Universidad del Desarrollo
Facultad de Ingeniería

VARIANT FINDER: Un procesador de datos genéticos para usuarios no expertos

POR: JUAN FRANCISCO CALDERÓN GIADROSIC, PhD

Proyecto de grado presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

Profesor guía:

MARÍA PAZ RAVEAU MORALES, PhD

Diciembre 2023

Santiago, chile

Dedicado a Isabella, mi *Tamalito*, que vino a cambiarlo todo y a todos.

Dedicado también a María de los Ángeles, *Mari*, por ser la tremenda compañera que es.

AGRADECIMIENTOS

Quisiera agradecer en primer lugar, a mi familia, por aguantarme siempre las locuras académicas (y otras) que se me ocurre hacer. Sin su amoroso apoyo incondicional no podría haber cursado este Magister y no haber fallado rotundamente en el intento.

En particular, agradecer a mi hija Isabella y a mi compañera Mari, por saber aceptar todas las veces que robé tiempo en familia con tareas y trabajos.

A mis colegas y autoridades del Centro de Genética y Genómica y del Instituto de Ciencias e Innovación en Medicina por su apoyo incondicional desde el momento en que les planteé la idea de cursar este Magister. A mis amigos en el Centro y el Instituto por darme ánimo y cariño cuando se necesitaba perseverar. ¡Gracias a todos!

A los buenos amigos que descubrí en el transcurso del Programa: Adolfo, Patricio y Catalina. Nuestros sábados AM fueron mucho más llevaderos gracias a ustedes y su buena compañía. Espero que conservemos esta naciente amistad.

A María Paz Raveau, mi profesora guía en la ejecución de este *Capstone Project*, por su esfuerzo genuino para conversar de un tema que no es el suyo y hacerlo con máxima rigurosidad y comprensión por las circunstancias en que se ejecutó este trabajo.

Tabla de contenidos

1. Resumen

La generación de datos genómicos ha aumentado exponencialmente y, lejos de mostrar signos de estancarse, cada vez los costos son menores y las aplicaciones de genómica, transcriptómica, epigenómica y otras basadas en HTS aumentan[1]. Sin ir más allá, al año 2016 el repositorio de datos genómicos más grande del mundo (SRA) tenía datos almacenados por $4x10^{15}$ bytes[2] y su tamaño actual se estima en $2,3x10^{16}$ bytes[3]. Por tanto, la capacidad de cómputo que se ha tenido que implementar para un correcto manejo y análisis de los datos generados es también, significativa. Así mismo, la bioinformática pasó de ser una herramienta a ser una disciplina de estudio en sí misma[2,4].

Este trabajo es una aproximación para generar una herramienta diseñada para facilitar el uso e interpretación de datos genéticos y ponerla a disposición de colegas investigadores no expertos en manejo de datos genéticos. La propuesta busca que los datos sean usados de una manera eficiente y dinámica sin necesidad de implementar códigos de programación para los que el/la investigadora pueden ser ajenos y no ser objetivo central de la investigación para la que los datos genéticos fueron generados.

En este sentido, la propuesta muestra un prototipo de interfaz web en la que el usuario ingresará un archivo de variantes genéticas (*Variant Call File* o *.vcf) e ingresará palabras clave para que el código implementado en el *backend* haga una búsqueda en *Pubmed*. A partir de esta búsqueda, el código implementado obtendrá los genes mencionados en los resúmenes de los trabajos indexados en Pubmed y filtrará el archivo de variantes (vcf) para mostrarle al investigador solamente las variantes presentes en genes que han sido previamente asociadas a los conceptos ingresados en la búsqueda de Pubmed.

La propuesta pretende facilitar el análisis personalizado de datos genómicos desde una perspectiva de usuario no experto. Es decir, pone a disposición una interfaz de fácil acceso (web) y con un flujo de información simple y lógico, pero que a la vez permite que el usuario final tenga una herramienta robusta de análisis de datos que, de otra manera, no podría hacer por sí solo.

2. Introducción

2.1. Contexto General

La producción de datos genómicos ha crecido exponencialmente en las últimas décadas. De hecho, junto con la generación de imágenes médicas ha sido uno de los puntales de la incorporación del término "*big data*" en el ámbito de la salud. Sin embargo, este crecimiento ha sido posterior al salto en volumen de datos generados por otras actividades o disciplinas humanas. Si considerásemos la creación de YouTube (año 2005) como el momento en que se comienza a romper el paradigma clásico de almacenamiento local de datos para pasar a almacenamiento masivo remoto de ellos, entonces notaremos que llevamos 20 años trabajando en un paradigma donde el límite físico que impone el almacenamiento de datos es cada vez más tenue, por no decir que ha desaparecido del todo.

La publicación del primer borrador del Genoma Humano[5] trajo consigo una transformación significativa e irreversible de la manera en que se estudiaba la genética. De hecho, el concepto de bioinformática nace a partir de la necesidad nueva de manejar datos asociados principalmente a genómica y que podían ser generados cada vez más fácilmente y, por tanto, en un volumen mayor. Curiosamente, el concepto nace incluso antes de que la comunidad científica entendiera el grado de variación que había entre un humano y otro y, por tanto, que documentar esa variación iba a tornarse extremadamente importante. Aquel primer borrador del Genoma Humano, de hecho, fue realizado con unas pocas muestras de donantes anónimos, y se había asumido, sin mayor evidencia, que esa secuencia leída iba a ser extremadamente conservada. No fue sino hasta el 2005 que otro consorcio internacional haría una primera estimación del grado de variación existente entre genomas y de la relevancia de comenzar a almacenar esa información sistemáticamente y con mayor representatividad de distintas etnias y ancestrías[6].

