UNIVERSIDAD ADOLFO IBÁÑEZ

Genoma

PROYECTO DE PASANTÍA

Realizado en Genoma Mayor SpA.

Creación de Pipeline informático para la detección de variantes genéticas con reportes personalizados.

Valeria Pichott Román.

Proyecto para optar al título de Ingeniería Civil Informática e Ingeniería Civil en Bioingeniería de la Facultad de Ingeniería y Ciencias de la Universidad Adolfo Ibáñez.

Santiago, Chile.

2023

FACULTAD DE INGENIERÍA Y CIENCIAS

UNIVERSIDAD ADOLFO IBÁÑEZ

Índice

1. Resumen ejecutivo 3
2. Abstract 4
3. Introducción 5
   a. Contexto de la Empresa 5
   b. Oportunidad 6
   d. Cuantificar el problema 7
4. Objetivos 8
   a. Objetivo general 8
   b. Objetivos específicos 8
5. Estado del Arte 9
6. Solución escogida 13
7. Metodología 14
8. Medidas de desempeño 14
9. Desarrollo y Plan de implementación 15
   a. Fase 1 proyecto 15
   b. Fase 2 proyecto 18
   c. Fase 3 proyecto 19
10. Análisis de riesgo 22
11. Evaluación Económica 22
12. Resultados 24
13. Discusión 27
14. Conclusión 28
16. Anexo 31
   Contexto Conceptual 31
   a. Glosario 43

1. Resumen ejecutivo

En el avance de la lucha contra el cáncer, la detección de genes mutados emerge como un pilar fundamental, logrando personalizar el enfoque de diagnósticos y tratamientos mejorando las tasas de supervivencia. Este proyecto, desarrollado en colaboración con la empresa Genoma Mayor, se enfocó específicamente en la implementación de soluciones para la detección de variantes genéticas asociadas al cáncer, utilizando los datos de pacientes proporcionados por el centro en formato FASTQ.

Este trabajo se centró en el desarrollo de un sistema integral para detectar variantes genéticas relacionadas con el cáncer y generar informes subsiguientes. A través de la implementación de pipelines informáticos, programados en Python, se logró analizar datos genéticos provenientes de secuenciación de nueva generación, detectando estas variantes con una alta sensibilidad de un 87%.

Además, para asegurar la organización y accesibilidad de la información generada, se generó una base de datos con MySQL Workbench para almacenar los resultados obtenidos del pipeline. Esta base de datos se convirtió en un repositorio centralizado para la información relevante, facilitando la posterior extracción de datos para su análisis y referencia.

Como parte final del proyecto, con Rmarkdown, se logró implementar la automatización de la generación de reportes utilizando la información recopilada de la base de datos, proporcionando un panorama claro y conciso de las variantes identificadas, junto con datos contextualizados para apoyar la toma de decisiones clínicas.

La exitosa integración de las tres fases de este proyecto no solo ha permitido detectar de manera eficaz variantes genéticas asociadas al cáncer, sino que también ha llevado a la sistematización del proceso. Ahora, con solo indicar el archivo de entrada, el sistema genera el producto final, agilizando considerablemente la generación de informes para su implementación en entornos clínicos. Este proyecto logró reducir el tiempo de obtención de reportes en un 48% y disminuir los costos asociados en un 50%.

Palabras clave: Medicina de precisión, Detección de Variantes, Genómica, Base de datos, Cáncer, ADN.

2. Abstract

In the advancement of the fight against cancer, the detection of mutated genes emerges as a fundamental pillar, achieving the customization of diagnostic and treatment approaches, thereby improving survival rates. This project, developed in collaboration with Genoma Mayor, focused on implementing solutions for the detection of genetic variants associated with cancer using patient data provided by the center, which included FASTQ sequencing files.

This project focused on the development of a comprehensive system to detect genetic variants related to cancer and generate subsequent reports. Through the implementation of computer pipelines, programmed in Python, it was possible to analyze genetic data from next-generation sequencing, detecting these variants with a high sensitivity of 87%.

Additionally, to ensure the organization and accessibility of the generated information, a database was created using MySQL Workbench to store the results obtained from the pipeline. This database became a centralized repository for relevant information, facilitating the subsequent extraction of data for analysis and reference.

As a final part of the project, with Rmarkdown, the automation of report generation was successfully implemented using the information collected from the database, providing a clear and concise overview of the identified variants, along with contextualized data to support clinical decision-making.

The successful integration of the three phases of this project has not only allowed for the effective detection of genetic variants associated with cancer but has also led to the systematization of the process. Now, by simply indicating the input file, the system generates the final product, significantly streamlining report generation for implementation in clinical environments. This project achieved a 48% reduction in report generation time and a 50% decrease in associated costs.

Keywords: Precision Medicine, Variant Detection, Genomics, Database, Cancer, DNA.

3. Introducción

Antes de adentrarse en el contexto, en el anexo se encuentra una introducción conceptual por si se requiere. Además, cada concepto que tenga * a su lado es porque su definición se encuentra en el glosario que también está en el anexo.

a. Contexto de la Empresa

Genoma Mayor SpA. Es una empresa fundada el 2010 y derivada de la Universidad Mayor, conocida por ser un centro de vanguardia especializada en la tecnología de secuenciación de nueva generación (NGS). Se dedicada a la entrega de servicios tanto de bioinformática y secuenciación como de asesoría genética, esta última mediante la comercialización de pruebas genéticas estandarizadas las cuales requieren muestras biológicas tales como sangre o saliva, y que operan según secuenciación NGS*, con la finalidad de analizar variantes genéticas. Tras doce años de experiencia, se ha expandido a otras áreas de las tecnologías ómicas y ahora cuenta con múltiples plataformas para análisis en Genómica y Metabolómica; como también flujos de trabajo de laboratorio y de bioinformática desarrollados y validados.

Genoma se compone de tres áreas principales de desarrollo que, al combinarse, impulsan la empresa y contribuyen al avance en genética y medicina. La primera se centra en ensayos clínicos de medicina de precisión, donde se llevan a cabo estudios clínicos para mejorar el tratamiento de enfermedades. La segunda se enfoca en exámenes de medicina de precisión molecular, empleando tecnologías avanzadas como la NGS, Sanger* y qPCR*. Por último, la tercera área se dedica a la investigación e informática, aplicando la genómica y la bioinformática para respaldar investigaciones.

Por otro lado, y lo que se relaciona directamente con mi proyecto, Genoma Mayor ejerce el control administrativo del Centro de Oncología de Precisión de la Universidad Mayor (COP), el cual busca implementar un modelo de atención clínica, basado en medicina personalizada para pacientes con cáncer, utilizando herramientas traslacionales, como la ya mencionada, secuenciación NGS. Los servicios clínicos ofrecidos por esta división de la empresa constan de, Consultas Médicas de especialistas, en la cual se consigue atender diversos tipos de patologías. Asimismo, se cuenta con las Consultas de Asesoría Genética, un proceso disciplinario en el que pacientes son advertidos, prevenidas y tratadas clínicamente de acuerdo con el perfil genotípico que posean. Esta área se centra en ayudar a las personas que están afectadas o tienen un mayor riesgo de enfermedades genéticas. Sin embargo, el servicio que más destaca es sin duda el Comité Oncológico Molecular, el cual es un grupo de profesionales de la salud, compuesto por oncólogos, genetistas, patólogos moleculares y

otros expertos en oncología, que se reúnen para discutir y tomar decisiones sobre el tratamiento de pacientes con cáncer, especialmente aquellos cuyos casos involucran aspectos moleculares y genéticos. En estos comités se revisa la información molecular y genética de los tumores de los pacientes, utilizando pruebas como la secuenciación genómica, para personalizar y optimizar el plan de tratamiento. La toma de decisiones basada en la información molecular permite una atención más precisa y adaptada a las características específicas del cáncer de cada paciente, lo que se conoce como medicina oncológica de precisión.

b. Oportunidad

En relación a lo mencionado anteriormente, Genoma presenta una gran oportunidad de mejora en el Comité Oncológico Molecular proveída por **COP**. En estos comités se analiza el contenido de los reportes entregados por una empresa externa. Esta empresa se encarga de secuenciar las muestras entregadas por Genoma y detectar las variantes genéticas con el propósito de entregar un reporte con los genes mutados más importantes que fueron identificados en el paciente, para luego, ser analizados en el comité. El objetivo de este equipo es discutir cuál sería el mejor tratamiento, ya sea quimioterapia, radioterapia, inmunoterapia, entre otros. También se evalúa el nivel de invasividad que estos tratamientos podrían tener y se analizan los pasos a seguir para los pacientes con cáncer, considerando sus resultados genéticos.

La dificultad que presentan en esta parte del trabajo es que el tiempo que transcurre desde que es enviada la muestra al momento de la devolución de los resultados en el reporte es excesivamente largo. Con una espera mínima de 31 días para la disponibilidad de información respecto a los genes mutados, a lo que hay que sumarle el tiempo que se demora en tener el comité y su posterior análisis. Es de gran importancia este último punto porque la espera prolongada, no solo genera una menor eficiencia, también impide a la empresa actuar rápido en casos médicos importantes, siendo el cáncer un gran ejemplo de esto.

Además, la identificación precisa de genes adquiere una importancia crucial en el ámbito oncológico. Los tratamientos para el cáncer son inherentemente costosos, y enfrentarse al riesgo de que no funcionen representa una pérdida considerable. Al conocer qué genes están mutados, la posibilidad de un tratamiento inefectivo disminuye, ya que se puede determinar de antemano cuál sería el enfoque terapéutico más efectivo y personalizado según la configuración genética del paciente. Por esta razón, es esencial establecer asociaciones con empresas que realicen identificaciones de genes de manera precisa. Sin embargo, este proceso a veces conlleva ciertas incertidumbres, dado que los métodos específicos empleados por cada empresa para identificar genes

son confidenciales, lo que limita el conocimiento sobre el procedimiento exacto de identificación genética. Considerando todo lo expuesto, ha surgido el anhelo de independizarse de la empresa externa, por una serie de criterios que afectan en la eficiencia y calidad del servicio entregado, buscando establecer un proceso propio de identificación genética que tome menos tiempo.

c. Cuantificar el problema

Para cuantificar el problema, se llevará a cabo un análisis de todos los tiempos y etapas involucrados en la obtención del reporte final. El proceso se inicia el primer día con la obtención de la muestra tumoral, seguido de la histopatología, donde se realizan cortes histológicos y se evalúa patológicamente la suficiencia del porcentaje tumoral, requiriendo entre 2 a 3 días. A continuación, se agenda una nueva cita con el paciente para obtener la muestra de sangre (muestra de control), lo cual agrega un día adicional al proceso. Las muestras se envían posteriormente a una empresa externa en China, añadiendo una espera de 1 semana solo por el envío.

En China, se llevan a cabo el Wetlab y el Drylab, donde se extrae el ADN de las muestras y se realiza la secuenciación y, por otro lado, se analizan los datos con softwares para identificar variantes genéticas. Estos dos procesos tienen una duración de 1 semana y media. Posteriormente, se suma el tiempo necesario para enviar el reporte a Chile y realizar la revisión de calidad de los reportes y los genes detectados.

Considerando que cada una de estas fases se realiza únicamente en días hábiles y bajo la suposición de una ejecución continua e ininterrumpida, el proceso en su totalidad toma las 4 semanas mencionadas anteriormente. No obstante, este lapso, en la mayoría de las ocasiones, se ve prolongado por contratiempos comunes, tales como retrasos en el envío, insuficiencia de la muestra tumoral, demoras causadas por días festivos, citas de pacientes que se aplazan, entre otros. Este aspecto tiene una importancia significativa, ya que, aunque sucede en pocas ocasiones, ha pasado que fallece un paciente antes de que se pueda establecer el tratamiento adecuado, a la par de situaciones donde los resultados de los informes carecen de coherencia, llevando a solicitar nuevas pruebas. Por ejemplo, un caso de cáncer de mama, y no encontrar genes relacionados con dicho cáncer como el BRCA1, generando incertidumbre y la necesidad de una nueva evaluación.

4. Objetivos

a. Objetivo general

Agilización y automatización en la generación de reportes con la detección de variantes genéticas, reduciendo el tiempo de generación en un 25%.

b. Objetivos específicos

1. Gestionar archivos de datos crudos provenientes de secuenciadores masivos en formato FASTQ, ya sean de plataforma Ilumina y/o MGI.
2. Construir y definir un proceso tecnológico que le genere la independencia a Genoma en la detección de variantes genéticas.
3. Crear un sistema de almacenamiento de información y que funcione como repositorio para la creación de reportes clínicos con las mutantes detectadas.
4. Implementar la automatización de procesos para la creación de reportes clínicos.
5. Reducir los costos asociados en la generación de reportes en a lo menos un 30%.

