. doi:10.1016/B978-0-12-404748-8.00009-5.
Service, R.F. (2006) «Gene sequencing. The race for the $1000 genome.», *Science (New York, N.Y.)*, 311, pp. 1544-1546. doi:10.1126/science.311.5767.1544.
Shendure, J. *et al.* (2017) «DNA sequencing at 40: past, present and future», *Nature*, 550(7676), pp. 345-353. doi:10.1038/nature24286.
Sherry, S.T., Ward, M. y Sirotkin, K. (1999) «dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation», *Genome Research*, 9(8), pp. 677-679. doi:10.1101/gr.9.8.677.

Snijders, A.M. *et al.* (2001) «Assembly of microarrays for genome-wide measurement of DNA copy number», *Nature Genetics*, 29(3), pp. 263-264. doi:10.1038/ng754.

Solomon, E. y Bodmer, W.F. (1979) «Evolution of sickle variant gene», *The Lancet*, 313(8122), p. 923. doi:10.1016/S0140-6736(79)91398-9.

Speicher, M.R. y Carter, N.P. (2005) «The new cytogenetics: blurring the boundaries with molecular biology», *Nature Reviews Genetics*, 6(10), pp. 782-792. doi:10.1038/nrg1692.

Staden, R. (1979) «A strategy of DNA sequencing employing computer programs», *Nucleic Acids Research*, 6(7), pp. 2601-2610. doi:10.1093/nar/6.7.2601.

Stark, Z. *et al.* (2017) «Prospective comparison of the cost-effectiveness of clinical whole-exome sequencing with that of usual care overwhelmingly supports early use and reimbursement.», *Genetics in medicine : official journal of the American College of Medical Genetics*, 19, pp. 867-874. doi:10.1038/gim.2016.221.

Stark, Z., Schofield, D., *et al.* (2019) «Does genomic sequencing early in the diagnostic trajectory make a difference? A follow-up study of clinical outcomes and cost-effectiveness», *Genetics in Medicine*, 21(1), pp. 173-180. doi:10.1038/s41436-018-0006-8.

Stark, Z., Dolman, L., *et al.* (2019) «Integrating Genomics into Healthcare: A Global Responsibility», *The American Journal of Human Genetics*, 104(1), pp. 13-20. doi:10.1016/j.ajhg.2018.11.014.

Stark, Z. *et al.* (2021) «Scaling national and international improvement in virtual gene panel curation via a collaborative approach to discordance resolution», *The American Journal of Human Genetics*, 108(9), pp. 1551-1557. doi:10.1016/j.ajhg.2021.06.020.

Stein, L. (2001) «Genome annotation: from sequence to biology», *Nature Reviews Genetics*, 2(7), pp. 493-503. doi:10.1038/35080529.

Stein, L.D. (2010) «The case for cloud computing in genome informatics», *Genome Biology*, 11(5), p. 207. doi:10.1186/gb-2010-11-5-207.

Stenson, P.D. *et al.* (2008) «Human Gene Mutation Database: towards a comprehensive central mutation database.», *Journal of medical genetics*, 45, pp. 124-126. doi:10.1136/jmg.2007.055210.

Strozzi, F. *et al.* (2019) «Scalable Workflows and Reproducible Data Analysis for Genomics», en Anisimova, M. (ed.). New York, NY: Springer (Methods en Molecular Biology), pp. 723-745. doi:10.1007/978-1-4939-9074-0_24.

Tan, N.B. *et al.* (2020) «Evaluating systematic reanalysis of clinical genomic data in rare disease from single center experience and literature review», *Molecular Genetics & Genomic Medicine*, 8(11), p. e1508. doi:10.1002/mgg3.1508.

Tandy-Connor, S. *et al.* (2018) «False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care», *Genetics in Medicine*, 20(12), pp. 1515-1521. doi:10.1038/gim.2018.38.

The 1000 Genomes Project Consortium (2010) «A map of human genome variation from population-scale sequencing», *Nature*, 467(7319), pp. 1061-1073. doi:10.1038/nature09534.