Uno de los productos más inmediatos del Proyecto Genoma Humano (HGP) es que se pudo calcular con mayor exactitud el tamaño del genoma, el tamaño y numero de genes que lo componían y establecer algunos hechos científicos que, antes de este evento, eran estimados de manera no siempre acuciosa. Algunos "datos" del genoma humano estimados por el HGP pueden ser encontrados aquí[7]. Interesantemente, la estimación del tamaño del Genoma Humano rápidamente convergió en torno a los $3x10^9$ pares de bases (3Gb o gigabases) y el número de genes estimados en el genoma humano bajó de 100.000 (un número que fue estimado a partir del número de proteínas estimadas) a los 30.000-35.000 genes en ese primer borrador y, a través de los años, se ha fijado en 20.000 – 25.000 genes. Veremos más adelante que a partir de este mapa de ruta o *roadmap* del Genoma Humano, el desarrollo de herramientas para el manejo de datos genómicos tuvo que lidiar siempre con un marco de referencia que no es estático,

sino que, con cada nuevo hallazgo en el campo de la genómica, varía y se reestablece en torno a lo que surge a partir de la nueva evidencia.

Otro punto de inflexión en el desarrollo de la genómica ocurrió en el 2008. Hasta entonces, la secuenciación de material genético ocurría mediante una reacción de amplificación por reacción en cadena de la polimerasa (PCR) que permitía que para cada templado (secuencia de DNA a ser leída o secuenciada) el usuario pudiera leer la secuencia en ambos sentidos, denominados *forward* y *reverse*. Esto significaba que, para cada nucleótido contenido en esta secuencia, la *profundidad* con la que se leía era de un máximo de 2X (1x en una dirección y 1x en la opuesta). Pero a partir del 2008, una nueva tecnología denominada "secuenciación de siguiente generación" o "*next generation sequencing NGS",* en inglés, permitió que la profundidad con que se leía un nucleótido de una secuencia cualquiera que ingresara al proceso podía aumentar hasta 20X, 30X o más. Más detalles respecto a estas tecnologías puede ser encontrado en Mardis E. et al[8] y en Hu et al[1].

El impacto más inmediato que tuvo el advenimiento del NGS (Actualmente llamada *High Throughput Sequencing* o *HTS*) fue que los costos de secuenciación, medidos en U$D por Mb ($10^6$ pares de bases) comenzaron a caer exponencialmente. Si inmediatamente terminado el primer borrador del Genoma Humano se estimaba que secuenciar *de novo* un genoma humano cualquier costaba U$D95 millones, en enero del 2008 este monto ascendía a U$D3 millones. En enero del 2009 "un genoma costaba" U$D200.000 y en 2023, hay empresas que ofrecen el servicio de secuenciación de genoma completo (WGS de la sigla *Whole Genome Sequencing*) por apenas 500 dólares, a una profundidad de 30X en promedio. La mejor manera de ilustrar el impacto de esta tendencia es compararla con el principio que rige el aumento del poder de cómputo, más conocido como Ley de Moore. En general, esta ley indica que el poder de cómputo empujado por la tecnología en esa disciplina no puede más que duplicarse cada dos años. Se extrapola, además, que cualquier otro desarrollo tecnológico que sea capaz de seguir el ritmo impuesto por la Ley de Moore es un desarrollo tecnológico que está creciendo a su máxima capacidad y cuya industria "goza de extremadamente buena salud". El siguiente gráfico muestra el comportamiento de los costos de secuenciación de DNA comparado con lo predicho por la Ley de Moore[9].

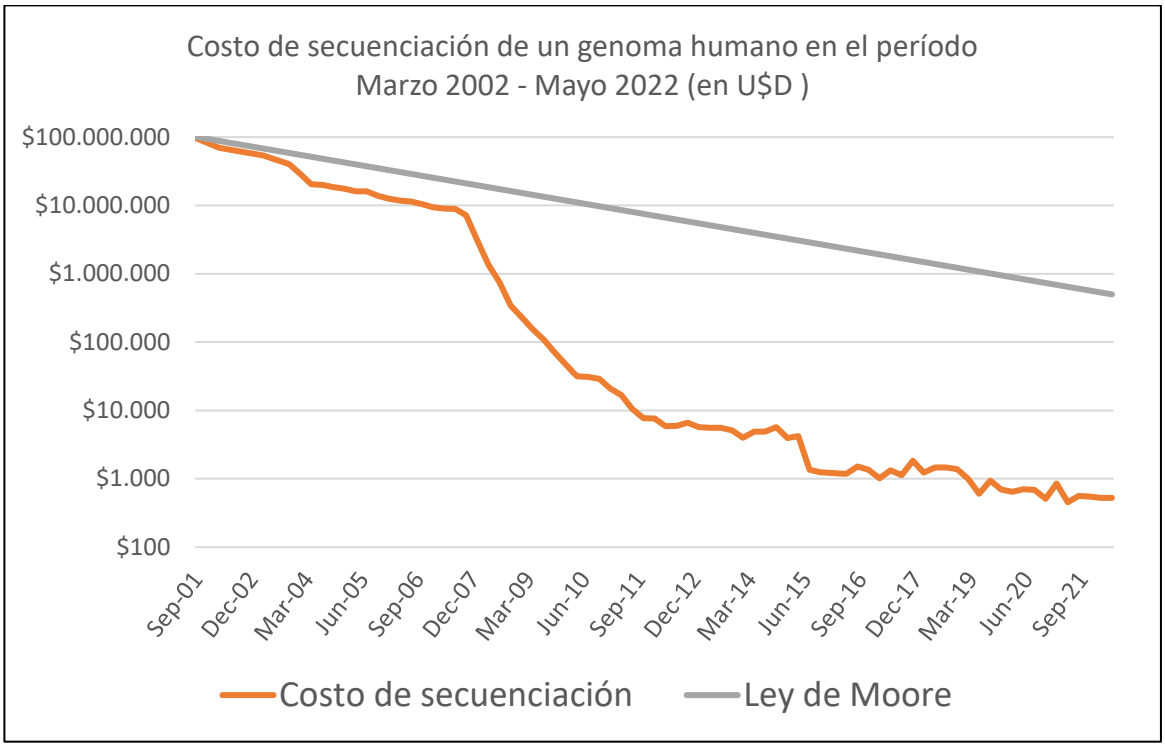

Figura 1. Caída en los costos de secuenciación en dólares estadounidenses vs. predicción de caída de costos de cómputo según lo predicho por la Ley de Moore. Datos obtenidos desde NHGRI[9]

