UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales Departamento de Química Biológica

Desarrollo y aplicaciones bioinformáticas para el procesamiento y análisis de información genómica humana

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires en el área de Química Biológica.

Germán Biagioli

Director de tesis: Dr. Marcelo Martí

Consejero de Estudios: Dr. Alejandro Nadra

Lugar de trabajo: Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.

Buenos Aires, 2021

Desarrollo y aplicaciones bioinformáticas para el procesamiento y análisis de información genómica humana

Resumen

La medicina del futuro (el futuro ya es hoy) requiere de la capacidad de procesar rápidamente, integrar y analizar cuidadosamente la enorme cantidad de datos que se generan en el plano de la salud. El cuerpo humano comienza a verse entonces como una fuente de datos, comenzando por el genoma humano. En particular, la genómica está cambiando el paradigma de la medicina en sus tres aspectos fundamentales: la prevención, el diagnóstico y el tratamiento, potenciando la revolución de la medicina de precisión. Una de las cuestiones clave para que la medicina de precisión incremente su adopción en la práctica clínica es contar con una herramienta sólida de manipulación y análisis de los datos genómicos que sirva de apoyo a la toma de decisiones mediante el análisis de los datos y presente los resultados en un formato fácilmente interpretable para los profesionales de la salud.

En este contexto, la presente tesis se planteó como objetivo general el desarrollo y diseño de un marco de referencia tanto tecnológico como conceptual para la gestión, procesamiento y análisis de información genómica humana derivada de experimentos de secuenciación (ngs y arrays), para su aplicación a los tres pilares de la práctica clínica, diagnóstico, tratamiento y prevención. Además se planteó realizar una prueba de concepto de su aplicación en casos reales de cada uno de ellos.

Los resultados muestran que fuimos capaces de desarrollar, por un lado, un protocolo bioinformático que permite de manera eficiente y precisa llevarnos desde los datos crudos obtenidos de un experimento de NGS (lecturas) hasta un archivo, que posee un gran contenido de información biológica y clínica. Esta información es la que luego es utilizada por los profesionales para incorporar en la clínica. Para ello, y por otro lado, hemos desarrollado una plataforma de software que permite por parte de los profesionales el análisis profundo y detallado de la información genómica en cualquiera de los tres contextos de interés mencionados (diagnóstico, tratamiento, prevención). Finalmente, los resultados de las pruebas de concepto demuestran que los desarrollos mencionados permiten a partir de datos genómicos, por ejemplo un exoma, llegar a un diagnóstico molecular preciso en el caso de un paciente que posee una sospecha clínica de una enfermedad mendeliana, u obtener un detallado perfil del riesgo relativo de una persona de desarrollar diversas enfermedades complejas. Más aún, el trabajo realizado con los presentes desarrollos en colaboración con diversos profesionales de la salud, muestra la importancia de contar con este tipo de herramientas que facilitan el acceso, son intuitivas y se encuentran disponibles en el ámbito local.

Development and bioinformatics applications for the processing and analysis of human genomic information

Abstract

The medicine of the future requires the ability to quickly process, integrate and carefully analyze the enormous amount of data generated in the field of health. We are beginning to see the body as a source of data, starting with the human genome. In particular, genomics is changing the paradigm of medicine in its three fundamental aspects: prevention, diagnosis and treatment, empowering the precision medicine revolution. A key issue for precision medicine to increase its adoption in clinical practice is to have a robust tool for manipulation and analysis of genomic data that supports decision making through data analysis and presents the results in a format easily interpretable by healthcare professionals.

In this context, the general objective of this thesis was to develop and design a technological and conceptual framework for the management, processing and analysis of human genomic information derived from sequencing experiments (ngs and arrays), for its application to the three pillars of clinical practice, diagnosis, treatment and prevention. In addition, a proof of concept of their application in real cases of each of them was proposed.

The results show that we were able to develop a bioinformatics protocol that allows us to efficiently and accurately take the raw data obtained from an NGS experiment (reads) to a file, which has a high content of biological and clinical information. It is this information that is then used by professionals to incorporate into the clinic. For this purpose, we have developed a software platform that allows professionals to analyze in depth and in detail the genomic information in any of the three contexts of interest mentioned above (diagnosis, treatment, prevention). Finally, the results of the proof-of-concept tests demonstrate that the aforementioned developments make it possible to use genomic data, for example an exome, to arrive at a precise molecular diagnosis in the case of a patient with a clinical suspicion of a Mendelian disease, or to obtain a detailed profile of an individual's relative risk of developing various complex diseases. Moreover, the work carried out with the present developments in collaboration with various health professionals shows the importance of having such tools that are easy to access, intuitive and locally available.