5. Estado del Arte

Se llevó a cabo una investigación para desarrollar este proyecto, explorando herramientas informáticas para la detección de variantes genéticas. La revisión de la literatura identificó dos categorías principales: las herramientas de Bioinformática Tradicionales y el uso de Deep Learning, cada una con enfoques y metodologías diversas.

Las herramientas de Bioinformática Tradicionales, establecidas y validadas en genética y genómica, se dividen en dos tipos. Primero, las herramientas de ensamblaje, que buscan construir un genoma a partir de secuencias de ADN, funcionando como un rompecabezas (Parres, 2022). En este enfoque, se trabajan con fragmentos cortos de secuencias de ADN para formar una secuencia más larga y completa. Segundo, las herramientas con enfoque Bayesiano, que mapean las lecturas secuenciadas sobre un genoma de referencia*, identificando variantes candidatas mediante algoritmos de alineamiento (Parres, 2022). Este método busca diferencias en la cantidad de bases nitrogenadas en lecturas confiables de ADN en comparación con un genoma de referencia.

Entre las herramientas más utilizadas para el ensamblaje genómico se encuentran ABySS, DNASTAR y Newbler, mientras que para el enfoque bayesiano en la detección de variantes genéticas destacan Samtools y GATK. Estas herramientas se emplean en pipelines informáticos, combinándolas para adaptarse a objetivos y requerimientos específicos. La flexibilidad de esta metodología permite obtener resultados distintos según la combinación de herramientas utilizada. Numerosos trabajos científicos exploran y comparan eficiencias de diversas combinaciones de herramientas, como se evidencia en casos como el siguiente.

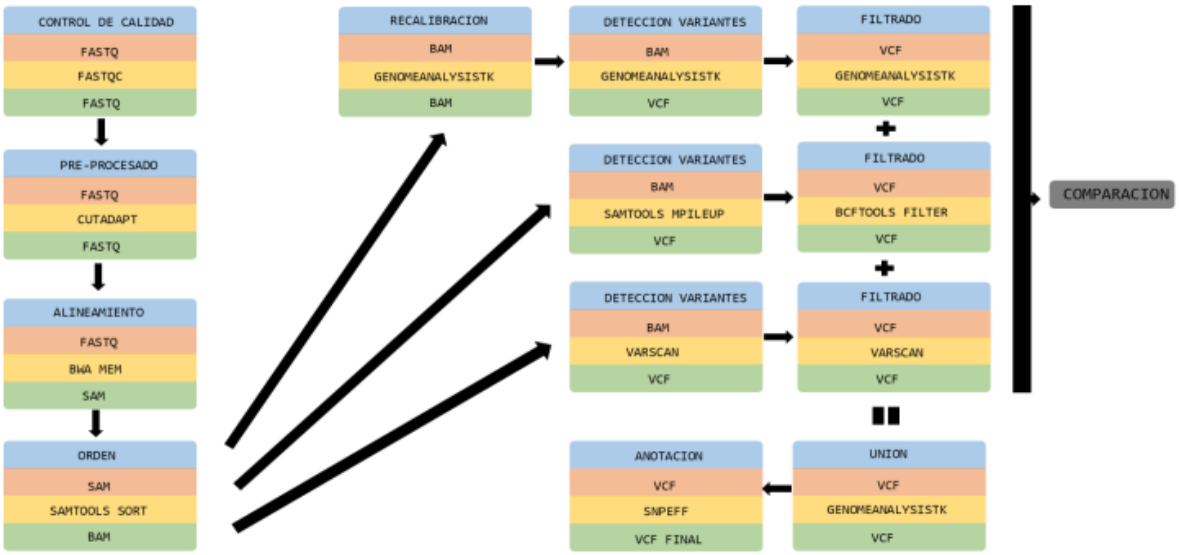


Figura 3: Esquema de distintas combinaciones de herramientas bioinformáticas tradicionales, donde las casillas azules indica el nombre del proceso, el marrón el nombre del archivo input, el amarillo el programa o herramienta bioinformática utilizado y en verde el archivo output. (Barquín,2017)

Aunque cada pipeline tiene un conjunto único de herramientas, generalmente siguen una serie de pasos comunes. Primero se realiza un control de calidad de los archivos FASTQ provenientes de secuenciadores masivos de nueva generación contieniendo información genética esencial, y su calidad es fundamental para el análisis del ADN. En segundo lugar, se lleva a cabo el preprocesamiento, que varía según el pipeline, pero tiene como objetivo principal transformar los archivos FASTQ en archivos BAM alineados a un genoma de referencia. Esta alineación es crucial para comparar las bases nitrogenadas con el "orden estándar" y detectar mutaciones. Luego, se ordena e indexa el archivo BAM para la detección de variantes, conocida como Variant Calling. Finalmente, se realiza un post procesamiento que difiere según el pipeline, con el propósito de filtrar las variantes encontradas, ya que en el ADN humano se encuentran miles de genes mutados, pero solo algunos son considerados patógenos. El resultado final es el archivo VCF, que contiene las variantes detectadas y analizadas.

Figura 4: Flujo de pasos estándar en un pipeline bioinformático para la detección de genes mutados (elaboración propia).

Después de revisar varios artículos científicos que abordan diversas combinaciones de herramientas, se resumen las más utilizadas en cada etapa en la tabla 1.

Etapa del Pipeline	Herramienta utilizada
Control de calidad	• FASTQC
Preprocesamiento	• Trimmomatic • PRINSEQ • Cutadapt • Picard
Mapeo y alinear	• BWA • Bowtie • MOSAIK • SHRiMP2 • novoalign
Ordenar	• Samtools
Variant Calling	• GATK HaplotypeCaller • Samtools • Varscan2 • Mutect2 • Consensus2 • MuSE • Consensus3 • SomaticSniper
Post procesamiento	• Samtools • Picard • GATK • SnpEff

Tabla 1: Tabla tipo resumen con distintas herramientas bioinformáticas. (Barquín, 2017), (Parres, 2022), (Mayordomo, 2023), (Koboldt, 2020) (Varela, 2019), (Carlos A, 2022).

La otra vía de exploración en la detección de variantes genéticas es el Deep Learning, siendo su primera propuesta el Deep Variant, diseñada para trabajar con datos provenientes de las NGS. Consta de dos pasos: "Primero busca variantes candidatas y codifica la información de dicha región en una imagen, denominada Pileup. En segundo paso, se encarga de introducir dicha imagen en un modelo neuronal llamado InceptionV2 que detecta si hay variante o no" (Parres, 2022).

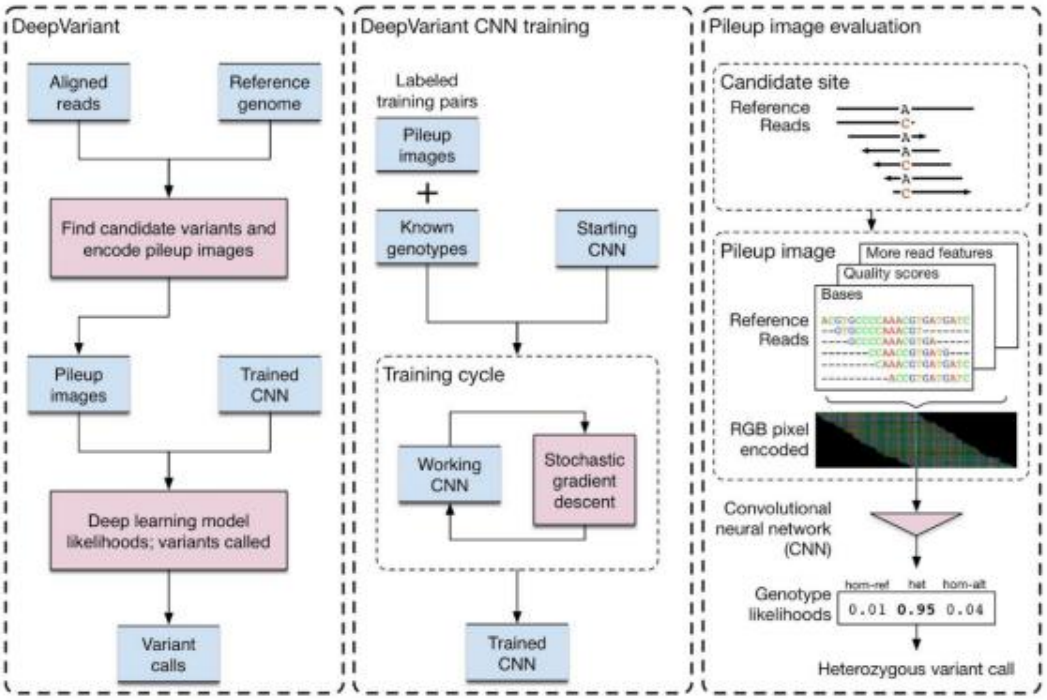


Figura 5: Flujo de trabajo de Deep Variant (Parres, 2022).

En todos los casos estudiados se vio que se utiliza el Deep Learning para transformar el problema de detección de variantes, en un problema de Computer Vision*. En este contexto se usan como datos los patrones en las secuencias de letras que conforman el ADN. De esta manera identificar patrones y variaciones en las secuencias de ADN, simplificando el proceso de análisis y detección de variantes genéticas.

Su funcionamiento parte de la misma forma que las herramientas tradicionales, con archivos FASTQ, y su alineamiento. Posteriormente se genera la conversión de los datos en Pileups*. Estas imágenes facilitan el análisis visual de las lecturas y la detección de variantes. Luego se entrenar el modelo de Deep Learning con datos previamente anotados para que el modelo aprenda reconocer patrones asociados a diferentes tipos de variantes genéticas. Finalmente se analizan los Pileups generados a partir de los datos reales. Así detecta las variantes genéticas de los datos, informándolos el archivo VCF.

En la actualidad, existen tres enfoques principales para el almacenamiento de información, cada uno con sus propias características distintivas. En primer lugar, las bases de datos relacionales (RDBMS) ofrecen un método tradicional y estructurado, destacándose por su capacidad de consulta sólida y la garantía de integridad referencial (López, 2016). En el ámbito de la genómica, donde los datos suelen tener una estructura bien definida, las RDBMS facilitan la organización de información genética en tablas con filas y columnas, simplificando la representación de relaciones genéticas y la

asociación de datos específicos con individuos, genes o variantes. Por otro lado, las bases de datos NoSQL proporcionan flexibilidad en el esquema, permitiendo manejar datos no estructurados y escalar horizontalmente para gestionar grandes volúmenes de información (Castillo, 2017). Este enfoque resulta beneficioso para conjuntos de datos genómicos masivos y altamente variables. Por último, las bases de datos en la nube ofrecen escalabilidad dinámica y accesibilidad global al permitir el despliegue de instancias en diversas ubicaciones geográficas. Esta combinación de flexibilidad, escalabilidad y accesibilidad las convierte en herramientas fundamentales para la gestión eficiente de datos genéticos en proyectos de investigación y aplicaciones clínicas a gran escala. La elección entre estas opciones dependerá de diversos factores, como la estructura de los datos, la complejidad de las relaciones genéticas y las necesidades específicas de la aplicación.

6. Solución escogida

En primer lugar, se descartaron las herramientas tradicionales basadas en ensamblaje debido a su lento tiempo de ejecución y alta demanda de recursos, ya que contradecían el objetivo principal del proyecto de agilizar la obtención de reportes.

Tras analizar Deep Variant y herramientas tradicionales con enfoque Bayesiano, se optó por estas últimas debido a la disponibilidad de recursos y documentación. Aunque Deep Variant ofrece capacidades avanzadas respaldadas por redes neuronales, su ejecución es costosa computacionalmente. Dada la limitación de recursos, se consideró más práctica la opción de herramientas tradicionales con enfoque Bayesiano.

Para el Pipeline y las herramientas, se eligió seguir las mejores prácticas de GATK. Es crucial destacar que se necesitan herramientas y pipelines distintos para el análisis de variantes somáticas* y germinales* debido a las diferencias inherentes en su análisis. Se aplicaron las mejores prácticas de GATK específicas para cada caso, programadas con Python, ya que GATK ha demostrado brindar excelentes resultados, especialmente en el contexto de genes cancerígenos. La amplia utilización de GATK en investigación genómica y aplicaciones clínicas proporciona una documentación sólida, lo que asegura confianza y eficacia para superar desafíos en el desarrollo del proyecto.

Se decidió crear una base de datos SQL para centralizar y gestionar la información generada por los pipelines. Esta elección facilita la generación de informes, ya que una base de datos SQL estructura la información de manera que permite extraer datos específicos mediante consultas, garantizando una obtención eficiente y precisa de la información.