La generación de datos genómicos a partir del 2008 creció exponencialmente y no parece detenerse. Al contrario, ni las predicciones más optimistas que se hicieron a 10 años del primer borrador de HGP pudieron prever la explosión en la generación de datos genómicos[10]. Este fenómeno ha impuesto desafíos de manejo y análisis de dichos datos que la comunidad científica ha enfrentado con mayor o menor eficiencia dependiendo de las herramientas disponibles. Más adelante veremos que a mayores volúmenes de datos, mayores son los desafíos en cuanto al manejo mismo del set creado, pero también en cuanto a la interpretación del significado biológico de los datos obtenidos. Esto, que ya es difícil en un contexto de datos experimentales obtenidos en un laboratorio, es mucho más exigente cuando el entorno en el que se generan los datos genómicos es clínico y/o involucra la entrega de información diagnóstica a seres humanos.

Este trabajo es una aproximación para generar una herramienta diseñada para facilitar el uso e interpretación de datos genéticos y ponerla a disposición de colegas investigadores no expertos en manejo de datos genéticos. La propuesta busca que los datos sean usados de una manera eficiente y dinámica sin necesidad de implementar códigos de programación para los que el/la investigadora pueden ser ajenos y no ser

objetivo central de la investigación para la que los datos genéticos fueron generados. En este sentido, la propuesta muestra un prototipo de interfaz web en la que el usuario ingresará un archivo de variantes genéticas (*Variant Call File* o *.vcf) e ingresará palabras clave para que el código implementado en el *backend* haga una búsqueda en *Pubmed*. A partir de esta búsqueda, el código implementado obtendrá los genes mencionados en los resúmenes de los trabajos indexados en Pubmed y filtrará el archivo de variantes (vcf) para mostrarle al investigador solamente las variantes presentes en genes que han sido previamente asociadas a los conceptos ingresados en la búsqueda de Pubmed.

Esta herramienta para el uso de datos es un prototipo que deberá ser mejorado en el futuro. Sin embargo, es un punto de partida para facilitar el acceso y mejorar el entendimiento de datos genómicos que pueden haber sido obtenidos sin necesariamente contar con la preparación para manejarlos.

Finalmente, y a pesar de que no es el objetivo central del presente trabajo, en la parte final de éste discutiremos los desafíos futuros para el manejo de datos genómicos desde las aristas regulatorias y éticas.

2.2. Impacto de los Datos Genómicos en Salud

Las expectativas generadas cuando se completó el primer borrador del HGP fueron enormes. Estas expectativas incluían el deseo ferviente de agrupaciones de pacientes con enfermedades genéticas de que la cura para sus padecimientos estuviera más cerca, pero también incluyó especulaciones bursátiles respecto al precio de acciones de farmacéuticas y empresas biotecnológicas, que prometían trabajar en mejorar la salud de poblaciones vulnerables, la mayoría de las veces, sin fundamento[11].

Diez años después, en el 2011, ya se visualizaba la utilización de las tecnologías de secuenciación de última generación (HTS) para resolver problemas de salud o más bien dicho, para comenzar a entender la biología y la fisiopatología que subyace a muchos fenotipos con modos de herencia conocidos o al menos signos de agregación familiar que sugiriera una arquitectura genética[11,12]. A esas alturas, más de 2.800 enfermedades monogénicas o mendelianas (causadas por variantes en un solo gen y, por tanto, que muestran modos de herencia como aquellos descritos por Gregor Mendel en 1865) habían sido “entendidas” debido a que se había podido identificar el gen causante[11]. Una de las razones para este auge en la utilización de los datos genómicos que cada vez se podían obtener a más bajo costo fue que se desarrollaron algoritmos de análisis que permitieron identificar las variantes genéticas patogénicas de manera certera y acuciosa[13].

Este fenómeno dio origen a consorcios cuyo único propósito era develar las variantes genéticas detrás de los aproximadamente 10.000 fenotipos de aparente causa monogénica[14,15]. Esto fue de extrema importancia, debido a que al mismo tiempo que diferentes grupos de investigación alrededor del mundo buscaban estas variantes monogénicas, nacía la necesidad de establecer mecanismos de compartir datos que hicieran compatibles los esfuerzos de estos investigadores.

3 años antes, en Estados Unidos había sido promulgada la Ley de No-Discriminación por Información Genética (GINA, por sus siglas en inglés). Esta ley buscaba regular la manera en que se almacenaban y compartían datos genéticos de manera que no se pudieran utilizar para discriminar a nadie en función de riesgos de desarrollo de alguna enfermedad o fenómenos similares de esta naturaleza[16].

Ya en 2011, Greenbaum y cols. sugerían que el naciente campo de la bioinformática y quienes lo estaban implementando, sufrían de una paradoja estructural: el desarrollo de la disciplina se debía principalmente a las políticas de fuente abierta (*open source*) y de datos abiertos (*open data*) que habían permitido usar estos *datasets* gigantes para el desarrollo de algoritmos y *pipelines* de trabajo. Sin embargo, dada la naturaleza inédita de los datos genómicos (generalmente asociados a datos clínicos de pacientes que tienen derecho a la privacidad de estos), este desarrollo podía verse limitado por el deber ético de proteger la información sensible de salud[17].