Indice

Introducción

De la genética a la genómica, el legado del Proyecto Genoma humano

La doble hélice del ADN

La palabra polímero proviene del griego poly: muchos y mero: parte o segmento, y se usa para denominar a grandes moléculas formadas por unidades estructurales más pequeñas repetidas, llamadas monómeros.

En biología hay dos polímeros muy importantes: los ácidos nucleicos, cuyos monómeros son los nucleótidos, y las proteínas cuyos monómeros son los aminoácidos que las componen.

Los nucleótidos son los monómeros que conforman la cadena de ADN y el ARN de los seres vivos. Los nucleótidos a su vez, están formados por tres partes: un azúcar, una base nitrogenada (que puede ser adenina → A, timina → T, citosina → C, guanina → G o uracilo → U) y un grupo fosfato que actúa como enganche con el siguiente nucleótido. Los nucleótidos se unen así, formando polímeros que codifican información biológica, los ácidos nucleicos.

Toda célula viva en el planeta almacena su información hereditaria en moléculas de ADN (Ácido Desoxirribonucleico) doblemente enlazadas. Como mostraron en esa hermosa “estructura” Watson y Crick [1] hace más de 60 años, cada molécula de ADN está formada por una cadena de polímeros compuestas de siempre los mismos 4 nucleótidos (A,T,C,G)

De acuerdo a una estricta regla definida por la complementariedad de sus estructuras las bases se asocian de tal manera que A siempre se aparea con T, y C siempre se aparea con G dando lugar de este modo a las dos hebras complementarias que componen una molécula de ADN. Las dos hebras de ADN entrelazadas una sobre la otra forman la doble hélice. [2]

Si bien la complementariedad de las bases es esencial para el proceso de replicación (copia) del ADN, y por lo tanto el paso clave de la transmisión de la información genética -y de la herencia-, desde una perspectiva “computacional” la clave del ADN está en su contenido de información, que es una propiedad emergente - y está almacenada y codificada - en la secuencia lineal de bases a lo largo de una hebra. En otras palabras, el orden de las “diferentes” bases a lo largo de una hebra de una molécula de ADN es lo que se conoce como su secuencia, por ejemplo:

5’ - ACTGTTAGCTAGCTAGCTAGCATCGATCAGCATCAGCTCAGCTGGCAATTTAG - 3’

(Nota: por convención la secuencia del ADN se leen/escriben siempre de 5’ a 3’).

El ADN en una célula no suele ser (salvo organismos unicelulares) una sola molécula larga, sino que se divide en varios segmentos de longitudes desiguales, denominados cromosomas. Los cromosomas pueden ser observados al microscopio óptico cuando la célula se encuentra en proceso de división, y el ADN se compacta y super enrolla dando lugar a los clásicos cromosomas. Cada especie tiene un número determinado de cromosomas. Por ejemplo, los humanos tienen 46 cromosomas (23 pares), las plantas de arroz tienen 24 cromosomas y los perros 78.

El ARN, o ácido ribonucleico, es un ácido nucleico similar en estructura al ADN pero con algunas diferencias sutiles. A diferencia del ADN, el ARN es de cadena simple. Hay diferentes tipos de ARN en la célula: ARN mensajero (ARNm), ARN ribosomal (ARNr) y ARN de transferencia (ARNt). Más recientemente, se han encontrado algunos ARN de tamaño pequeño que están involucrados en la regulación de la expresión génica.

La célula utiliza el ARN para diferentes tareas; por ejemplo el ARNm, es la molécula de ácido nucleico cuya traducción transfiere información del genoma a las proteínas. Otra forma de ARN es el ARNt o ARN de transferencia, y moléculas de ARN no-codificantes de proteínas que físicamente llevan los aminoácidos al sitio dónde se lleva a cabo la traducción y permiten que sean ensamblados en cadenas de aminoácidos para formar las proteínas en dicho proceso.

¿Qué son los genes? ¿Qué es el genoma?

Un ***gen***, en una de sus definiciones más comúnmente utilizadas, es la secuencia de ADN que proporciona a la célula las instrucciones necesarias para producir una proteína específica, que luego llevará a cabo una función particular en el organismo. En otras palabras, *un gen es la secuencia de ADN que **"codifica" para una proteína***. Las proteínas son polímeros lineales de aminoácidos. En el cuerpo humano podemos distinguir unos 20 tipos de aminoácidos que forman las proteínas y son denominados proteinogénicos. El ***código genético humano*** *(Figura 1)* es la tabla de equivalencia entre las secuencias de ARN y de proteínas. Las secuencias de ADN codificantes son leídas de a tripletes (codones) donde cada uno de los codones "codifican" para un aminoácido dado. El *código genético humano* se dice degenerado dado que existen 64 codones posibles para 20 aminoácidos (más de un codón es capaz de codificar el mismo aminoácido). Además, existen 3 codones denominados STP que indican que se ha terminado de "traducir" la proteína. Las proteínas son las principales responsables de realizar las funciones biológicas de un organismo. Las proteínas forman las estructuras del cuerpo como los órganos y tejidos, así como las reacciones químicas que controlan y llevan señales entre las células.