Se optó por utilizar R con su librería R Markdown para generar informes, dada su flexibilidad y capacidad de personalización como también en la facilidad para integrar y orquestar herramientas, simplificando el flujo de trabajo.

Se consideró la opción de realizar el proceso en la nube, pero se descartó debido a las limitaciones de recursos en la empresa. Aunque se discutió la viabilidad con la empresa, se optó por llevar a cabo todo de forma local sin contratar servicios en la nube como AWS, aunque se pueda considerar esta opción en el futuro.

7. Metodología

Para alcanzar los objetivos del proyecto, se dividió en tres fases interrelacionadas: la primera, centrada en detectar variantes genéticas cancerígenas; seguida por una etapa de almacenamiento, que garantiza la integridad y accesibilidad de los datos derivados de la detección; y, finalmente, la generación y automatización de informes con las variantes identificadas.

Se implementó la metodología ágil Scrum para abordar eficientemente estas etapas, proporcionando flexibilidad para adaptarse a cambios, entregas incrementales y regulares para obtener retroalimentación temprana, y una comunicación continua con las partes interesadas. Con esta forma de trabajo, se desarrolla el pipeline informático en sprints cortos, probando de forma incremental su funcionalidad con cada script que se genere. También se diseñar y alimentar la base de datos en incrementos, asegura la integridad de los datos. Por último, se promueve la retroalimentación de la generación de los reportes, mejorando el proyecto de manera iterativa.

8. Medidas de desempeño

En relación con las medidas de desempeño, se destacan dos principales. En primer lugar, el KPI del porcentaje de reducción de tiempo en obtener el reporte final evalúa el éxito del objetivo global del proyecto como también los 4 primeros objetivos específicos, ya que cuantifica la eficacia del análisis, almacenamiento y automatización de la data. Se calcula comparando el tiempo actual para obtener el informe de variantes detectadas (*ti*) con el tiempo que tomaría sin utilizar el servicio de la empresa externa (*tf*).

$$\% \text{ Reducción tiempo} = ((ti - tf)/ ti) * 100$$

El KPI de porcentaje de reducción de costos evalúa la eficacia en la disminución de gastos, proporcionando una medida del beneficio económico del proyecto y confirmando el último objetivo

específico. Este indicador se calcula comparando el costo actual para obtener el informe de variantes detectadas (*ci*) con el costo asociado sin la utilización del servicio de la empresa externa (*cf*).

$$\% \text{ Reducción Costos} = ((ci - cf) / ci) * 100$$

9. Desarrollo y Plan de implementación

a. Fase 1 proyecto

Para el desarrollo de este proyecto se utilizó la data cruda de las máquinas Ilumina de 139 pacientes del COP de este año 2023. Para cada paciente, se generan dos archivos FASTQ* (Read 1 y Read 2) * para las secuencias tumorales y otros 2 FASTQ para las muestras de control.

Como se mencionó previamente, la primera fase del proyecto es el Pipeline, y su etapa inicial implica el formateo de los datos convirtiendo los archivos FASTQ en archivos BAM (Binary Alignment/Map), necesarios para las herramientas de detección de variantes. Para esto, lo primero es evaluar la calidad de las bases nitrogenadas en ambos pares de FASTQ con la herramienta FASTQC, visualizando la distribución de calidad a lo largo de las secuencias (figura 6). Estos datos informan sobre la confiabilidad de los datos, guiando las decisiones en el procesamiento subsiguiente. Un diagrama de flujo, adjunto en el anexo (Anexo 2), muestra toda la serie de pasos.

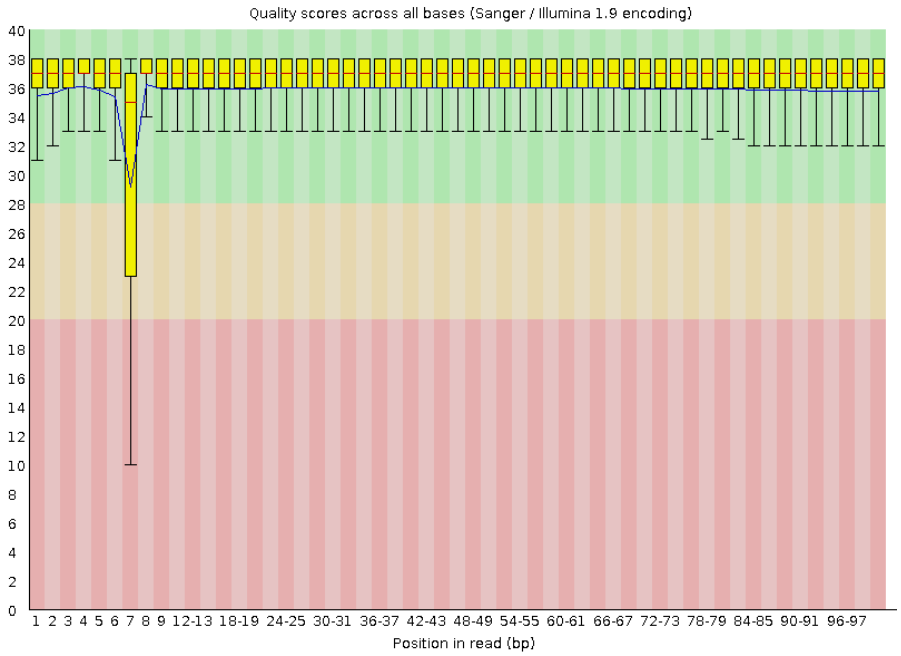


Figura 6: Grafico obtenido por la herramienta FASTQC. Representación de la calidad de las bases en cada posición de la secuencia (elaboración propia, extraído del pipeline).

Posteriormente para cada FASTQ es necesario cortar las bases de mala calidad, como también los adaptadores* que fueron añadidos durante la preparación de la biblioteca para la secuenciación. Para este proyecto se consideró un umbral de Q30 (quality score) para el corte, lo que se traduce en que cada base en la secuencia tiene una probabilidad del 99.9% de ser correcta. Se escogió este umbral porque es un estándar común para garantizar datos de alta calidad y es muy utilizado en estudios que requieren alta precisión, como es en el caso de la detección de variantes cancerígenas. Esto se realizó con la herramienta Trim Galore.

El siguiente paso, con la herramienta Bowtie2, es alinear estas secuencias al genoma de referencia. Esto se hace por separado para las muestras tumorales y de control, generando un archivo SAM (Sequence Alignment/Map) para cada caso. La alineación consiste en mapear las bases secuenciadas, emparejando las secuencias cortas de ADN con el genoma de referencia. Este proceso es crucial porque, por un lado, al alinear al genoma de referencia se puede saber la ubicación exacta de las bases que se están analizando y con ello saber de qué genes* estamos hablando, para, posteriormente, comparar el BAM tumoral con el de control para saber que bases difieren entre sí y, por ende, saber qué genes están mutados y así encontrar las variantes genéticas. Luego, se utilizan varias transformaciones de archivos con Samtools convirtiendo los archivos SAM a BAM.

La segunda etapa del pipeline corresponde el preprocesamiento de los datos. Una vez obtenido los archivos BAM, con la herramienta AddOrReplaceReadGroups, agrego información sobre grupos de lecturas, mejorando la organización y la interpretación de los datos de secuenciación. Seguidamente, es necesario la eliminación de los duplicados de lecturas*, para ello, primero son identificados con la herramienta MarkDuplicates y eliminados con Samtools. Este paso es importante, ya que la presencia de duplicados puede afectar negativamente la interpretación de los datos y la precisión de los resultados.

Por último, con BaseRecalibrator y ApplyBQSR se mejora la precisión de las llamadas de variantes. Estas herramientas están diseñadas para identificar los errores sistemáticos que pueden ocurrir durante la secuenciación y el mapeo, ajustando las puntuaciones de la calidad de los datos de secuenciación y mejorar la precisión de las llamadas de variantes.

La tercera Etapa del pipeline corresponde al llamado de variantes, realizado con Mutetc2. Primero se llama a las variantes en la muestra tumoral, y luego, se filtran las variantes basándose en la información de la muestra normal. Después de la llamada inicial de variantes, Mutect2 puede aplicar filtros basados en la presencia de las variantes en bases de datos conocidas. Este paso ayuda a reducir la cantidad de falsos positivos al eliminar variantes que son comunes en la población general y no

específicas del cáncer. De esta forma, desde los archivos BAM, se genera un archivo VCF (Variant Call Format) que contiene las variantes detectadas y un archivo F1R2* con información detallada sobre las frecuencias de las bases en las muestras tumorales y normales, así como la cantidad de veces que cada base se observa en las lecturas. Este archivo F1R2 ayuda a Mutect2 a ajustar su análisis considerando los errores específicos de secuenciación de cada muestra, mejorando así la precisión en la identificación de variantes genéticas y reduciendo los falsos positivos.

De forma paralela, es recomendable calcular la contaminación. Esto se hace al utilizar getPileupSummaries, con el cual se obtienen resúmenes de pilas* que proporcionan información importante sobre la cobertura*, la frecuencia de alelos* y otros parámetros que son esenciales para la llamada de variantes. Y, posteriormente se utiliza CalculateContamination, para eliminar la contaminación, que se refiere a la presencia de alelos de muestras ajenas en los datos de secuenciación.

Por último, se realiza el post procesamiento que consta de 3 pasos. Primero se aplica learnReadOrientationModel, en donde con el archivo F1R2 se aprende y utiliza un modelo para corregir sesgos en la orientación de las lecturas* de secuenciación, mejorando así la precisión en la detección de variantes genéticas. Luego se aplica FilterMutectCalls, el cual aplica filtros a las variantes detectadas por Mutect2, con la información obtenida de CalculateContamination y learnReadOrientationModel. En la etapa final del análisis, Funcotator genera un archivo MAF (Mutation Annotation Format) donde asigna funciones biológicas y evalúa el impacto genético y celular de las variantes detectadas. Utiliza información integrada de diversas bases de datos, como dbNSFP, RefSeq y COSMIC, para ofrecer una comprensión detallada de las consecuencias funcionales de las variantes en proteínas y elementos genómicos. Esta anotación funcional facilita la interpretación del significado biológico de las variantes. Se adjunta una tabla en el anexo (Anexo 4) que resume las herramientas utilizadas y sus funciones correspondientes para una mejor comprensión.

Algo importante a destacar es que el análisis de variantes somáticos y germinales es distinto principalmente por la naturaleza de la muestra. Para variantes somáticas, se analizan muestras de células tumorales específicas, mientras que para variantes germinales se requieren muestras de sangre, ya que estas variantes están presentes en todo el cuerpo y no en ubicaciones específicas como en las variantes somáticas. Es por ello que se realizó un pipeline aparte para la detección de variantes germinales. Un diagrama de flujo, adjunto en el anexo (Anexo 3), muestra toda su serie de pasos, lo cuales son muy parecidos al pipeline somático. Difieren únicamente en la fase de detección de variantes y en el post procesamiento. En lugar de utilizar Mutetc2, se utiliza HaplotypeCaller que

produce un archivo GVCF (genomic VCF) en vez de un VCF. En relación con el post procesamiento, se lleva a cabo en cuatro pasos. En el primero, mediante la herramienta ReblockGVCF, se realiza una compresión del archivo GVCF. En este proceso, se comprimen las regiones donde las variantes son homocigotas para la referencia, es decir, aquellas áreas en las que no hay variantes presentes. Esta compresión se efectúa de acuerdo con los nuevos parámetros de GQ (calidad del genotipo), los cuales son asignados automáticamente por HaplotypeCaller. Luego se aplica ValidateVariants, el cual se usa para verificar la coherencia y la validez del formato de las variantes en el archivo generado. Después es necesario aplicar GenotypeGVCFs para transformar archivos GVCF a VCF, permitiendo la genotipificación conjunta de múltiples muestras en lugar de hacerlo por separado. Este proceso tiene la ventaja de mejorar la precisión de las llamadas genotípicas al considerar la información de todas las muestras simultáneamente. Lo cual es útil en regiones con baja cobertura o en muestras con información limitada. Por último, se aplica funcotator, al igual que en el pipeline para la detección de variantes somáticas.

Toda la primera fase del proyecto se implementó completamente en el lenguaje de programación Python, desarrollando funciones específicas para cada una de las herramientas involucradas.

b. Fase 2 proyecto

La segunda fase del proyecto se enfocó en la creación de la base de datos que almacena toda la información generada por los pipelines. Este proceso incluyó el diseño conceptual, seguido por el diseño lógico y, finalmente, la implementación del diseño físico utilizando MySQL Workbench. Los esquemas completos de cada una de estas fases se encuentran detallados en el anexo (Anexo 5, Anexo 6 y Anexo 7).