Durante la década que pasó, es decir, hasta 20 años después de la publicación del primer borrador de HGP la comunidad científica realizó grandes esfuerzos para lograr un cuerpo de gobernanza que permitiera compartir datos de manera segura pero que apoyara el avance de la investigación en genómica y salud de la manera más decidida posible. Esto, debido a que las distintas plataformas de HTS generaban datos levemente distintos entre sí, y que, si no se lograban complementar y reunir en una misma base de datos, serían imposibles de comparar. Para ello, primeramente, se debió establecer un formato interoperable de compartir información acerca de la comparación de datos genómicos generados *de novo* con respecto al genoma *de referencia* (es decir, un constructo "promedio" de todos los genomas secuenciados para una determinada especie). Este formato se llamó Variant Call Format / File y lo discutiremos en detalle en la próxima sección.

3. Trabajo Relacionado

En una revisión hecha en el 2015, Stephens et al indican que el ciclo de vida de un set de datos se compone de cuatro etapas: adquisición, almacenamiento, distribución y análisis[18]. Al mismo tiempo, ellos presentan una comparación entre cuatro dominios de datos que son un excelente ejemplo que dará pie a la fundamentación del presente proyecto: datos astronómicos, datos de X (Twitter), datos de YouTube y datos genómicos. La comparación está en la tabla 1.

Fase de ciclo de vida	**Datos astronómicos**	**Datos de X (Twitter)**	**Datos de YouTube**	**Datos genómicos**
Adquisición	25 zetta-bytes / año	0.5 – 15 10^9 tweets / año	500-900 millones de horas / año	1 zetta-byte / año
Almacenamiento	1 EB / año	1-17 PB /año	1 – 2 EB/año	2-40 EB / año
Análisis	Volúmenes masivos en tiempo real	Análisis de metadata NPL	Muy limitado	Heterogéneo 2 a 10.000 trillones de horas CPU
Distribución	600 Tb / seg	Pequeño	Uso primario del ancho de banda de internet	Transmisión masiva de pequeños subsets de datos

Tabla 1. Características principales de cuatro dominios de generación de datos. Adaptado de Stephens et al[18].

Tal como mencionamos en la sección anterior, la generación de datos genómicos ha aumentado exponencialmente y, lejos de mostrar signos de estancarse, cada vez los costos son menores y las aplicaciones de genómica, transcriptómica, epigenómica y otras basadas en HTS aumentan[1]. Sin ir más allá, al año 2016 el repositorio de datos genómicos más grande del mundo (SRA) tenía datos almacenados por $4x10^{15}$ bytes[2] y su tamaño actual se estima en $2{,}3x10^{16}$ bytes[3]. Por tanto, la capacidad de cómputo que se ha tenido que implementar para un correcto manejo y análisis de los datos generados es también, significativa. Así mismo, la bioinformática pasó de ser una herramienta a ser una disciplina de estudio en sí misma[2,4].

3.1. Adopción de estándares comunes en el manejo de datos genómicos

El estándar de datos ha convergido ya en los últimos 8 a 10 años a un conjunto de formatos en los que los datos genómicos deben ser producidos y almacenados. Describir toda la trayectoria que ha seguido la comunidad excede el objetivo del presente trabajo, pero baste con mencionar algunas de las iniciativas más relevantes en términos de estandarizar métodos de trabajo, protocolos de generación y almacenamiento de datos y mecanismos para compartir grandes sets de datos genómicos entre distintos operadores.

La referencia central, siempre pensando en datos genómicos humanos, es el *Global Alliance for Genomics and Health.* Este consorcio internacional se formó en 2013 y actualmente incluye más de 500 organizaciones que han construido un mapa de ruta estratégico que incluye la completa estandarización de los elementos antes mencionados[19,20]. El objetivo central de la *Global Alliance* es aunar fuerzas para que la implementación de la medicina genómica como practica rutinaria en los sistemas de salud sea una realidad.

La adopción de estándares comunes, sin embargo, antecede a la formación de G*lobal Alliance*. De hecho, desde el advenimiento de las tecnologías de HTS se viene trabajando en estándares de archivos y datos genómicos. Una de las referencias claves para entender el desarrollo de esta disciplina es la que muestra las características del archivo inicial en un experimento de secuenciación masiva: el archivo FASTQ[21]. A partir de la convergencia en los archivos de salida de las distintas tecnologías de secuenciación, entonces la adquisición de distintos *pipelines* de análisis sería más fácil. En la práctica, todo experimento de secuenciación masiva HTS genera archivos FASTQ que contiene las secuencias cortas que luego se alinearan y mapearan contra el genoma de referencia para generar un genoma experimental. Un set estandarizado de herramientas se utiliza para estos procesos. Más detalle de su origen e implementación se puede encontrar en el artículo publicado por Danecek y otros[22].

Anteriormente mencionamos que los datos obtenidos por un secuenciador HTS podrían provenir de distintas fuentes experimentales: genómicas (lectura de la secuencia nucleotídica del ADN de una muestra biológica); transcriptómicas (lectura y cuantificación relativa de la cantidad de ARNm de una muestra biológica); epigenómica (lectura de secuencias de ADN que poseen ciertas modificaciones adicionales a su estructura nucleotídica); entre otras. **Para efectos del presente trabajo, nos enfocaremos única y exclusivamente en datos genómicos. Es decir, lectura de secuencia de ADN.**

En el caso de uso que veremos en este trabajo, trabajaremos con un tipo de archivo denominado *Variant Call Format* o Archivo *vcf*. El archivo VCF debe cumplir con un formato acordado por la comunidad y cuya principal característica es que contiene todas las *diferencias* entre la secuencia de ADN leída en una muestra determinada en comparación con el genoma de referencia contenido en la base de datos correspondientes. Más detalles se pueden ver en la figura 2, tomada de Danecek y otros[23].