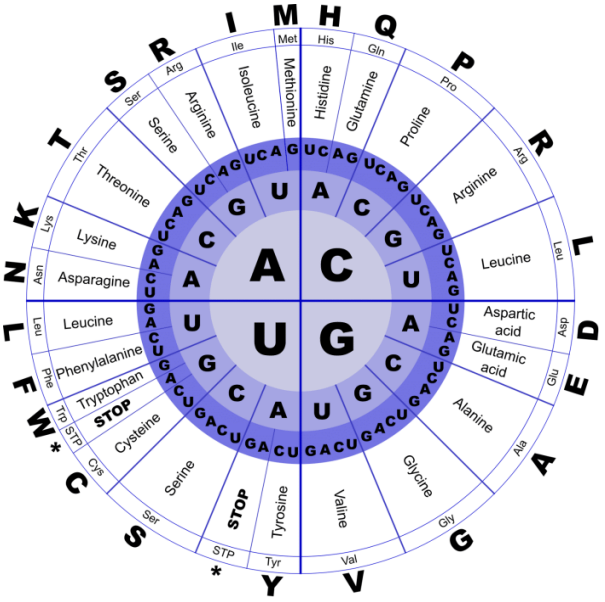

Figura 1: Código genético del ARN.[86]

*El **genoma** es la secuencia del conjunto completo de ADN de un organismo dado*. Todos los organismos vivos tienen su propio genoma. Cada genoma contiene la información necesaria para construir y mantener ese organismo a lo largo de su vida. El mismo contiene no sólo los genes, sino también otras secuencias (no codificantes para proteínas). Algunas de ellas codifican para los diferentes tipos de ARN con variadas funciones biológicas, otras regiones contienen la información para la expresión de los genes y otras cumplen funciones asociadas a la integridad estructural de la molécula de ADN, a los procesos de replicación, etc. Lo importante, es que todas estas funciones dependen de la información que está contenida en la *secuencia* y, *es con esta información con la trabajaremos en esta tesis.*

Todos los individuos de una misma especie poseen un genoma que los define como especie, y en ese contexto hay un genoma de referencia y todos los humanos tienen el mismo genoma de referencia, y por ende los mismos genes dispuestos espacialmente en una misma posición, además de tener más del 99,9% de su secuencia de ADN idéntica a la de cualquier otro ser humano. Aún así, somos diferentes. Estas diferencias son suficientes para cambiar la forma y función de una proteína, cuánta proteína se produce, cuándo se produce o dónde se produce, etc. *Estos cambios, entre otros, influyen en todo lo que somos, y en particular, siendo el foco principal de estas tesis, en diferentes aspectos de nuestra salud*.

¿Qué es la genómica? ¿En qué se diferencia de la genética?

La genética es un término que se refiere tradicionalmente al estudio de los genes y sus funciones en la herencia, en otras palabras, la forma en que ciertos rasgos o condiciones se transmiten de una generación a otra. Una visión más amplia de la genética implica todos los estudios científicos de los genes y sus efectos. Históricamente, los genes, principalmente aquellos asociados a enfermedades genéticas (como la fibrosis quística, la enfermedad de

Huntington, etc.), se estudiaban de a uno. Se buscaba su ubicación (su locus) en el genoma y luego se determinaba su secuencia.

Con el avance de las tecnologías de secuenciación se hizo posible determinar la secuencia, no de solo un gen, sino de todo el genoma de un organismo en un solo experimento. Entonces determinar y luego *estudiar "todo un genoma" dio lugar a una nueva disciplina llamada genómica.*

La genómica es un término más reciente y amplio que describe el estudio de todos los genes de un organismo (el genoma), incluyendo las interacciones entre ellos y con el entorno del mismo. La genómica, de alguna manera, incluye la genética pero también avanza, en el caso de la salud humana, más allá de las enfermedades tradicionalmente llamadas genéticas (técnicamente llamadas Mendelianas, y usualmente asociadas a un solo gen), sobre aquellas enfermedades complejas como las cardíacas, el asma, la diabetes y el cáncer, cuya causa surge de la compleja combinación de factores genéticos y ambientales.

La genómica está, en este contexto y como veremos a continuación, revolucionando la práctica médica, ofreciendo nuevas posibilidades para la prevención, el diagnóstico y tratamiento tanto de las enfermedades mendelianas, como las complejas.

¿Qué es la secuencia del ADN?