La estructura de la base de datos se realizó con el objetivo principal de facilitar la extracción de información para la generación de informes. En el esquema resumen adjunto (figura 7), se visualiza la organización de los datos, dividida en 17 entidades distribuidas en tres secciones principales.

En primer lugar, la entidad "Paciente" contiene la información necesaria para el análisis médico, como la edad, peso, hábitos, etc. Esta entidad se conecta con entidades clave como "Médico", "Antecedentes Familiares" y "Exámenes Genoma", que albergan información relevante para la empresa. Además, se optó por crear una entidad independiente, "Datos Paciente", que almacena la información privada del paciente, como nombre, Rut, teléfono, etc. Esta separación se diseñó considerando la confidencialidad del paciente, permitiendo el acceso de terceros a información relevante para ensayos y estudios clínicos.

En segundo lugar, la entidad "Variante" almacena las variantes encontradas junto con sus atributos correspondientes. Esta entidad se relaciona con "Gen", ya que cada variante está asociada a un gen mutado. Además, tanto "Variante" como "Gen" establecen conexiones con entidades que contienen información de otras bases de datos, ampliando la posibilidad de establecer conexiones adicionales, tales como "HGNC", "CGC", "ClinVar", "Gnomad Exome", "Gnomad Genome" y "dbSNP".

Por último, la entidad "Secuencia" sirve como una entidad conectora entre la información del paciente, sus variantes genéticas y los detalles de la secuenciación de las muestras control y tumoral. Este diseño integral proporciona una estructura robusta para gestionar y analizar la información compleja generada por los pipelines.

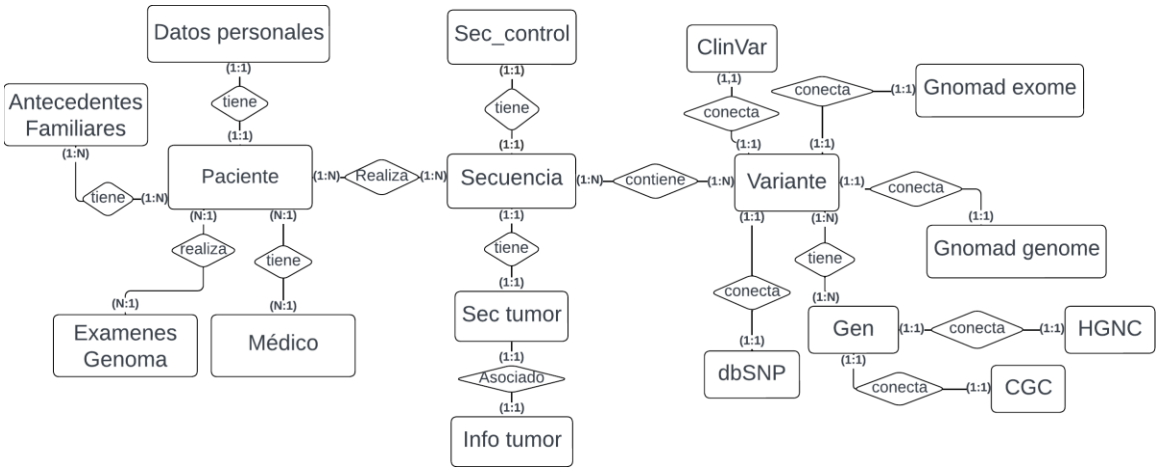


Figura 7: Resumen de diseño conceptual de la base de datos, representando las entidades (elaboración propia).

Respecto al almacenamiento de datos, se implementó un script en Python que se ejecuta automáticamente tras la detección de variantes genéticas, una vez ingresados los archivos iniciales en la fase 1 del proyecto.

c. Fase 3 proyecto

La fase final del proyecto consistió en la automatización de informes, implementada a través de R y la librería R Markdown. En esta etapa se diseñó un script en R que extrae la información directamente de la base de datos, que luego deja los parámetros importantes en el encabezado YAML, en la sección de 'params' que luego se utilizan en el cuerpo del documento Markdown, con el cual se creó una planilla del reporte a generar.

En el diagrama que representa la estructura de la solución (figura 8), se observa claramente que el proceso de implementación del proyecto se desarrolló en tres fases interconectadas que, al unirlas, una vez que los archivos FASTQ son ingresados, se logra de forma automática la ejecución del análisis de las variantes, el almacenamiento de la información resultante y la generación del informe

final. Este enfoque elimina la necesidad de realizar acciones adicionales, creando así un flujo de los datos y sin interrupciones en la obtención de resultados.

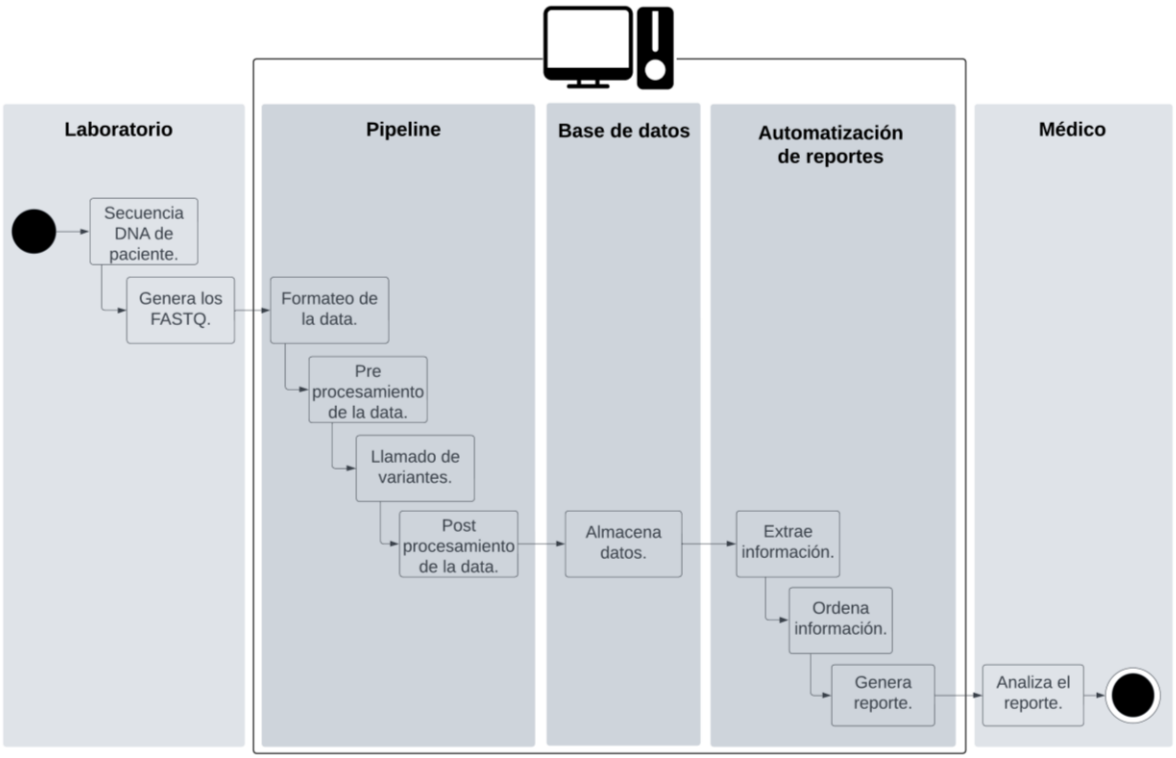


Figura 8: Diagrama de estructura de solución. (Elaboración propia)

En cuanto al código, se implementó un sistema compuesto por siete scripts que cubren diversas etapas. A continuación, se detalla cada script y su función dentro del flujo de trabajo:

- Alimentar_BBDD.py: Ingresa la información del paciente y genera IDs necesarios para su almacenamiento en la base de datos.
- PL_somático.py: Detecta variantes somáticas.
- PL_germinal.py: Detecta variantes germinales.
- Extraer_info.py: Extrae información relevante de los archivos MAF y VCF generados por los scripts PL_somático.py y PL_germinal.py.
- Conexion_BBDD.R: Extrae información almacenada en la base de datos.
- Reporte.Rmd: Genera un informe utilizando la información proporcionada.

Estos seis scripts están orquestados por el script principal main.py. Para una mejor comprensión del flujo de trabajo, se incluye en el Anexo 9 un diagrama de actividades que ilustra cada acción, desde el ingreso del paciente hasta la generación del informe. Además, en el anexo se

encuentra la Carta Gantt (Anexo 1) que muestra de forma detalla cómo se implementó el proyecto en el transcurso de 4 meses y medio.

10. Análisis de riesgo

Riesgo	Probabilidad	Impacto	Estrategia de mitigación
Falta de recursos	Alta	Alto	Solicitar más recursos si es necesario. Dar prioridad a las tareas más importantes y utilizar los recursos disponibles de manera eficiente.
Poca coincidencia con las variantes de referencia	Mediana	Alto	Implementar un proceso de revisión continua que incluya la comparación con las bases de datos genéticas más recientes.
Fallos en los pipelines	Baja	Alto	Revisar cada herramienta/función por separado. Establecer puntos de control en el pipeline
Problemas en la base de datos	Baja	Mediana	Realizar copias de seguridad periódicas. Revisar de forma detallada los diseños conceptual y lógico viendo si no hay problemas de normalización o de ciclos.
interrupciones inesperadas	Mediana	Baja	Generar puntos de revisión en el pipeline para detectar el error.
Falta de tiempo	Mediana	Alta	Priorizar tareas críticas y gestionar las responsabilidades de acuerdo con las habilidades y la carga de trabajo individual

Tabla 2: Matriz de riesgo (elaboración propia)

11. Evaluación Económica

En cuanto a la evaluación económica, por un lado, la inversión inicial es de 0 pesos chilenos, ya que no se necesitó adquirir nuevos recursos. Genoma Mayor ya contaba con el equipo informático necesario y la infraestructura del laboratorio para la secuenciación genética, eliminando la necesidad de inversiones iniciales.

Hasta el momento, Genoma Mayor había subcontratado la generación de informes a un costo de 1,3 millones de pesos por paciente, lo cual incluye los gastos asociados al wetlab* y al drylab*. Considerando que, en promedio, Genoma Mayor cuenta con 18 pacientes cancerígenos al mes, se traduce en aproximadamente 23,4 millones de pesos gastados anualmente únicamente en la generación de reportes. Que se reducirá en un 50%, debido a que secuenciar de forma local el DNA, tiene un costo aproximado de 650 mil, lo que se traduciría en un gasto anual de 11,7 millones de pesos, por el costo de los implementos del laboratorio.

A continuación, se realizó un flujo de caja de aquí a 3 años con una tasa de descuento de 9,5% (tasa de descuento hasta la fecha 6/7/2023), y tomando en consideración que Genoma cobra 2,4 millones de pesos por hacer estos exámenes. Los ingresos se estiman en 43,200,000, basados en el supuesto de que el número promedio de pacientes por mes no ha cambiado. Con la siguiente formula se calculó el Valor presente.

$$VPt = \frac{CFt}{(1+R)^t}$$

$$CF_t = Ingresos_t - Egresos_t$$

Año	Ingresos	Egresos	Flujo de Efectivo Neto	Valor Presente (a 9,5%)
0	-	-	-	-
1	43.200.000	- 11.700.000	31.500.000	28.767.123,3
2	43.200.000	- 11.700.000	31.500.000	26.271.345,5
3	43.200.000	- 11.700.000	31.500.000	23.992.096,3

Tabla 3: Flujo de caja. Todo en pesos chilenos).

$$VAN = \sum_{t=0}^{T} \frac{CFt}{(1+r)^t} - I$$

Con la información de la tabla 3 y utilizando la fórmula anterior, se calculó el Valor Actual Neto (VAN), obteniendo un valor positivo de 79,030,565.1. Este resultado sugiere que el proyecto tiene el potencial de generar un beneficio neto aproximado de 70,000,000 en términos de valor presente. En resumen, la evaluación económica respalda la viabilidad y rentabilidad de la implementación de la solución seleccionada.

12. Resultados

Después de completar el proyecto, se evaluó la sensibilidad en la detección de variantes genéticas. Este análisis es fundamental ya que una sensibilidad alta refleja la efectividad de la metodología empleada, indicando la proporción de variantes identificadas correctamente en comparación con las detectadas por la empresa externa. Para ello, se compararon ambos resultados, analizando 31 de los 152 pacientes del 2023, cuyos resultados se detallan en la tabla 4. En la fórmula utilizada, los Verdaderos Positivos (VP) son variantes detectadas y coincidentes con las identificadas por la empresa externa, mientras que los Falsos Negativos (FN) son variantes no identificadas a pesar de ser detectadas por dicha empresa. Este análisis proporciona información crucial sobre la fiabilidad del proceso de identificación de variantes.

$$Sensibilidad = \frac{VP}{(VP + FN)}$$

Tras calcular la sensibilidad a cada uno de los pacientes se llegó a una sensibilidad promedio de 0,87. Esto indica que el método de identificación de variantes ha logrado detectar correctamente el 87% de las variantes que la empresa externa identificó.