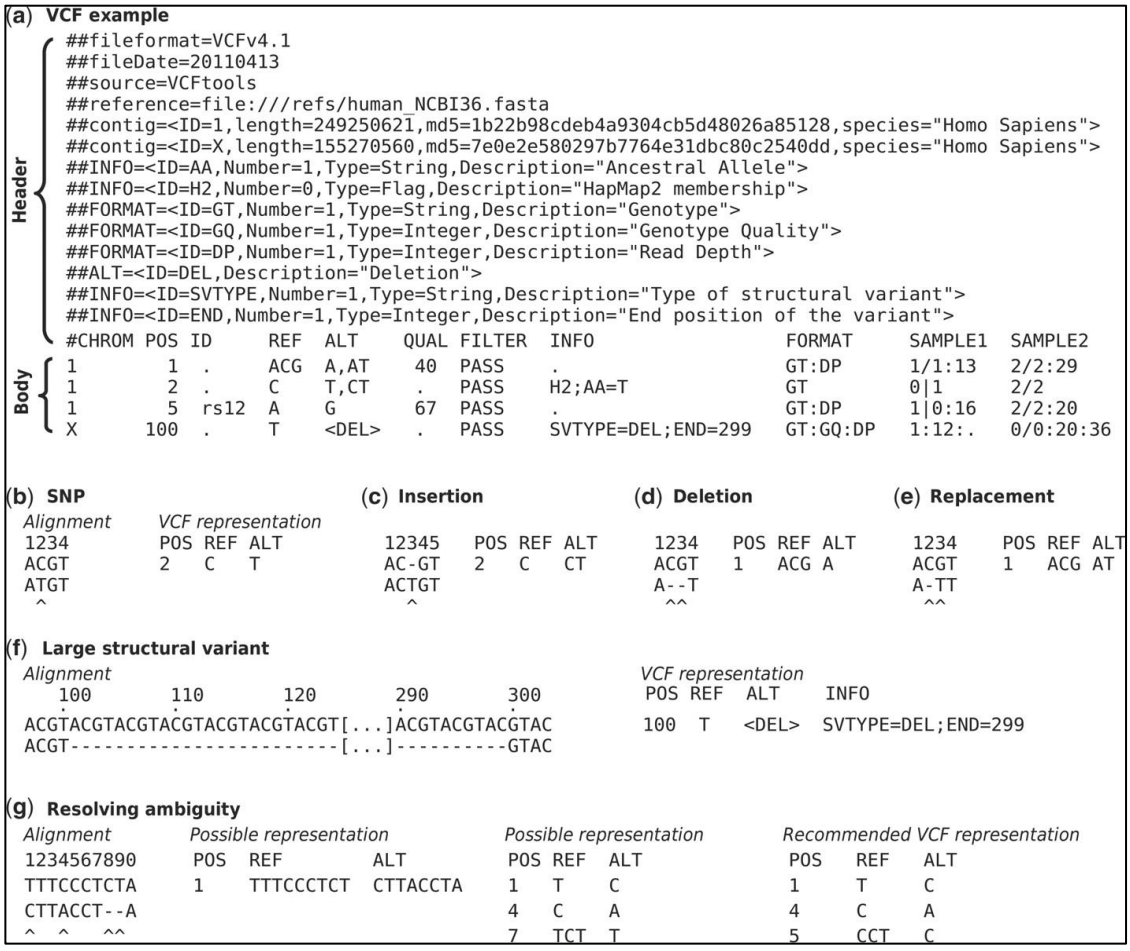

Figura 2. Estructura del archivo VCF. En el encabezado se encuentra todo el metadato necesario para verificar la compatibilidad e interoperabilidad de la información contenida en el archivo VCF. Adaptada de Danacek y otros[23].

Como se puede observar en la Figura 2, el archivo VCF es de un formato relativamente simple. Sin embargo, un archivo VCF que proviene de un experimento

con secuenciación de genoma completo o WGS, puede llegar a contener fácilmente 100.000 o 150.000 variantes o filas en su sección b).

Este fenómeno hace que existan numerosos intentos de facilitar la interpretación y uso de los datos contenidos en un archivo VCF. En algunos casos, la idea implementada tiene que ver con generar interfaces de visualización que se asemejen a los contenidos en UCSC Genome Browser o Ensembl. Estos dos portales de información genética son los más utilizados por la comunidad científica. Sin embargo, debido a las múltiples aplicaciones contenidas en estos portales se requiere cierto manejo de las capacidades de las plataformas y de la nomenclatura necesaria para navegarlas, lo que las hacen poco amigables para los usuarios no especializados.

En otros casos, se ha intentado implementar métodos de filtrado de variantes genéticas que sean más intuitivos (interfaces gráficas o web) y que no exijan conocimientos de programación y/o bioinformática en general.

Entre estos ejemplos podemos mencionar re-Searcher[24], BrowseVCF[25], VIVA[26], VCF.Filter[27], VCF-Miner[28], IGV[29], VCF-kit[30], entre otros.

Estas aplicaciones tienen aspectos en común, a saber: Generalmente son *standalone* (es decir, se ejecutan sin dependencias de software), La interfaz es línea de comandos o GUI, todas tienen distintas modalidades de filtrado de variantes y todas otorgan la posibilidad de guardar el resultado de las estrategias de filtrado implementadas por el usuario.

En algunos casos, la aplicación ha sido diseñada para poder asociar variantes con efectos patogénicos ya anotados en la literatura. Por ejemplo, BrowseVCF conecta y filtra un archivo VCF con una aplicación remota denominada *Variant Effect Predictor*[31] y que es parte de Ensembl. Tal como su nombre lo indica, VEP prioriza las variantes genéticas según su potencial para causar un efecto de pérdida o ganancia de función de la proteína para la cual codifica un gen cualquiera, lo que se traduce en un alto potencial patogénico para el individuo que sea portador.

VIVA, por otro lado, es una herramienta que se enfoca en integrar no sólo mecanismos de filtrado de variantes genéticas de acuerdo a diferentes criterios de análisis sino que además puede generar visualizaciones de alta calidad de elementos que son usuales en publicaciones de la disciplina: calidad de la secuencia, asociaciones genotipo-fenotipo y análisis preliminares de ligamiento y asociación mediante generación de archivos compatibles con software tipo PLINK[32].