El término secuencia de ADN se utiliza para referirse al orden exacto de las bases (o pares de bases) a lo largo de una cadena de ADN. Nótese, que debido a que las bases existen como pares, y la identidad de una de las bases en el par determina al otro miembro del par (por complementariedad), no es necesario que los investigadores reporten ambas bases del par, y se suele simplemente escribir la secuencia de una sola cadena (siempre del extremo 5' al 3'). Por ejemplo, la siguiente, representa una secuencia de ADN

5' - ACTGTTAGCTAGCTAGCTAGCATCGATCAGCATCAGCTCAGCTGGCAATTTAG - 3'

La secuencia de ADN, como cualquier secuencia de caracteres, es una entidad capaz de almacenar información. De este modo del genoma de un organismo, podemos decir que es la secuencia de ADN que contiene toda la información necesaria para crear a ese organismo y darle un conjunto de instrucciones para que interactúe y responda a su entorno. En otras palabras, para que viva. Es por ello que conocer la secuencia del genoma de un organismo es tan importante, y por ello desde la década del 80, hemos estado determinando cada vez con mayor detalle la secuencia de nuestro(s) genomas(s).

El proyecto Genoma humano y su legado.

El Proyecto Genoma Humano (HGP, por su sigla en inglés, Human Genome Project) [3] fue el esfuerzo de un equipo internacional de investigación para determinar la secuencia de ADN de todo el genoma humano. Los científicos que lo llevaron adelante, buscaban secuenciar y mapear todos los genes de los miembros de nuestra especie, el *Homo sapiens*. El Consorcio Internacional de Secuenciación del Genoma Humano (International Human Genome Sequencing Consortium) [4] publicó el primer borrador del genoma humano en la revista Nature en febrero de 2001 con la secuencia de los tres mil millones de pares de bases del genoma humano completo en un 90%. Desde allí en adelante, se ha

continuado determinando cada vez un mayor porcentaje del mismo, con mayor precisión y detalle, y más recientemente, prestando atención a la varianza entre los individuos (o sea la varianza poblacional).

El HGP ha revelado que probablemente haya aproximadamente unos 20.500 genes humanos. Este producto final del HGP ha dado al mundo un recurso de información detallada sobre la estructura, organización y función del conjunto completo de genes humanos. Esta información puede considerarse, como ya mencionamos, como el conjunto básico de "instrucciones" heredadas para el desarrollo y la función de un ser humano.

Tras la publicación de la primera versión del genoma en febrero de 2001, Francis Collins, el entonces director del Instituto Nacional de Investigación sobre el Genoma Humano [5], señaló que el genoma podría considerarse en términos de un libro con múltiples usos: *"Es un libro de historia - una narración del viaje de nuestra especie a través del tiempo. Es un manual, con un plano increíblemente detallado para construir cada célula humana. Y es un libro de texto transformador de la medicina, con conocimientos que le darán a los proveedores de atención médica inmensos poderes para tratar, prevenir y curar enfermedades".*
Los protocolos, herramientas y técnicas desarrolladas para la realización del HGP sentaron las bases de la genómica y se expandieron rápidamente a otros organismos, como por ejemplo los ratones, las moscas de la fruta y los gusanos planos. Estos esfuerzos se apoyan mutuamente, porque la mayoría de los organismos tienen muchos genes similares, u "homólogos", con funciones similares. Por lo tanto, la identificación de la secuencia o función de un determinado gen en un organismo modelo, por ejemplo, tiene el potencial de explicar un gen homólogo en seres humanos, o en uno de los otros organismos.

Desde una perspectiva técnica, el genoma humano, representa un mapa que nos permite ordenar el funcionamiento de nuestro organismo desde una perspectiva molecular. Tal es así, que luego de determinar su secuencia, el esfuerzo de la comunidad biológica pasó a ser determinar y comprender el funcionamiento de los diferentes segmentos o componentes del mismo (genes, regiones regulatorias, regiones estructurales, etc.). Visto de otro modo, podemos decir que una parte importante de la biología lo que busca es comprender qué significa cada una de las partes del genoma humano.

Por otro lado, es importante destacar, que una vez determinado genoma humano de referencia, el objetivo pasó a ser determinar cómo el genoma de un individuo difiere del que hoy consideramos “de referencia”. Más aún, en el contexto de la salud, el objetivo consiste en determinar cómo es que estas diferencias (denominadas variantes) afectan nuestra salud.

¿Qué es una variante?

Cómo mencionamos en el párrafo anterior, la versión pública del genoma humano, es lo que actualmente tomamos como *genoma humano de referencia.* El mismo, no representa la secuencia del genoma de un individuo particular, sino que es un consenso que, valga la

redundancia, se utiliza como mapa de referencia para describir las diferencias que presentan las secuencias de los genomas de individuos particulares respecto de la misma.