Paciente	N.º Variantes detectadas por empresa externa	N.º Variantes detectadas por pipeline generado	VP	FN	Sensibilidad
1	27	29	24	3	0,89
2	15	18	12	2	0,86
3	28	34	25	3	0,89
4	18	26	16	2	0,89
5	25	29	22	3	0,88
6	32	37	27	5	0,84
7	20	22	19	1	0,95
8	11	17	9	2	0,82
9	19	23	16	3	0,84
10	18	19	15	3	0,83
11	27	33	25	2	0,93
12	24	24	21	3	0,88
13	29	36	24	5	0,83
14	30	33	27	3	0,9
15	34	38	30	4	0,88
16	21	28	20	1	0,95
17	11	16	9	2	0,82
18	15	17	13	2	0,87
19	18	21	14	4	0,78
20	25	27	20	5	0,8
21	16	19	12	4	0,75
22	28	29	23	5	0,82
23	21	26	19	2	0,9
24	23	27	18	5	0,78
25	22	26	19	3	0,86
26	27	30	25	2	0,93
27	31	33	29	2	0,94
28	27	36	25	2	0,93
29	14	17	11	3	0,79
30	28	31	27	1	0,96
31	27	33	25	2	0,93

Tabla 4: Tabla comparativa de resultados en la detección de variantes genéticas.

Por otro lado, el porcentaje de reducción de costos pasó de ser 23,4 millones de pesos anualmente a 11,7 millones de pesos anualmente, lo que se traduce en un porcentaje de reducción de 50%.

*%Reducción de costo= ((Costo inicial – Costo Actual) / Costo inicial) * 100*

*%Reducción de costo= ((23.400.000 – 11.700.000) / 23.400.000) * 100* = 50%

En cuanto al tiempo, antes de implementar el proyecto todo el proceso de obtención de reportes pasó de ser de 23 días hábiles a 12 días hábiles de forma redondeada, lo cual se ve en la tabla 5. Esto se traduce en pasar de 31 días (contando fin de semanas) a 16 días en total. Lo cual significa un porcentaje de reducción de 48%.

*% Reducción de tiempo = ((Tiempo Inicial – Tiempo Actual) / Tiempo Inicial) * 100*

*% Reducción de tiempo = ((31 días – 16 días) / 31 días) * 100* ≈ 48%

Etapa	Tiempo antes Proyecto (en días hábiles)	Tiempo después Proyecto (en días hábiles)
Toma muestra tumoral	1 día	1 día
histopatología	3 días	3 días
Toma muestra de sangre	1 día	1 día
Envío a China	7 días	0 días
Wet lab y dry lab	10 días (en conjunto)	6 días (Wet lab) 11,5 horas (Dry lab)
Revisión de calidad	1 día	1 día
Tiempo total en días hábiles	23 días	11 días y 11,5 horas

Tabla 5: tabla comparativa de los tiempos antes y después de implementar el proyecto.

En el anexo (Anexo 8) se encuentra una tabla con cada una de las etapas del pipeline con su respectivo tiempo promedio, en donde se ve de forma detallada como se llegó a las 11,5 horas en la detección de variantes genéticas.

Respecto a la base de datos se logró generar una base de datos que integra toda la información proporcionada por los pipelines como también información extra relevante para la empresa, con la cual se puede saber de forma rápida consultas sobre los datos que contiene y poder responder preguntas como; ¿ Cuáles son todos los paciente que tienen X gen variado?, ¿Cuál es el gen que más se repite entre todos los pacientes del presente año?, ¿Cuántas variantes genéticas son encontradas en promedio en pacientes con cáncer gastrointestinal?, etc. Además de generar querys complejas, la base de datos también cumplió su objetivo de ser un repositorio de información para la generación de reportes.

13. Discusión

Tras los resultados expuestos, el pipeline, a pesar de tener buenos resultados, con una especificidad de un 87%, hay aspectos a mejorar, ya que, en el contexto en el que está sumergido el proyecto, es crucial poder detectar el 100% de las variantes genéticas. Algo a destacar es que, a lo largo del proyecto, se ha incrementado progresivamente la sensibilidad. Al principio, se empezó con un 21%, lo cual es muy bajo, pero al hacer cambios y pruebas constantes, como ajustar el trimming a un valor de calidad mayor o conectar a diferentes bases de datos, ha mejorado los resultados. Es por eso que hay que seguir encontrando mejoras para subir el porcentaje de sensibilidad. De igual manera, esto es algo que constantemente va evolucionando.

Un aspecto que no se pudo evaluar, debido a la falta de datos, es la especificidad. En lugar de medir la proporción de casos positivos reales que son identificados correctamente por el pipeline, la especificidad se centra en la proporción de casos negativos reales que son identificados correctamente. Esta métrica tiene gran importancia, ya que refleja la eficacia del pipeline para descartar genes normales sin generar "falsas alarmas". Lamentablemente, no se pudo calcular este parámetro debido a la ausencia de los falsos positivos proporcionados por la empresa externa, los cuales corresponden a todas las variantes descartadas que no están disponibles para nuestro análisis.

En términos de costos, una vez que la implementación de mi proyecto se complete, se alcanzará exactamente el objetivo establecido en la evaluación económica: una reducción del 50% en los costos asociados a la generación del informe. En cuanto al tiempo, no solo se logró el objetivo general del proyecto, sino que también se superó casi al doble, con una reducción del tiempo en un 48%.

Un aspecto a destacar es que la implementación del proyecto se llevó a cabo de manera local. Esta decisión fue discutida con la empresa antes de iniciar el proceso, durante la cual se les recomendó considerar el desarrollo del proyecto en la nube. Sin embargo, para adoptar esta sugerencia, sería necesario contratar una plataforma de servicios en la nube, una medida a la que Genoma no estaba dispuesto a comprometerse en ese momento. A pesar de esta elección, desde mi perspectiva, a corto o mediano plazo, la transición de la implementación local a la nube sigue siendo recomendado debido a consideraciones de escalabilidad y espacio. Es importante señalar que la empresa ya tiene conocimiento de esto y planea realizarlo.

El último tema para abordar es el perfeccionamiento de la planilla de los reportes. A pesar de que la conexión entre la base de datos y la generación de informes fue exitosa, Falta perfeccionar la planilla en donde las variables iteran, la cual ha estado en un proceso de iterativos feedback.

14. Conclusión

En resumen, se logró con éxito agilizar y automatizar la generación de informes con detección de variantes genéticas, reduciendo el tiempo de generación en más del 25%, superando la meta en un 23%. Además, se hizo un buen manejo de archivos FASTQ, que incluyó el desarrollo de un pipeline informático y la creación de una base de datos y la implementación de la automatización de informes, lo que redujo el 50 % en los costos asociados a este proceso. Este logro, combinado con la revisión final de la planilla y las mejoras continuas en el pipeline, facilita que Genoma Mayor logre la independencia de la empresa externa, obteniendo así su sistema de generación de informes para detectar variantes genéticas. Una vez logrado lo anterior, se espera abordar nuevos desafíos, como lo ya mencionado, el traspaso a la nube, en donde se espera traspasar la base de datos como también el almacenamiento de los archivos y la orquestación de los distintos componentes.

Una vez alcanzado este hito, el siguiente desafío será la transición a la nube, que implica trasladar la base de datos, el almacenamiento de archivos y la orquestación de los diferentes componentes aprovechando así la escalabilidad, accesibilidad y eficiencia de este tipo de plataforma. Además, a futuro, será necesario implementar en cada etapa del proceso de extracción de información y establecer capas de seguridad para garantizar la seguridad y confidencialidad de los datos de los pacientes.

15. Bibliografía

(1) Parres, D. (2022). *Detección de Variantes Genómicas utilizando Deep Learning*. Universidad Politécnica de València. 1-55 https://riunet.upv.es/bitstream/handle/10251/185358/Parres%20-%20Deteccion%20de%20Variantes%20Genomicas%20utilizando%20Deep%20Learning.pdf?sequence=1

(2) Barquín, M. (2017). *Desarrollo de un protocolo de análisis de datos de NGS y comparación de algoritmos de detección de variantes.* Universidad Oberta de Catalunya. 1-35 https://openaccess.uoc.edu/bitstream/10609/74385/6/mbarquinTFM0117memoria.pdf

(3) Mayordomo, A. (2023). *Identificación de variante genética causal para síndromes de cáncer colorrectal hereditario: secuenciación masiva en paralelo y aplicación de herramientas bioinformáticas*. Facultad de ciencias agrarias Universidad Nacional de Rosario, 1-69. http://biblioteca.puntoedu.edu.ar/bitstream/handle/2133/26301/MAYORDOMO%2C%20Constanza%20-%20TESIS%20.pdf?sequence=3&isAllowed=y

(4) Koboldt, D. (2020). *Best practices for variant calling in clinical sequencing*. Genome Medicine. 1-13. https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00791-w

(5) Varela, D. (2019). *Diseño e implementación de un flujo de trabajo bioinformático en la nube para la identificación de variantes oncogénicas a partir de datos genómicos*. Universidad Eia Ingeniería Biomédica Envigado. 1-59. https://repository.eia.edu.co/server/api/core/bitstreams/db4c7ffc-b465-448e-9c89-641bc3d91436/content

(6) Carlos A Garcia-Prieto, Francisco Martínez-Jiménez, Alfonso Valencia, Eduard Porta-Pardo, Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools, *Bioinformatics*, Volume 38, Issue 12, June 2022, Pages 3181–3191, https://doi.org/10.1093/bioinformatics/btac306

(7) National Human Genome Research Institute. (2023). genome.gov https://www.genome.gov/es/genetics-glossary/ACGT

(8) Broadinstitute.org. Recuperado el 19 de octubre de 2023, de https://gatk.broadinstitute.org/hc/en-us/articles/360035894731-Somatic-short-variant-discovery-SNVs-Indels-

(9) Broadinstitute.org. Recuperado el 19 de octubre de 2023, de https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

(10) Gracia del Busto, I. H., & Yanes Enríquez, I. O. (2013). BASES DE DATOS NoSQL. *Telemática*, *11*(3), 21–33. Retrieved from https://revistatelematica.cujae.edu.cu/index.php/tele/article/view/74

(11) Castillo, J. N., Garcés, J. R., Navas, M. P., Jácome Segovia, D. F., & Armas Naranjo, J. E. (2017). Base de Datos NoSQL: MongoDB vs. Cassandra en operaciones CRUD (Create, Read, Update, Delete). *Revista Publicando*, *4*(11(1), 79-107. Recuperado a partir de https://revistapublicando.org/revista/index.php/crv/article/view/398

(12) López Herrera, P. (2016). Comparación del desempeño de los Sistemas Gestores de Bases de Datos MySQL y PostgreSQL. https://core.ac.uk/download/pdf/80528621.pdf

16. Anexo

a. Contexto Conceptual

Hoy en día, una innovadora tecnología ha transformado el campo de las ciencias biológicas: la Secuenciación de Nueva Generación (NGS). Esta herramienta nos permite analizar el DNA y leer el código genético de un organismo de manera rápida y precisa. El DNA está compuesto principalmente por la combinación de 4 moléculas fundamentales, conocidas como bases nitrogenadas: A (adenina), G (guanina), C (citosina) y T (timina). Una secuencia específica de estas bases forma lo que conocemos como un gen. Al secuenciar el ADN, lo que hacemos, en términos sencillos, es descifrar el orden en la que están las bases nitrogenadas (ver figura 1), lo cual es esencial porque si este orden cambia se traduce en cambios en las funciones y características del organismo.

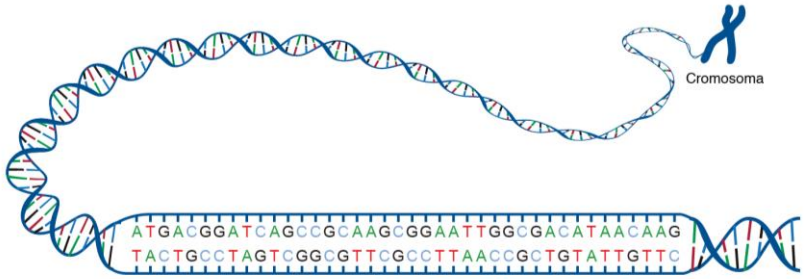


Figura 1: Representación gráfica de las bases nitrogenadas. (National Human Genome Research Institute, 2023)

Por otro lado, el cáncer surge de mutaciones y desórdenes en estas bases nitrogenadas, perturbando el funcionamiento normal de las células. Estas mutaciones/ variantes genéticas pueden ser identificadas comparando la secuencia de ADN con la de referencia, la cual es conocida como el genoma de referencia humano. Este genoma representa una secuencia completa y detallada del ADN humano, considerado como el patrón estándar que se espera que todos los seres humanos posean en cuanto a la disposición de sus bases nitrogenadas. Por esta razón, se utiliza como punto de referencia en investigaciones y estudios genéticos.