The 1000 Genomes Project Consortium (2015) «A global reference for human gene-

tic variation», *Nature*, 526(7571), pp. 68-74. doi:10.1038/nature15393.
The International HapMap Consortium (2003) «The International HapMap Project», *Nature*, 426(6968), pp. 789-796. doi:10.1038/nature02168.
The NICUSeq Study Group (2021) «Effect of Whole-Genome Sequencing on the Clinical Management of Acutely Ill Infants With Suspected Genetic Disease: A Randomized Clinical Trial», *JAMA Pediatrics*, 175(12), pp. 1218-1226. doi:10.1001/jamapediatrics.2021.3496.
Thevenon, J. *et al.* (2016) «Diagnostic odyssey in severe neurodevelopmental disorders: toward clinical whole-exome sequencing as a first-line diagnostic test», *Clinical Genetics*, 89(6), pp. 700-707. doi:10.1111/cge.12732.
Thuriot, F. *et al.* (2018) «Clinical validity of phenotype-driven analysis software PhenoVar as a diagnostic aid for clinical geneticists in the interpretation of whole-exome sequencing data.», *Genetics in medicine : official journal of the American College of Medical Genetics* [Preprint]. doi:10.1038/gim.2017.239.
Thusberg, J., Olatubosun, A. y Vihinen, M. (2011) «Performance of mutation pathogenicity prediction methods on missense variants», *Human Mutation*, 32(4), pp. 358-368. doi:10.1002/humu.21445.
Treangen, T.J. y Salzberg, S.L. (2012) «Repetitive DNA and next-generation sequencing: computational challenges and solutions», *Nature Reviews Genetics*, 13(1), pp. 36-46. doi:10.1038/nrg3117.
Van Ness, B. (2008) «Genomic Research and Incidental Findings», *Journal of Law, Medicine & Ethics*, 36(2), pp. 292-297. doi:10.1111/j.1748-720X.2008.00272.x.
Venter, J.C. *et al.* (2001) «The Sequence of the Human Genome», *Science*, 291(5507), pp. 1304-1351. doi:10.1126/science.1058040.
Voelkerding, K.V., Dames, S.A. y Durtschi, J.D. (2009) «Next-Generation Sequencing: From Basic Research to Diagnostics», *Clinical Chemistry*, 55(4), pp. 641-658. doi:10.1373/clinchem.2008.112789.
Vujkovic, M. *et al.* (2020) «Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis», *Nature Genetics*, 52(7), pp. 680-691. doi:10.1038/s41588-020-0637-y.
Wagner, A.H. *et al.* (2021) «The GA4GH Variation Representation Specification: A computational framework for variation representation and federated identification», *Cell Genomics*, 1(2), p. 100027. doi:10.1016/j.xgen.2021.100027.
Wallace, D.C. (2018) «Mitochondrial genetic medicine», *Nature Genetics*, 50(12), pp. 1642-1649. doi:10.1038/s41588-018-0264-z.
Walsh, M. *et al.* (2017) «Diagnostic and cost utility of whole exome sequencing in peripheral neuropathy», *Annals of Clinical and Translational Neurology*, 4(5), pp. 318-325. doi:10.1002/acn3.409.
Wang, H. *et al.* (2020) «Clinical utility of 24-h rapid trio-exome sequencing for critically ill infants», *npj Genomic Medicine*, 5(1), pp. 1-6. doi:10.1038/s41525-020-0129-0.
Wang, K., Li, M. y Hakonarson, H. (2010) «ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data», *Nucleic Acids Research*, 38(16), p. e164. doi:10.1093/nar/gkq603.
Wang, X. *et al.* (2019) «Phenotype-Driven Virtual Panel Is an Effective Method

to Analyze WES Data of Neurological Disease», *Frontiers in Pharmacology*, 9. doi:10.3389/fphar.2018.01529.

Watson, J.D. y Crick, F.H.C. (1953) «Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid», *Nature*, 171(4356), pp. 737-738. doi:10.1038/171737a0.