En cierto modo, todos los ejemplos antes indicados, tienen un factor común: la idea de hacer accesible para público no experto en programación, la posibilidad de profundizar el análisis de sus datos (*data mining*) genéticos en la forma de archivos VCF.

El presente trabajo propone una interfaz web en la que se filtren archivos VCF ingresados por el usuario para mostrarle solamente las variantes contenidas en los genes que hayan sido previamente asociados en la literatura con criterios de búsqueda que también serán ingresados por el usuario.

4. Hipótesis y Objetivos

4.1. Hipótesis

La interfaz web implementada en este trabajo filtrará eficientemente las variantes genéticas contenidas en un archivo VCF de acuerdo con los criterios de búsqueda en la literatura que ingrese el usuario.

4.2. Objetivo General

El Objetivo General del presente trabajo es implementar una interfaz de fácil uso para usuarios no especializados que necesiten trabajar con archivos VCF para interpretar los resultados obtenidos en sus experimentos.

4.3. Objetivos Específicos

1. Implementar un motor de búsqueda de literatura científica que identifique eficientemente genes mencionados en los resúmenes o *abstracts* de publicaciones indexadas en Pubmed según criterios de búsqueda del usuario.
2. Implementar una interfaz web que permita al usuario entregar un archivo VCF a la aplicación creada.
3. Implementar una estrategia de filtrado del archivo VCF ingresado de acuerdo con la lista de genes obtenidos y que permita entregar archivos de salida en formatos convencionales de fácil lectura para el usuario.

5. Datos y Metodología

5.1. Datos

El sistema está diseñado para aceptar archivos VCF en su formato convencional (es decir, generado por VCFtools[23]. En el repositorio que acompaña esta entrega hay ejemplos de archivos VCF que fueron utilizados para probar la aplicación implementada.

Por otra parte, el sistema está diseñado para obtener desde UCSC Genome Browser un archivo con todos los genes de un organismo que esté secuenciado y almacenado en este portal. En el caso de uso implementado, este organismo es *Homo sapiens*, es decir, trabajaremos con datos genómicos provenientes de humanos.

Finalmente, se implementó una búsqueda de texto en la que el sistema ingresa vía API a la *National Library of Medicine*, que alberga a Pubmed[33], un portal que contiene más de 36 millones de citas de literatura biomédica incluyendo libros y artículos revisados por pares e indexados de manera sistemática. Esta búsqueda obtiene y almacena texto de un número variable de los resúmenes que sean positivos con los criterios de búsqueda ingresados por el usuario en la interfaz web.

5.2. Metodología

La Metodología para implementar el proyecto consiste en utilización de Lenguaje Python en un entorno de Anaconda Navigator[34] y a través de cuadernos de Jupyter Notebook[35]. Luego, este código será compilado en módulos distintos, de manera de que sean ejecutados de la manera más limpia y eficiente. Este paso se realizará utilizando Visual Studio Code.

Los distintos módulos o funciones serán implementados en una interfaz web que será el contacto con el usuario. El modelo puede ser observado en la Figura 3:

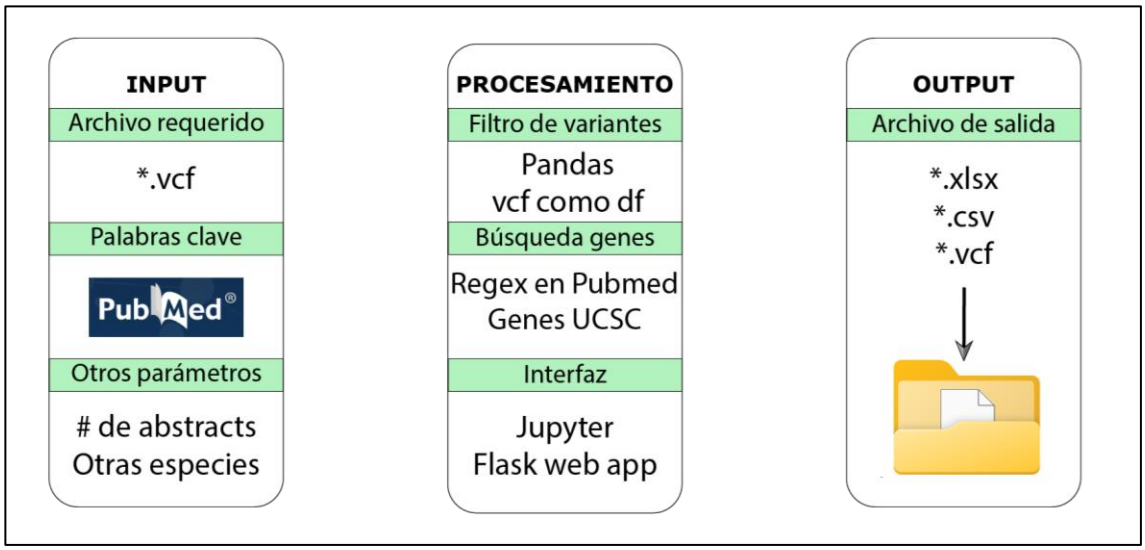

Figura 3. Esquema de implementación del *Variant Browser 1.0*. El *Variant Browser* es una interfaz web que requiere un archivo de variantes genéticas (*.vcf) y términos de búsqueda. A partir de estos dos parámetros obligatorios y otros opcionales, se genera una lista de genes que se utilizará como insumo para filtrar el archivo de variantes de manera de ofrecer como *output* un set de variantes pertinentes única y exclusivamente a los genes asociados con los términos de búsqueda inicial.