Las mutaciones pueden verse reflejadas de varias formas. Pueden consistir en variaciones de un solo nucleótido o puntuales (SNVs), que corresponden cuando se altera un solo nucleótido (A, T, G o C). También pueden ser variantes de inserción-eliminación (indels) que ocurre cuando uno o más pares de bases están o no están en el genoma. Otra forma es una variante de inversión, que es cuando se invierte el orden de las bases y por último pueden ser variantes del número de copias (CNVs). Todos estos tipos de variantes se pueden observar mejor en la figura 2 donde se ve cómo se compara el fragmento de un DNA con uno de referencia.

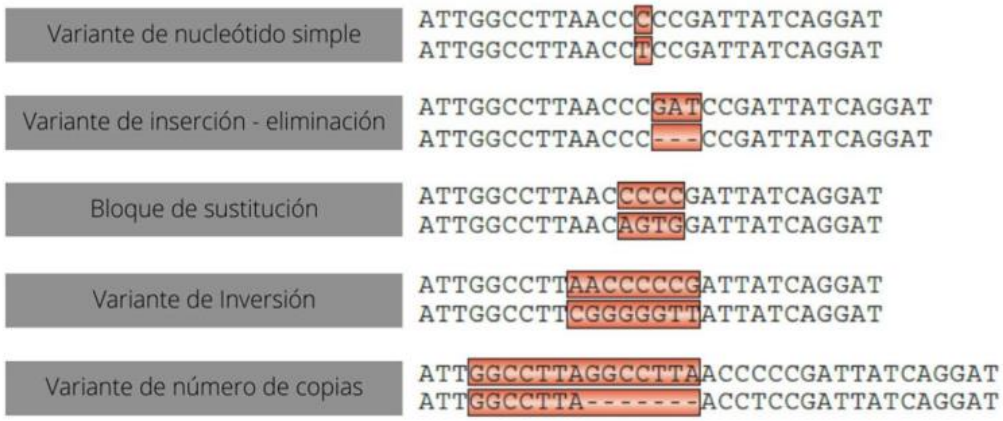
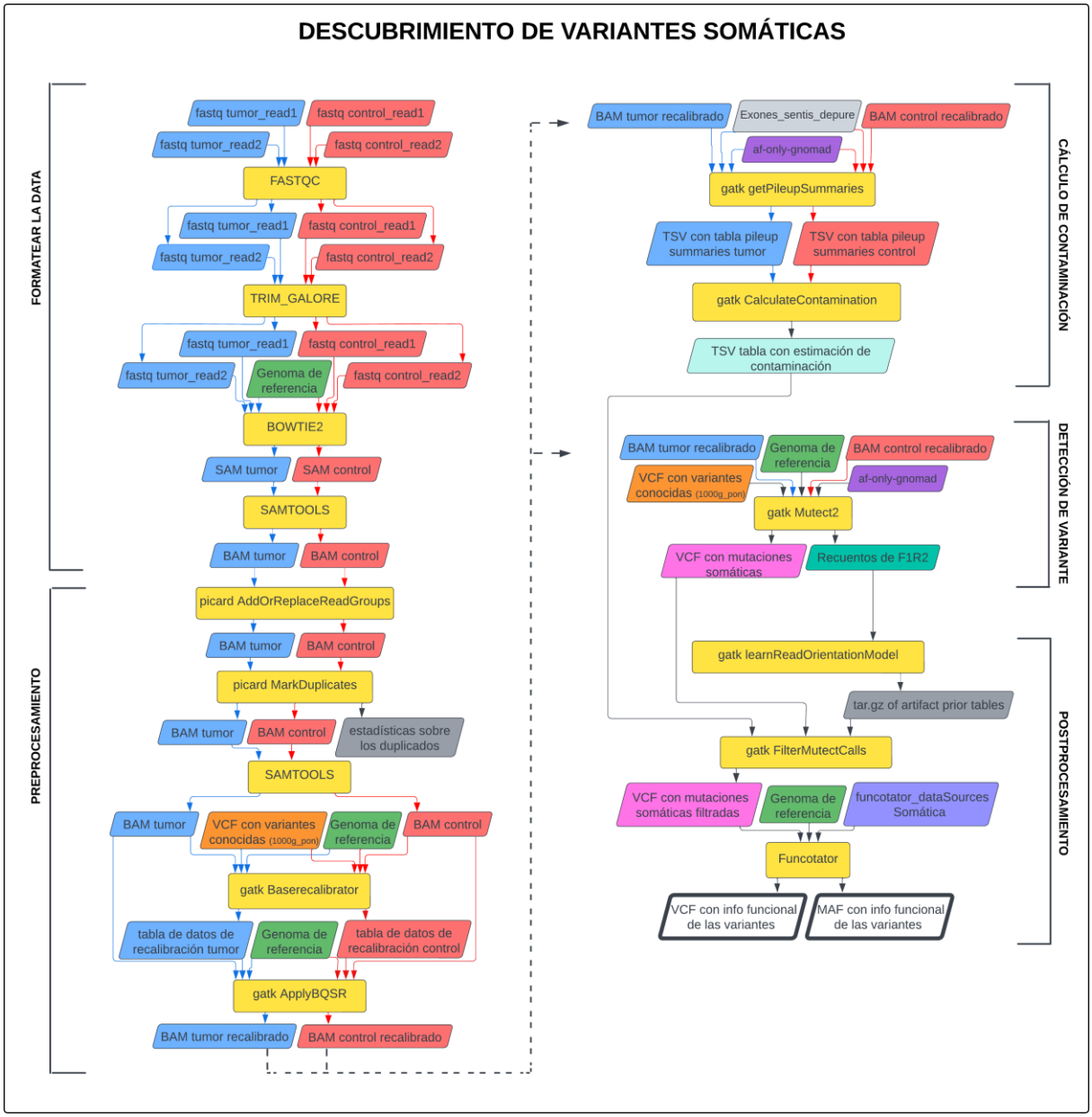


Figura 2: Distintos tipos de variaciones genéticas (Mayordomo,2023).

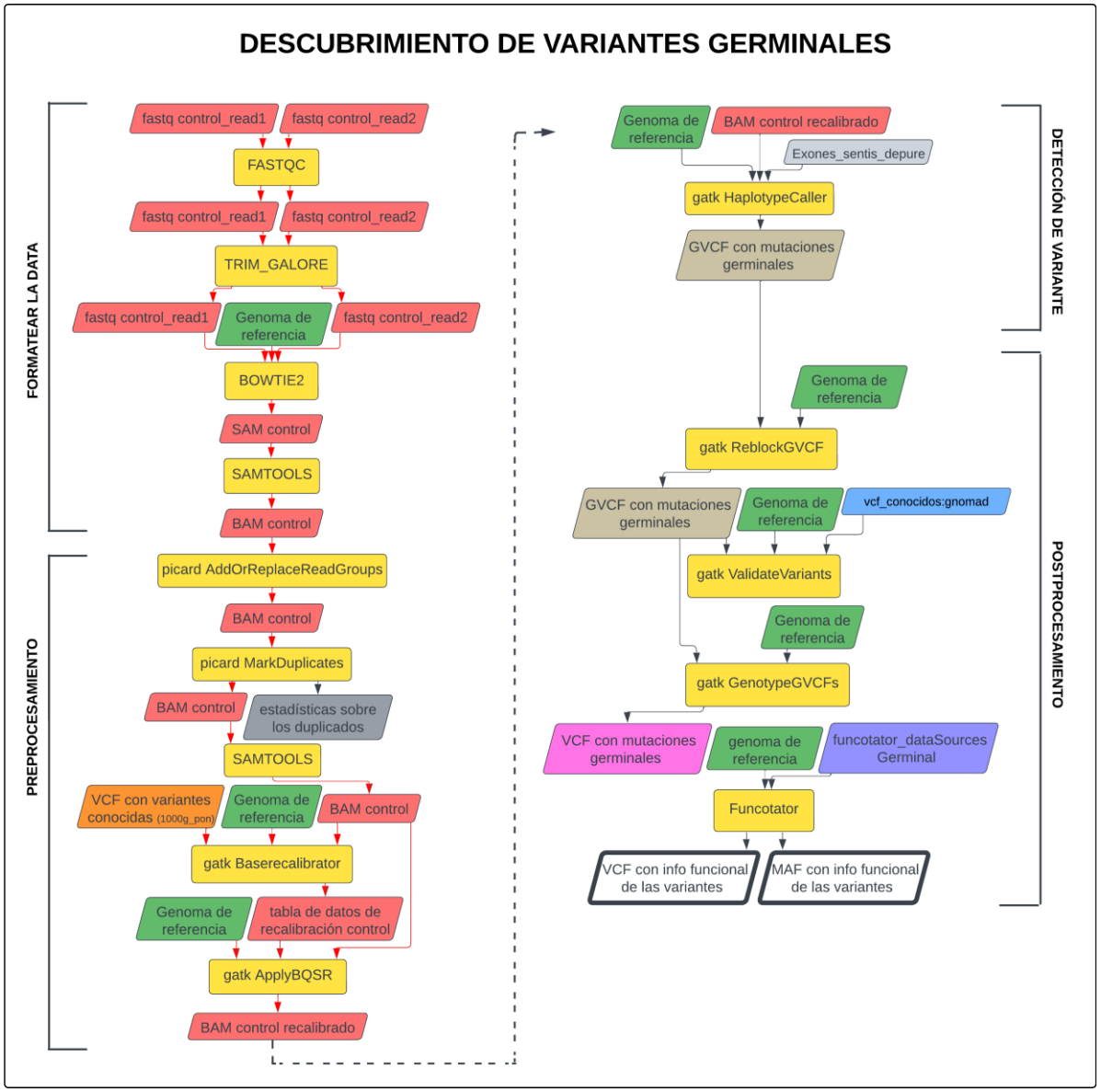
Es importante considerar estos tipos de variaciones porque se han descrito como causantes de la aparición de tumores. Es por ello que la correcta identificación se ha convertido en un punto importante para el diagnóstico y tratamiento de pacientes con cáncer. Además, hay que destacar la relevancia de la identificación especifica de los genes mutados, ya que esta información permite la selección de tratamientos altamente personalizados que se ajustan específicamente a las condiciones genéticas de cada paciente, ya que a veces, por tener ciertos genes específicos mutados se sabe que algunos tratamientos no hacen efecto. Por lo tanto, este tipo de procedimiento no solo se traduce en garantía de que el tratamiento seleccionado será el más idóneo para la situación, sino que también tiene una implicación financiera considerable por los costos de los tratamientos oncológicos.

Actividad	Mes																					
	Jul				Ago				Sep				Oct				Nov				Dic	
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2
Capacitaciones																						
Descarga de programas																						
TMB (Junta de comité)																						
Realización de flujo del pipeline																						
Creación pipeline variantes Somáticas																						
- Formateo de data																						
- Preprocesamiento de la data																						
- Detección variantes Somáticas																						
- Cálculo de contaminantes																						
- Post Procesamiento de la data																						
Creación pipeline variantes Germinales																						
- Formateo de data																						
- Preprocesamiento de la data																						
- Detección variantes Germinales																						
- Post Procesamiento de la data																						
Automatización de pipelines																						
Presentación 1/ Informe 1																						
Diseño base de datos																						
- Diseño conceptual																						
- Diseño lógico																						
- Diseño físico																						
Generación base de datos																						
Presentación 2/ Informe 2																						
Almacenamiento de los datos																						
Generación de Reporte																						
Presentación final/ Informe Final																						

Anexo 1: Carta Gantt (Elaboración propia).