Wenger, A.M. *et al.* (2017) «Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers», *Genetics in Medicine*, 19(2), pp. 209-214. doi:10.1038/gim.2016.88.

Wert, G. de *et al.* (2021) «Opportunistic genomic screening. Recommendations of the European Society of Human Genetics», *European Journal of Human Genetics*, 29(3), pp. 365-377. doi:10.1038/s41431-020-00758-w.

Wetterstrand, K.A. (2022) «DNA Sequencing Costs: Data». NHGRI Genome Sequencing Program (GSP). Disponible en: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (Accedido: 29 de mayo de 2022).

Whiffin, N. *et al.* (2017) «Using high-resolution variant frequencies to empower clinical genome interpretation», *Genetics in Medicine*, 19(10), pp. 1151-1158. doi:10.1038/gim.2017.26.

Wilcox, E. *et al.* (2021) «Creation of an Expert Curated Variant List for Clinical Genomic Test Development and Validation: A ClinGen and GeT-RM Collaborative Project», *The Journal of Molecular Diagnostics*, 23(11), pp. 1500-1505. doi:10.1016/j.jmoldx.2021.07.018.

Wilson, J.M.G. y Jungner, G. (1968) «Principles and practice of screening for disease». World Health Organization. Disponible en: https://apps.who.int/iris/handle/10665/37650.

Wolf, S.M. *et al.* (2018) «Navigating the research–clinical interface in genomic medicine: analysis from the CSER Consortium», *Genetics in Medicine*, 20(5), pp. 545-553. doi:10.1038/gim.2017.137.

World Economic Forum (2020) «Precision Medicine Vision Statement: A Product of the World Economic Forum Global Precision Medicine Council», *World Economic Forum*. Disponible en: https://www.weforum.org/reports/precision-medicine-vision-statement-a-product-of-the-world-economic-forum-global-precision-medicine-council/ (Accedido: 29 de mayo de 2022).

Worthey, E.A. *et al.* (2011) «Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease.», *Genetics in medicine : official journal of the American College of Medical Genetics*, 13, pp. 255-262. doi:10.1097/GIM.0b013e3182088158.

Wortmann, S.B. *et al.* (2015) «Whole exome sequencing of suspected mitochondrial patients in clinical practice», *Journal of Inherited Metabolic Disease*, 38(3), pp. 437-443. doi:10.1007/s10545-015-9823-y.

Wright, C.F. *et al.* (2018) «Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders.», *Genetics in medicine : official journal of the American College of Medical Genetics*, 20, pp. 1216-1223. doi:10.1038/gim.2017.246.

Xiang, J. *et al.* (2020) «Reinterpretation of common pathogenic variants in Clin-

Var revealed a high proportion of downgrades», *Scientific Reports*, 10(1), p. 331. doi:10.1038/s41598-019-57335-5.

Yanofsky, C. (2007) «Establishing the Triplet Nature of the Genetic Code», *Cell*, 128(5), pp. 815-818. doi:10.1016/j.cell.2007.02.029.

Yeung, A. *et al.* (2020) «A cost-effectiveness analysis of genomic sequencing in a prospective versus historical cohort of complex pediatric patients.», *Genetics in medicine : official journal of the American College of Medical Genetics* [Preprint]. doi:10.1038/s41436-020-0929-8.

Yohe, S.L. *et al.* (2015) «Standards for Clinical Grade Genomic Databases.», *Archives of pathology & laboratory medicine*, 139, pp. 1400-1412. doi:10.5858/arpa.2014-0568-CP.

Yu, Y. *et al.* (2012) «Exome and Whole-Genome Sequencing as Clinical Tests: A Transformative Practice in Molecular Diagnostics», *Clinical Chemistry*, 58(11), pp. 1507-1509. doi:10.1373/clinchem.2012.193128.

Zhao, X. *et al.* (2004) «An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays», *Cancer Research*, 64(9), pp. 3060-3071. doi:10.1158/0008-5472.CAN-03-3308.

Zook, J.M. *et al.* (2014) «Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls», *Nature Biotechnology*, 32(3), pp. 246-251. doi:10.1038/nbt.2835.