El genotipo de un organismo es el conjunto de sus genes. Mientras que su fenotipo son todas sus características observables - que son influenciadas tanto por su genotipo como por el medio ambiente. El fenotipo puede referirse a cualquier aspecto de la morfología, comportamiento o fisiología de un organismo. A veces, cuando se utiliza la palabra genotipo se hace referencia a todo el genoma de un organismo y otras veces solo se refiere a los alelos de una posición particular del genoma.

Las diferencias en la secuencia de ADN presentes en un individuo respecto del genoma de referencia, se denominan variantes.

La mayoría de las veces las variantes no tienen ningún efecto significativo sobre el fenotipo. Pero, a veces, el efecto es perjudicial, y sólo una letra que falte o que haya cambiado puede resultar en una proteína dañada, la ausencia de la proteína, o el cambio en la cantidad producida de la misma, lo que puede traer graves consecuencias para nuestra salud.

Entonces en función de la definición de variante y teniendo en cuenta que, como mencionamos anteriormente, todos los seres humanos tenemos más del 99% de nuestra secuencia de ADN idéntica a la de cualquier otro ser humano, podemos decir que ***el genoma de un individuo es (o se representa como) una lista de variantes, respecto del genoma de referencia****.*

Las variantes, se pueden clasificar en diversos tipos. Por ejemplo, una variante que corresponde a la sustitución de una citosina (C) por una timina (T), es lo que se denomina una variante de nucleótido único (SNV del inglés). En cambio una variante que agregue, por ejemplo 4 bases, ACTT, se la llama inserción; mientras que aquella que corresponda a la pérdida de una o más bases se la denomina *deleción*. Por otro lado, a veces las variantes corresponden a cambios en segmentos grandes de un cromosoma (de miles de pares de bases) y se denominan usualmente rearreglos estructurales. En particular, cuando grandes segmentos son eliminados o duplicados, se las denomina alteraciones en el número de copias (CNV, del inglés Copy Number Variations) ya que usualmente alteran el número de alelos (copias de uno o más genes).

Las variantes se producen normalmente a lo largo de todo el genoma de una persona. En promedio, cada persona tiene entre 3 y 5 millones de variantes en su genoma, la mayoría de las cuales, por lo general, no tienen ningún efecto en la salud (o eso se supone). La mayoría de estas variantes se encuentran en el ADN entre genes (variantes intergénicas). Cuando las variantes se producen dentro de un gen o en una región reguladora, tiene un impacto mayor ya que pueden afectar la función del gen y suelen desempeñar un papel más directo en nuestra salud.

Alelos y frecuencia del alelo minoritario

Podemos definir a los alelos, como a las formas alternativas de un gen (o simplemente un par de bases específicas del genoma sin función definida y que denominaremos marcador)

que pueden ocurrir en un mismo locus, (posición dentro del genoma). Es decir, cada alelo representa una de las alternativas de un segmento de secuencia de ADN en un lugar particular de un cromosoma.

Relacionando este concepto con el de variante, es claro que una variante es necesariamente un alelo alternativo, siendo la secuencia del genoma de referencia el “alelo de referencia”. En cierto sentido, variante y alelo son sinónimos. Lo importante es que el concepto de alelo tiene una larga tradición de uso en genética/genómica de poblaciones. Cuando el mismo alelo es observado en diversos individuos de una población, uno puede calcular su frecuencia. El alelo de mayor frecuencia suele ser la referencia, y por ello nos referimos al alelo de menor aparición, como al menos frecuente.

Basado en el concepto de alelo menor, hablamos de variante rara, cuando se trata de una variante que ocurre en menos del 0,5% de la población. Llamamos variantes de baja frecuencia, aquellas que ocurren entre el 0,5% y el 5% de la población. Y las variantes más comunes son aquellas que ocurren en más del 5% de la población.

El término polimorfismo significa técnicamente una mutación que ha alcanzado una frecuencia alélica superior al 1% de la población. Muchos autores usan el término mutación para cualquier alelo raro, y el término polimorfismo para cualquier alelo común (o sea con frecuencia superior al 5%).

Una serie de diferentes factores influyen en las frecuencias alélicas menores de una población, como puede ser la selección natural, es decir, si la mutación es deletérea o perjudicial y no se transmite de generación en generación o si es beneficiosa, tal vez aumente su frecuencia. También influye la migración de personas de una población a otra, llevando su acervo genético de un lugar a otro.

Diferentes tipos de variantes

Las variantes se suelen clasificar (o caracterizar), primero, según el impacto que puedan producir a nivel molecular.

Variantes ***sinónimas o silenciosas***, son aquellas mutaciones cuyo cambio de nucleótidos en el codón, no producen cambios en el aminoácido resultante. No tienen ningún efecto en la proteína resultante.