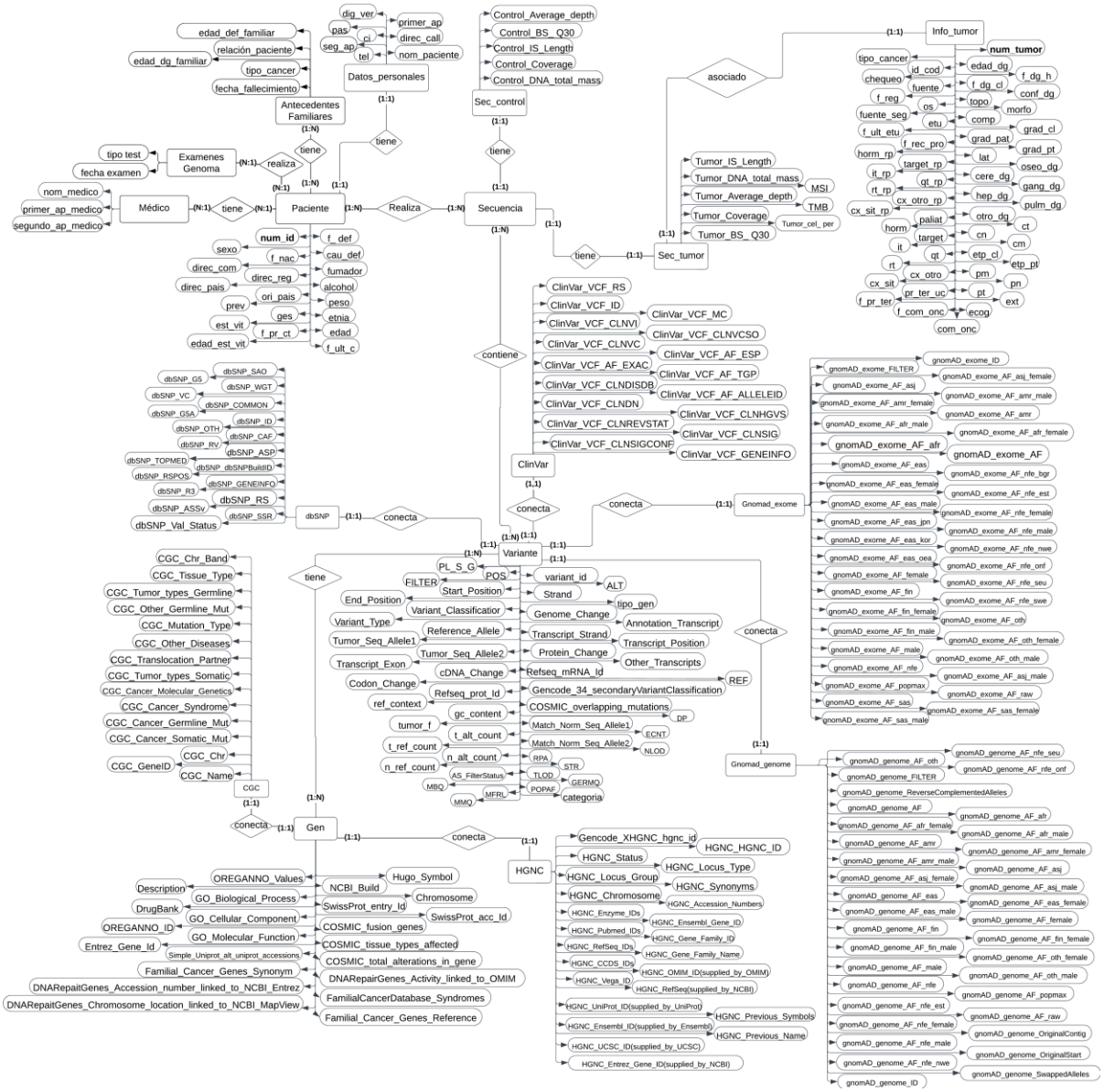
Anexo 2: Flujo de Archivos y herramientas utilizadas en Pipeline para la detección de variantes somáticas (elaboración propia). Los rectángulos amarillos corresponden a las distintas herramientas, mientras que los rectángulos que apuntas hacia ellas o desde ellas son los archivos de entrada y salida.



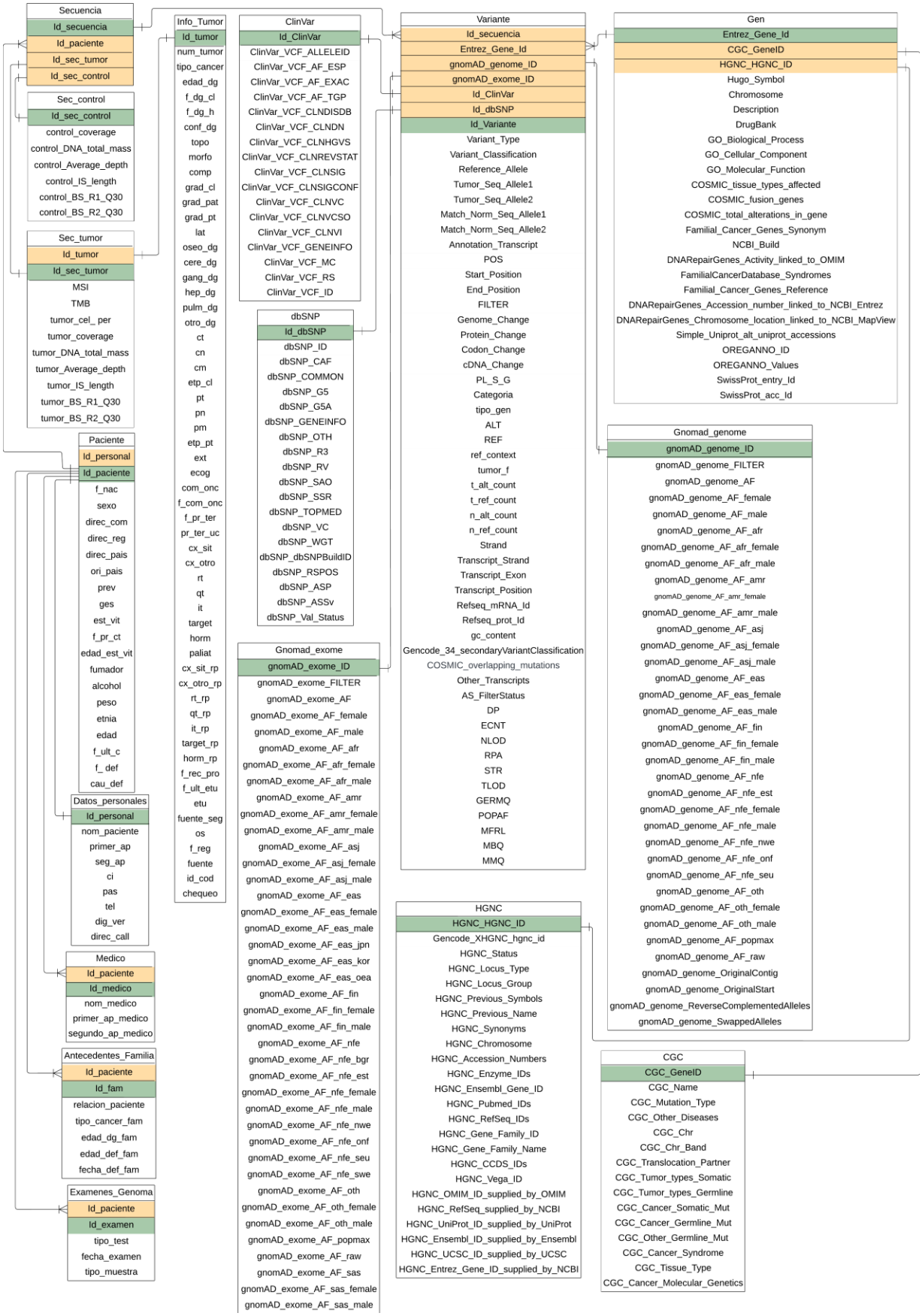
Anexo 3: Flujo de Archivos y herramientas utilizadas en Pipeline para la detección de variantes germinales (elaboración propia). Los rectángulos amarillos corresponden a las distintas herramientas, mientras que los rectángulos que apuntas hacia ellas o desde ellas son los archivos de entrada y salida.

Etapa	Herramienta	Función
Formateo de la data	*FASTQC*	Evalúa la calidad de las bases nitrogenadas.
	Trim Galore	Cortar las bases de mala calidad, como también los adaptadores.
	Bowtie2	Alinea secuencias al genoma de referencia.
	Samtools	Conversión de los archivos SAM a BAM.
Preprocesamiento	AddOrReplaceReadGroups	Agrega información sobre grupos de lecturas.
	MarkDuplicates	Identifica los duplicados de lecturas.
	Samtools	Elimina los duplicados de lecturas.
	BaseRecalibrator	Identifica y cuantifica errores sistemáticos en las puntuaciones de calidad.
	ApplyBQSR	Ajusta las puntuaciones de calidad.
Detección Mutantes	Mutect2	Detecta las variantes genéticas somáticas.
	HaplotypeCaller	Detecta las variantes genéticas germinales.
	getPileupSummaries	Proporciona resúmenes detallados de la cobertura y la frecuencia de alelos en sitios específicos del genoma a partir de pilas de lecturas.
	CalculateContamination	Estima y corrige la tasa de contaminación.
Post procesamiento	LearnReadOrientationMode	Corrige sesgos en la orientación de las lecturas de secuenciación.
	FilterMutectCalls	Aplica filtros a las variantes detectadas
	ReblockGVCF	Compresión del archivo GVCF.
	ValidateVariants	Verificar la coherencia y la validez del archivo GVCF.
	GenotypeGVCFs	Transforma archivos GVCF a VCF.
	Funcotator	Asigna funciones biológicas y evalúa el impacto genético.

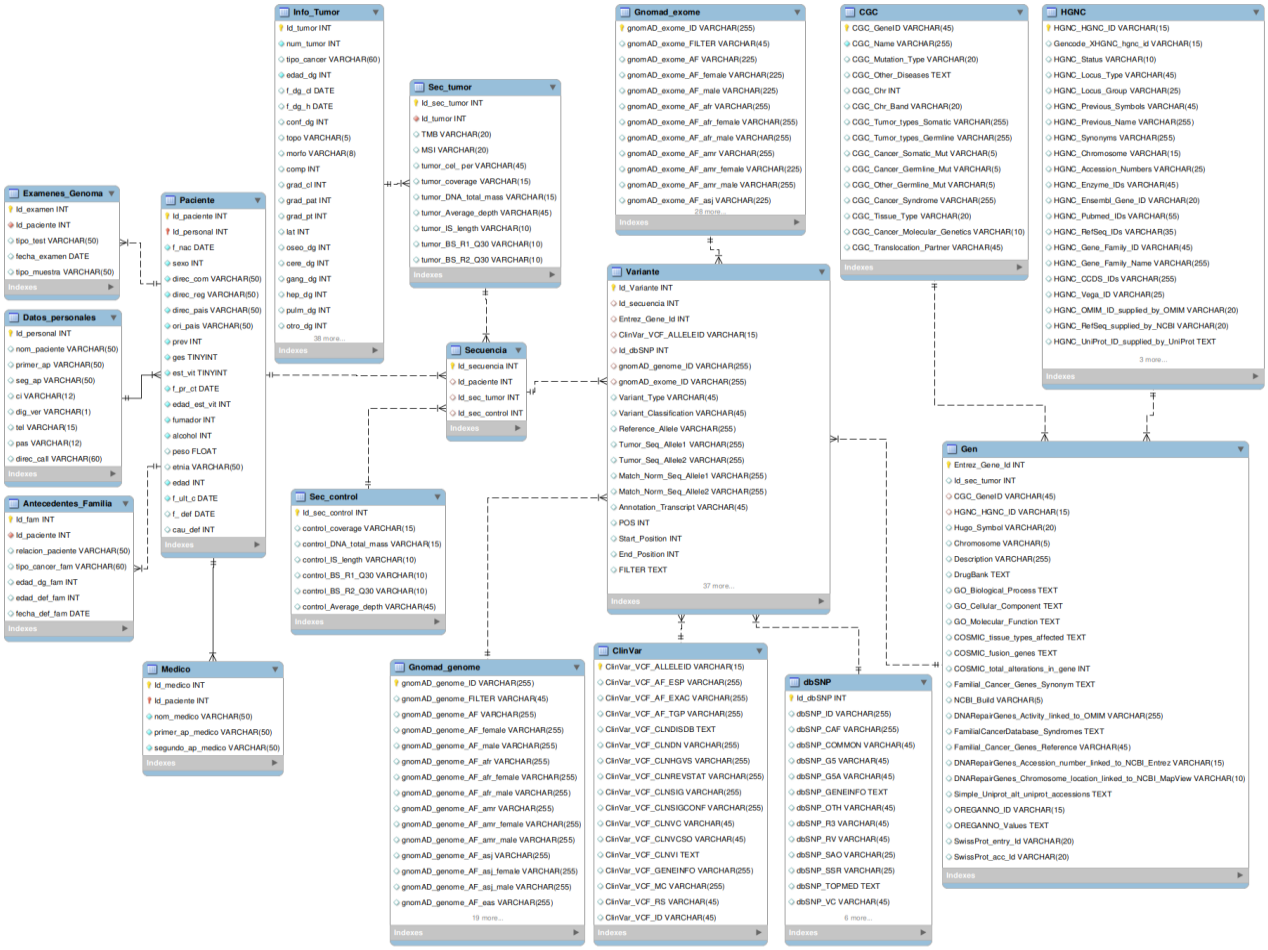
Anexo 4: Tabla descriptiva de las herramientas utilizadas en la creación del pipeline para detectar variantes somáticas y germinales.