6. Resultados

6.1. Búsqueda de genes en la literatura

La búsqueda de genes en la literatura fue realizada programando una extracción de datos desde Pubmed en la que se seleccionaban *abstracts* de trabajos en que las palabras clave ingresadas por el usuario en la interfaz web estuvieran presentes. Esto en general tuvo resultados satisfactorios. Se hicieron ocho búsquedas para distintas enfermedades monogénicas o con base genética conocida así como enfermedades multifactoriales y que pueden tener muchos genes asociados. En este sentido, la lista de genes tuvo un porcentaje variable de coincidencia con otras bases de datos (en particular, OMIM). Esta observación es clave considerando que hay condiciones monogénicas con arquitectura genética muy bien delineada versus las condiciones poligénicas o multifactoriales (denotadas por aquellas que tienen muchas entradas en OMIM) y esto parece impactar en la eficiencia de nuestro motor de búsqueda en la literatura científica. Estos resultados están en la Tabla 2:

Palabras clave	**Genes encontrados**	**Genes en OMIM**[36]	**% coincidencia**
"Marfan Syndrome"	FBN1 TGFBR2 EGF PRKG1	FBN1 TGFBR2	50%
"Loeys-Dietz Syndrome"	TGFB2 TGFB3 TGFBR1 TGFBR2 SKI SMAD2 SMAD3 FBN2 CAPN14 FBN1 CFI BGN	TGFB1 TGFB2 TGFB3 TGFBR1 TGFBR2 SMAD2 SMAD3	50%
"Duchenne Muscular Dystrophy"	DMD MARK2	DMD TCTEX1D1	50%
"Cystic Fibrosis"	CFTR	CFTR CFM1	100%
"Shprintzen-Goldberg Syndrome"	TGFBR1 B4GALT7 COL3A1 XYLT1 TGFBR2 SMAD2 SKI TGFB2 IPO8 MTHFR EGF CD4 ACE XYLT2 B3GALT6 FBN1 IGF1R TGFB3	SKI	5%

	B3GAT3 ADAMTSL4 SMAD3 FES		
“Cleft palate”	IRF6 SMCP KMT2D	N=1.909 entradas OMIM	100%
“hypothyroidism”	CA2 CD28 MET DIO2 SIRT2 AVP FAS CD4 ACE CRP MPI VIP CIT PRL	N=353 entradas OMIM	64%
“Tetralogy of Fallot”	TF FABP4 HP PVR VIP APC MB	NKX2-5 GATA4 ZFPM2 GATA6 JAG1 TBX1	0%

Tabla 2. Análisis de coincidencia de búsqueda de nombres de genes comparado con una base de datos (OMIM) que lista genes que comprobadamente están asociados a una enfermedad o palabra clave, como las de la columna 1.

6.2. Filtrado de archivos VCF

El filtrado de archivos VCF funcionó correctamente. El caso de uso fue realizado con archivos VCF provenientes de pacientes con un diagnóstico clínico conocido y con una variante genética previamente identificada, por lo que se conocía de antemano el resultado del VCF filtrado. Los archivos se encuentran en el repositorio de GitHub que contiene todo lo indicado en el anexo 1.

7. Conclusiones

La implementación de este producto de datos resultó un desafío interesante desde el punto de vista técnico. El hecho de que ya existan múltiples productos que intentan facilitar la interpretación y el análisis de archivos de variantes genéticas es indicativo de que hay un nicho que vale la pena explorar y (tratar de) cubrir.

En referencia a los objetivos planteados en este Proyecto, su cumplimiento fue parcial. A continuación se describen los niveles de cumplimiento de cada uno de los objetivos específicos.

Objetivo Específico 1: Implementar un motor de búsqueda de literatura científica que identifique eficientemente genes mencionados en los resúmenes o *abstracts* de publicaciones indexadas en Pubmed según criterios de búsqueda del usuario.

Este objetivo fue cumplido. Se implementó un código que permite, a través de la utilización de API de Pubmed y el uso de expresiones regulares, en conjunto con un "diccionario de genes" obtenido de UCSC Genome Browser, la extracción de listas comprensivas de genes mencionados en los *abstracts* explorados.

Objetivo Específico 2: Implementar una interfaz web que permita al usuario entregar un archivo VCF a la aplicación creada.

Este objetivo fue cumplido casi totalmente. El código escrito en lenguaje Python que implementaba las funciones relacionadas a los objetivos específicos 1 y 3 fue implementado de manera que se entregara al usuario final a través de una interfaz web (basada en HTML) en la que existen distintas páginas para cada uno de los pasos del proceso de filtrado de VCF. Sin embargo, considerando los resultados mostrados en la Tabla 2, la estrategia de búsqueda de genes basada en expresiones regulares que busquen dirigidamente palabras que cumplan con las reglas de nomenclatura de genes no arrojó resultados tan específicos como se hubiera esperado. Existen otras alternativas de paquetes en Python que son específicos para búsqueda de genes pero se optó por la búsqueda de expresiones regulares porque se decidió explorar una versión de código implementada 100% por el autor y no utilizar un paquete disponible. Es una mejora prioritaria para futuras versiones del software.

Objetivo Específico 3: Implementar una estrategia de filtrado del archivo VCF ingresado de acuerdo con la lista de genes obtenidos y que permita entregar archivos de salida en formatos convencionales de fácil lectura para el usuario.

Este objetivo fue cumplido. Se requirió definir varias funciones que acompañan esta parte del proceso y cada una pasó por un proceso de optimización para que funcionaran coordinadamente en función de generar estos archivos filtrados.