En contraposición, si el cambio introducido produce un aminoácido diferente en nuestra proteína, estos tipos de variantes se llaman mutaciones ***no sinónimas o missense***.

Si ocurre una mutación no sinónima y el cambio de nucleótido introducido produce que se genere una señal de stop en el proceso de traducción, de tal manera que en lugar de tener la proteína de longitud completa, la proteína queda truncada, o sea termina prematuramente, la variante se conoce como de ganancia de stop, stop prematuro o ***stop gained***. De manera similar, si el cambio de nucleótido produce que se pierda la señal de stop en el proceso de traducción, esta variante se denomina ***stop loss.*** Otra variante que también afecta el proceso de traducción ocurre cuando debido a la inserción o deleción de uno, o un par, de nucleótidos se produce un desplazamiento del marco de lectura, lo que se llama un ***frameshift.***

Finalmente, la expresión de un gen también puede verse afectada por variantes que ocurren en regiones intrónicas, ***intron variant***, lo que puede afectar el proceso de splicing de los genes produciendo distintas expresiones de un gen.

Si la variante se encuentra en regiones río arriba del gen, esta se denomina ***upstreams variant o 5' UTR variant***. De igual manera pueden ocurrir mutaciones en regiones regulatorias, regiones promotoras de la expresión de los genes, que si bien se encuentran en zonas alejadas del gen, cambios en la misma lo pueden afectar. Por ejemplo, las ***intergenic variants***.

Predicción del efecto de una variante

Además de conocer la ubicación de una variante y su efecto a nivel molecular, nuestro objetivo último es determinar qué efecto produce en el fenotipo, es decir, a nivel molecular *¿Qué significa para la función del gen/proteína que estamos analizando?* y a nivel organismo qué consecuencias fisiopatológicas produce.

Predecir el efecto (sobre todo a nivel organismo) de una variante, en un gen o proteína, puede ser bastante difícil. La complejidad del problema se debe principalmente a la heterogeneidad de los mecanismos moleculares que subyacen a las enfermedades, la mayoría de los cuales aún deben ser entendidos en profundidad para que sean útiles a nivel clínico.

Es importante destacar que *el impacto de una determinada variante depende principalmente de su contexto dentro del genoma*: su efecto en el organismo puede depender de la presencia de otras variantes en el mismo o en otros genes. Su efecto también varía según el contexto celular. De hecho, algunas variantes se producen en proteínas que desempeñan un papel esencial para la célula o el organismo, que no puede ser realizado por ninguna otra proteína. En tal caso, incluso una variante ligeramente desestabilizadora puede ser fuertemente deletérea.

Entonces podemos analizar el efecto de una variante desde dos perspectivas: la primera analizando la propia variante de manera independiente, y la segunda realizando un análisis más comprensivo analizando el entorno de la misma y otros datos clínicos.

Algunas de las características que podemos observar en una variante de manera independiente para tratar de predecir su efecto son: *la frecuencia alélica, el efecto de la variante sobre la proteína, la conservación evolutiva, es decir, que tan conservada está esa posición, así si el cambio de aminoácido afecta a una de estas posiciones altamente conservadas, es más probable que sea perjudicial, no significa que deba serlo, pero es una evidencia circunstancia*l. También se puede analizar la estructura de la proteína, propiedades fisicoquímicas de los aminoácidos involucrados en el cambio, las características funcionales de la misma a partir de estudios funcionales in vitro o in vivo.

Desde el punto de vista del organismo y en el contexto de una posible enfermedad se puede analizar el modo de herencia para ver si el genotipo de la variante se corresponde. Se puede analizar la segregación de la variante en los familiares del paciente. También se busca si existe asociación previa del gen con el diagnóstico presuntivo. La variante puede ser *de novo* pero si se determina que la variante es deletérea debido a su efecto sobre la proteína, y además, la enfermedad ha sido observada en otros individuos, con otras

variantes deletéreas en ese mismo gen, eso nos daría más evidencia de que esta variante está causando la enfermedad en el paciente.

Se pueden estudiar las vías metabólicas, las relaciones de los genes candidatos con otros genes. Se puede observar si un gen en particular, se encuentra dentro de una vía, donde otros genes, en esa vía, se han encontrado asociados con la enfermedad. Esto es, en esencia, una ***culpa por asociación***, que puede ser muy útil.

Clasificación de las variantes

Entonces, si nos focalizamos primero en el efecto de la variante sobre la función del gen/proteína aislada, se puede clasificar la variante en:

Probablemente perjudicial o dañina (Probably damaging) si por ejemplo:

- Nos encontramos con una terminación prematura (stop gained) en la expresión de la proteínas o si se pierde un codón de parada (stop loss).
- Si se produce un corrimiento en el marco de lectura (frameshift), lo que producirá un stop codon generalmente un número de bases río abajo (downstreams) de donde está el desplazamiento del marco de lectura.
- Si la variante es intrónica y altera el sitio de splicing puede pasar que un exón crítico sea excluido o no es retenido apropiadamente dentro del cDNA.