Anexo 5: Diseño conceptual de la base de datos.



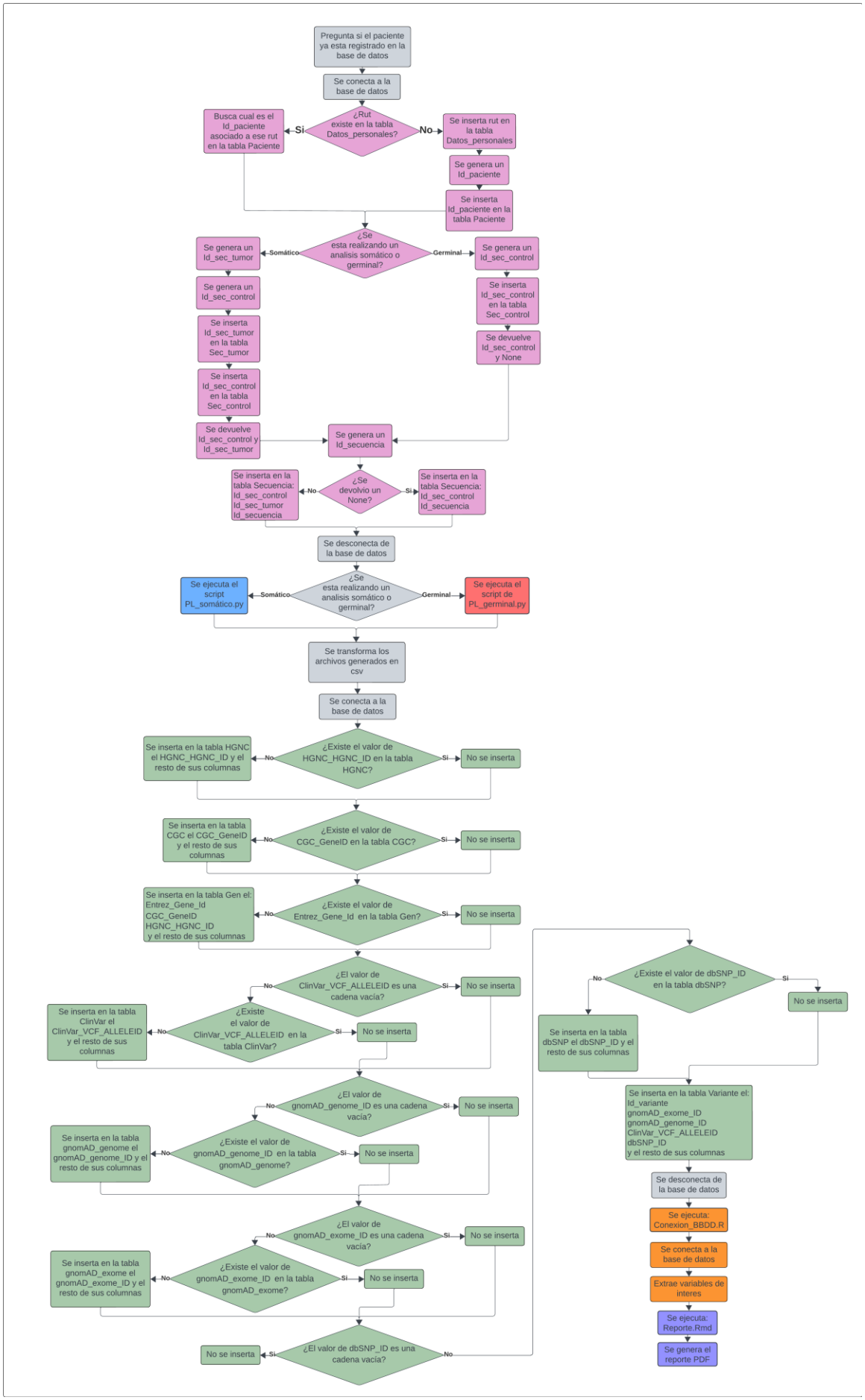
Anexo 6: Diseño lógico de la base de datos



Anexo 7: Diseño físico de la base de datos en MySQL Workbench

Etapa	Herramienta	Tiempo promedio
Formateo de la data	*FASTQC*	484 seg
	Trim Galore	1.899 seg
	Bowtie2	11.206 seg
	Samtools	358 seg
Preprocesamiento	AddOrReplaceReadGroups	382 seg
	MarkDuplicates	1.377 seg
	Samtools	448 seg
	BaseRecalibrator	464 seg
	ApplyBQSR	392 seg
Detección Mutantes	Mutect2	18.345 seg
	getPileupSummaries	498 seg
	CalculateContamination	471 seg
Post procesamiento	LearnReadOrientationMode	581 seg
	FilterMutectCalls	1.149 seg
	Funcotator	3.158 seg
Tiempo total		41.322 seg

Anexo 8: Tabla con los tiempos promedio de cada herramienta del pipeline para la detección de variantes somáticas.


Pregunta si el paciente ya esta registrado en la base de datos
Se conecta a la base de datos
¿Rut existe en la tabla Datos_personales?
Si
Busca cual es el Id_paciente asociado a ese rut en la tabla Paciente
No
Se inserta rut en la tabla Datos_personales
Se genera un Id_paciente
Se inserta Id_paciente en la tabla Paciente
¿Se esta realizando un analisis somático o germinal?
Somático
Se genera un Id_sec_tumor
Se genera un Id_sec_control
Se inserta Id_sec_tumor en la tabla Sec_tumor
Se inserta Id_sec_control en la tabla Sec_control
Se devuelve Id_sec_control y Id_sec_tumor
Germinal
Se genera un Id_sec_control
Se inserta Id_sec_control en la tabla Sec_control
Se devuelve Id_sec_control y None
Se genera un Id_secuencia
¿Se devolvio un None?
No
Se inserta en la tabla Secuencia: Id_sec_control Id_sec_tumor Id_secuencia
Si
Se inserta en la tabla Secuencia: Id_sec_control Id_secuencia
Se desconecta de la base de datos
¿Se esta realizando un analisis somático o germinal?
Somático
Se ejecuta el script PL_somático.py
Germinal
Se ejecuta el script de PL_germinal.py
Se transforma los archivos generados en csv
Se conecta a la base de datos
¿Existe el valor de HGNC_HGNC_ID en la tabla HGNC?
No
Se inserta en la tabla HGNC el HGNC_HGNC_ID y el resto de sus columnas
Si
No se inserta
¿Existe el valor de CGC_GeneID en la tabla CGC?
No
Se inserta en la tabla CGC el CGC_GeneID y el resto de sus columnas
Si
No se inserta
¿Existe el valor de Entrez_Gene_Id en la tabla Gen?
No
Se inserta en la tabla Gen el: Entrez_Gene_Id CGC_GeneID HGNC_HGNC_ID y el resto de sus columnas
Si
No se inserta
¿El valor de ClinVar_VCF_ALLELEID es una cadena vacía?
No
¿Existe el valor de ClinVar_VCF_ALLELEID en la tabla ClinVar?
No
Se inserta en la tabla ClinVar el ClinVar_VCF_ALLELEID y el resto de sus columnas
Si
No se inserta
Si
No se inserta
¿El valor de gnomAD_genome_ID es una cadena vacía?
No
¿Existe el valor de gnomAD_genome_ID en la tabla gnomAD_genome?
No
Se inserta en la tabla gnomAD_genome el gnomAD_genome_ID y el resto de sus columnas
Si
No se inserta
Si
No se inserta
¿El valor de gnomAD_exome_ID es una cadena vacía?
No
¿Existe el valor de gnomAD_exome_ID en la tabla gnomAD_exome?
No
Se inserta en la tabla gnomAD_exome el gnomAD_exome_ID y el resto de sus columnas
Si
No se inserta
Si
No se inserta
¿El valor de dbSNP_ID es una cadena vacía?
Si
No se inserta
No
¿Existe el valor de dbSNP_ID en la tabla dbSNP?
No
Se inserta en la tabla dbSNP el dbSNP_ID y el resto de sus columnas
Si
No se inserta
Se inserta en la tabla Variante el: Id_variante gnomAD_exome_ID gnomAD_genome_ID ClinVar_VCF_ALLELEID dbSNP_ID y el resto de sus columnas
Se desconecta de la base de datos
Se ejecuta: Conexion_BBDD.R
Se conecta a la base de datos
Extrae variables de interes
Se ejecuta: Reporte.Rmd
Se genera el reporte PDF

Anexo 9: Diagrama de actividad. Este diagrama muestra qué hace cada uno de los scripts involucrados en la generación del reporte final. Las casillas en gris son las acciones realizadas por el script main.py, las casillas rosadas son las acciones realizadas por el script Alimentar_BBDD.py, las casillas en verde son las acciones realizadas por el script Extraer_info.py, las casillas naranjas son las acciones realizadas por el script Conexión_BBDD.R y, por último, las casillas moradas son las acciones realizadas por el script Reporte.Rmd. En cuanto a las acciones realizadas por PL_somático.py y PL_germinal.py se encuentran en el anexo 2 y 3.

b. Glosario

- Secuenciación NGS: tecnología avanzada que permite analizar simultáneamente millones de fragmentos de ADN, transformando la investigación genética y la medicina al proporcionar datos genómicos a gran escala de manera eficiente.
- Sanger: método clásico y secuencial para determinar la secuencia de nucleótidos en una cadena de ADN, utilizando la síntesis de cadenas complementarias y la incorporación de didesoxinucleótidos marcados.
- qPCR: Técnica molecular que permite la amplificación y cuantificación precisa de ácidos nucleicos, brindando información sobre la cantidad inicial de material genético presente en una muestra.
- Variantes somáticas: cambios genéticos que ocurren específicamente en las células somáticas, que no son heredables.
- Variantes germinales: alteraciones genéticas que se heredan y se encuentran en las células germinales, como óvulos y espermatozoides, y pueden transmitirse a la descendencia.
- FASTQ: formato de archivo utilizado en bioinformática para almacenar datos de secuenciación de ADN, que incluye secuencias de nucleótidos y sus respectivas puntuaciones de calidad asociadas.
- Read 1 y Read 2: representan dos fragmentos de secuencia distintos obtenidos de una misma región de ADN, permitiendo una visión completa del material genético en términos de secuencia y orientación.
- Adaptadores: son secuencias cortas de ADN que se añaden a los extremos de los fragmentos de material genético durante la preparación de muestras para la secuenciación, facilitando la unión a las plataformas de secuenciación y permitiendo la amplificación y detección eficientes.
- Genoma de referencia: una secuencia completa y representativa del ADN de una especie particular, utilizada como estándar para comparar y analizar la variabilidad genética en individuos de esa especie.
- Gen: Conjunto de bases nitrogenadas.
- Duplicados de lecturas: copias idénticas de fragmentos de secuencia en datos de secuenciación.

- F1R2: F1 es la muestra de sangre normal o tejido no tumoral y R2 es la muestra del tejido tumoral. Entonces, "F1R2" indica las observaciones de bases (números de lecturas) en las que la variante se observa en la muestra del tejido tumoral (R2) y no en la muestra de sangre normal (F1). Esta información es útil para distinguir variantes somáticas (presentes solo en las células tumorales) de variantes germinales (heredadas y presentes tanto en el tejido normal como en el tumoral).
- Pila: Conjunto de lecturas que se alinean al mismo lugar en el genoma.
- Cobertura: indica cuántas veces, en promedio, un determinado nucleótido o región del genoma ha sido secuenciado.
- Frecuencia de alelos: proporción relativa de diferentes variantes alélicas (formas alternativas de un gen) dentro de una población, proporcionando información sobre la diversidad genética y la distribución de los alelos en dicha población.
- Orientación de las lecturas: Se refiere a la dirección en la que las secuencias de ADN son leídas durante el proceso de secuenciación.
- Pileup: Imagen que muestra cómo las secuencias de lectura se alinean con una referencia genómica.
- Computer Vision: Campo de la inteligencia artificial, donde se analizan datos visuales.
- Wetlab: la sección del laboratorio que abarca la secuenciación del ADN.
- Drylab: la parte bioinformática encargada de la detección de genes mutados.

con