El código que acompaña este informe es, sin duda, un prototipo sujeto a futuras mejoras. Algunos de los aspectos que podrían mejorar el desempeño de esta interfaz son los siguientes:

1. **Número de *abstracts* analizados**: En este prototipo, se fijó el número de abstracts analizados en n=100. Este número arbitrario muestra un buen rendimiento en términos de identificar robustamente los genes asociados a una condición (Ver Tabla 2). Sin embargo, y advirtiendo el costo computacional que esto podría implicar, idealmente el número de *abstracts* podría ser seleccionado por el usuario final y, tal vez más interesante aún, es que el cruce de palabras seleccionadas con diccionario de genes sea el que rija sobre el conteo de *abstracts*. De este modo, garantizaríamos que los *n abstracts* analizados tengan genes listados. No se implementó puesto que fue difícil predecir el comportamiento en caso de que la búsqueda ingresada en *Pubmed* fuera de una condición donde no hay literatura que asocie a algún gen en particular, en cuyo caso la búsqueda de un número mínimo de *abstracts* que contengan genes podría haber extendido el tiempo de funcionamiento más allá de lo razonable.
2. **Interfaz gráfica**: Sin ninguna duda, esto es parte fundamental de un Proyecto de Uso de Datos que contiene una Interfaz Web. Sabemos que esto es un tema clave al momento de evaluar adherencia y usabilidad de un motor de búsqueda como el que aquí se implementó.
3. **Visualización de resultados**: Este apartado tiene que ver con el formato de los resultados obtenidos. En esta propuesta, el formato de salida es archivos de texto, en formatos convencionales (*.xlsx, *.csv o un archivo *.vcf ya filtrado). Esto puede resultar poco atractivo en comparación con visualizar un archivo de variantes en alguno de los portales mencionados en la Introducción. Sin embargo, hay que recordar que este Producto de Datos apunta a una audiencia que requiere una solución simple y de fácil implementación.

4. **Otros desafíos de código**: Sin duda que futuras versiones de este proyecto debe tener algunos elementos propios de implementaciones web de calidad profesional. Incorporar algunos elementos de verificación que garanticen mayor estabilidad del sistema y generar una guía de instalación de dependencias para que sea fácilmente implementable por el usuario no experto son los próximos pasos por seguir.
5. **Privacidad de los datos**: En una futura versión, para ser compartida de manera pública, se deberá confirmar que los datos fueron accedidos bajo Consentimiento Informado del individuo cuya información genética se está analizando, de acuerdo con la legislación local donde se esté implementando esta solución.

8. Bibliografía

1. Hu, T., Chitnis, N., Monos, D. & Dinh, A. Next-generation sequencing technologies: An overview. *Hum Immunol* **82**, 801–811 (2021).
2. Muir, P. *et al.* The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol* **17**, 53 (2016).
3. Sequence Read Archive (SRA) Data Working Group | DPCPSI. https://dpcpsi.nih.gov/council/sradwg.
4. Levine, A. G. An explosion of bioinformatics careers. *Science (1979)* **344**, 1303–1306 (2014).
5. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature 2001 409:6822* **409**, 860–921 (2001).
6. Belmont, J. W. *et al.* A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
7. 2001: First Draft of the Human Genome Sequence Released. https://www.genome.gov/25520483/online-education-kit-2001-first-draft-of-the-human-genome-sequence-released.
8. Mardis, E. R. A decade's perspective on DNA sequencing technology. *Nature* **470**, 198–203 (2011).
9. DNA Sequencing Costs: Data. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.
10. Human genome at ten: The sequence explosion. *Nature* **464**, 670–671 (2010).
11. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature 2011 470:7333* **470**, 187–197 (2011).
12. Green, E. D. & Guyer, M. S. Charting a course for genomic medicine from base pairs to bedside. (2011) doi:10.1038/nature09764.
13. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30–35 (2010).
14. Centers for Mendelian Genomics. https://www.genome.gov/Funded-Programs-Projects/NHGRI-Genome-Sequencing-Program/Centers-for-Mendelian-Genomics-CMG.

15. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Unlocking Mendelian disease using exome sequencing. *Genome Biol* **12**, 1–11 (2011).

16. Genetic Information | HHS.gov. https://www.hhs.gov/hipaa/for-professionals/special-topics/genetic-information/index.html.

17. Greenbaum, D., Sboner, A., Mu, X. J. & Gerstein, M. Genomics and Privacy: Implications of the New Reality of Closed Data for the Field. *PLoS Comput Biol* **7**, (2011).

18. Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLoS Biol* **13**, (2015).

19. Framework for responsible sharing of genomic and health-related data – GA4GH. https://www.ga4gh.org/framework/.

20. Strategic Plan (2020) – GA4GH. https://www.ga4gh.org/document/strategic-roadmap-2020/.

21. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**, 1767 (2010).

22. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).

23. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

24. Karabayev, D. *et al.* re-Searcher: GUI-based bioinformatics tool for simplified genomics data mining of VCF files. *PeerJ* **9**, e11333 (2021).

25. Salatino, S. & Ramraj, V. BrowseVCF: a web-based application and workflow to quickly prioritize disease-causative variants in VCF files. *Brief Bioinform* **18**, 774–779 (2017).

26. Tollefson, G. A. *et al.* VIVA (VIsualization of VAriants): A VCF File Visualization Tool. *Scientific Reports 2019 9:1* **9**, 1–7 (2019).

27. Müller, H. *et al.* VCF.Filter: interactive prioritization of disease-linked genetic variants from sequencing data. *Nucleic Acids Res* **45**, W567–W572 (2017).

28. Hart, S. N. *et al.* VCF-Miner: GUI-based application for mining variants and annotations stored in VCF files. *Brief Bioinform* **17**, 346–351 (2016).

29. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26 (2011).

30. Cook, D. E. & Andersen, E. C. VCF-kit: assorted utilities for the variant call format. *Bioinformatics* **33**, 1581–1582 (2017).

31. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, (2016).

32. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575 (2007).

33. About - PubMed. https://pubmed.ncbi.nlm.nih.gov/about/.

34. Anaconda Navigator — Anaconda documentation. https://docs.anaconda.com/free/navigator/.

35. Perkel, J. M. Why Jupyter is data scientists' computational notebook of choice. *Nature* **563**, 145–146 (2018).

36. Home - OMIM. https://omim.org/.