Posiblemente perjudicial o dañina (Possibly damaging): el efecto quizás pueda no ser tolerado:

- Si el cambio de aminoácidos es no sinónimo altera un aminoácido importante.
- Si la variante es una inserción o deleción podría también ser posiblemente perjudicial.
- Si el cambio de aminoácido desestabiliza la estructura de la proteína.

Probablemente benigna: generalmente es benigna si la mutación

- afecta a las regiones UTR (untranslated region).
- es sinónima.
- si ocurre entre genes, intergénicas o si es intrónica, cuando está en una región del intrón que no afecte el sitio de splicing.
- si la mutación ocurre en pseudogenes.

Todas estas afirmaciones sobre sí una variante es probable o posiblemente perjudicial o si es probablemente benigna tienen muchas excepciones. Ninguna de ellas es completamente determinante.

Una vez que se ha reunido toda esta información, se intenta llegar a una evaluación final única, de lo que esta variante está haciendo en el individuo en relación con su salud. Así podemos decir que la variante es ***patogénica*** si estamos seguros de que causa la enfermedad.

En cambio si sospechamos fuertemente que la variante es patogénica, pero no tenemos un nivel de evidencia suficiente para estar absolutamente seguros, decimos que es ***probablemente patogénica***.

Decimos que una variante es de ***significado incierto (VOUS o VUS)*** cuando no podemos llegar a ninguna conclusión, no hay evidencia suficiente.

Como se puede inferir, en la genómica clínica, la determinación precisa de la patogenicidad de una variante es una tarea compleja. Durante los últimos años se han desarrollado diversos métodos semi-automáticos para facilitar el proceso de asignación de criterios, combinando los distintos tipos de evidencia, que permiten realizar la clasificación de las variantes descritas anteriormente.

En 2015, el American College of Medical genetics (ACMG) y la Asociación de Patología Molecular (AMP), publicó lo que se conoce usualmente como “las recomendaciones de la ACMG/AMP” [6], un conjunto sistemático de reglas para valorar y combinar la evidencia de cada variante, que permite clasificarla de acuerdo a su patogenicidad.

Estas recomendaciones de la ACMG se han convertido en el standard de facto en la clínica a la hora clasificar las variantes en patogénicas, probablemente patogénicas, de significado incierto y benignas, y es por ello que en esta tesis analizaremos los resultados en este contexto y utilizaremos estas guías para la clasificación.

Más adelante dedicaremos una sección específica a los métodos de valoración automática de las variantes de acuerdo a los criterios de la ACMG, ya que son una herramienta poderosa para la genómica clínica que se encuentra en pleno desarrollo y permite mejorar de manera significativa la eficiencia del análisis de datos genómicos.

En resumen, durante las últimas décadas la genómica ha buscado establecer el rol biológico de los millones de variantes que se han descrito en la humanidad.

El gran desafío de la genómica aplicada a la clínica es determinar cuáles de las variantes del individuo que se está estudiando son responsables y/o contribuyen al problema de salud que está bajo estudio (o sea cuáles son patogénicas). Pero antes de profundizar en este tema, veamos cómo se determinan hoy en día las variantes.

Generaciones de tecnologías de secuenciación

Secuenciación de ADN. Desde Sanger a las tecnologías de Secuenciación de Próxima Generación.

Como mencionamos anteriormente, la secuenciación de ADN, consiste en determinar el orden correcto de las bases o nucleótidos en una cadena de ADN. Las primeras técnicas de secuenciación se desarrollaron en la década del 80, fue Frederick Sanger, quien desarrolló la técnica que lleva su nombre, la secuenciación de Sanger [7] y que aún hoy en día es el “gold standard” para secuenciar ADN, y por la cual recibiera su segundo premio nobel (el primero lo recibió por determinar la estructura de la insulina).

En 2005, y fruto del proyecto genoma humano, se produjo un enorme avance tecnológico en los métodos de secuenciación, que dio lugar a la tradicionalmente denominada “Secuenciación de Próxima Generación” (NGS, Next Generation Sequencing). Más recientemente, nuevos avances asociados a la secuenciación en molécula única llevaron a reorganizar la nomenclatura, de acuerdo a lo que describiremos brevemente a continuación.

Tecnologías de secuenciación de primera generación: El método de Sanger.

El método de Sanger se conoce como el método del didesoxinucleótido o secuenciación por síntesis [8]. Consiste en usar una hebra del ADN cadena como molde y utilizar el ADN polimerasa para sintetizar todos los fragmentos posibles de diferente longitud de la cadena complementaria. Estos fragmentos, luego son separados por tamaño, y el conocimiento (identificación) de su última base contiene la información de la secuencia (ver figura 2.a).

La clave, de la secuenciación por Sanger, se encuentra en la utilización de nucleótidos químicamente modificados llamados dideoxinucleótidos (ddNTPs), que en la posición 3' no poseen el grupo OH. Esto inhibe el proceso de síntesis posterior a su adición por parte de la polimerasa y permite generar los fragmentos mencionados.

Los ddNTPs entonces se utilizan para interrumpir el proceso de elongación de nucleótidos. La elongación consiste en el agregado de nucleótidos en forma secuencial tras el agregado de la primera base, hasta que la polimerasa alcance el fin del molde a transcribir. Una vez incorporado el ddNTP en la cadena de ADN que se está transcribiendo la polimerasa interrumpe la elongación, entonces obtenemos fragmentos de ADN de diferentes tamaños terminados siempre por un ddNTP.

En la versión original del método, se realizaban cuatro reacciones de síntesis, cada una con uno de los ddNTP diferente, y los fragmentos generados eran separados en un gel de 4 calles, lo que permitía luego determinar la secuencia. Desarrollos posteriores llevaron a la marcación de cada uno de los ddNTP con un color fluorescente diferente, lo que permite hacer una sola reacción con los 4 ddNTP, y la electroforesis en gel fue reemplazada por electroforesis capilar.

La secuenciación de Sanger fue ampliamente utilizada durante tres décadas (fue la que se utilizó para el genoma humano) e incluso hoy en día es usado para casos puntuales, como un método de validación de mutaciones individuales, sin embargo, es difícil mejorar la velocidad de análisis, lleva mucho tiempo y no permite la secuenciación de genomas complejos.

Tecnologías de segunda generación

A partir de 2005 y en los años siguientes, ha surgido una nueva generación de secuenciadores para romper las limitaciones de la primera generación en cuanto al costo y el tiempo de análisis. Actualmente es la tecnología que más se utiliza en investigación y consiste en fragmentar el genoma en segmentos relativamente pequeños, amplificar estos fragmentos para luego secuenciar millones de ellos en paralelo, de forma masiva, usualmente por secuenciación por síntesis. Cada reacción de secuenciación, es realizada en un volumen muy pequeño lo que reduce el costo en reactivos. La detección, se realiza paso por paso, cuando un nucleótido es incorporado, ya sea mediante una reacción lumínica (equipos Roche o PGM), por cambio de pH (equipos ION) o por fluorescencia (Illumina). En esta tesis utilizamos principalmente datos de Illumina, por lo cual a continuación describiremos brevemente esta tecnología.

Los equipos de NGS más utilizados en la actualidad a nivel mundial son los de la compañía Illumina [9] en sus diferentes versiones según tamaño y capacidad, MySeq, NextSeq, HiSeq y NovaSeq. La técnica de secuenciación utilizada por estos equipos se basa en una paralelización masiva y miniaturización, lo que permite que millones de “lecturas” se secuencien simultáneamente [10].

El proceso de secuenciación de una muestra de ADN requiere de varias etapas. El primer paso en el proceso es la ***preparación de la biblioteca***, donde las muestras de ADN se fragmentan aleatoriamente en secuencias más cortas, se desnaturalizan en moléculas de una sola cadena y se ligan adaptadores a ambos extremos de cada fragmento. Estos adaptadores son los que se unen a los respectivos adaptadores complementarios que se encuentran fijados en un soporte sólido denominado “celda de flujo” (Flow cell) y que permiten al equipo procesarlos.

Durante el segundo paso, cada hebra de ADN sujeta en la celda de flujo es amplificada por "PCR bridge amplification” proceso que crea varias copias idénticas de cada secuencia formando ***clusters locales***. Cada cluster contiene millones de copias de la misma secuencia original (esta amplificación clonal es lo que permite aumentar la señal para que esta sea luego detectada). Cada cluster se secuenciará de manera independiente dando cada uno lugar a una “lectura”.

El tercer paso consiste en el proceso de ***secuenciación*** propiamente dicho (Figura 2.b), después de realizar la amplificación por bridge PCR de cada fragmento, se añaden los cuatro tipos (A/T/C/G) de nucleótidos que son nucleótidos de terminación reversibles (dNTPs), cada uno marcado fluorescentemente con un color diferente, primers genéricos y la ADN polimerasa.

En cada ciclo de secuenciación, los cuatro nucleótidos compiten entonces por los sitios de unión en cada fragmento de ADN a ser secuenciado. Cuando la polimerasa alarga la cadena con un dNTP etiquetado fluorescentemente, los grupos fluorescentes son excitados por una fuente de luz (láser) y el color es registrado por un detector óptico, permitiendo la identificación de la base incorporada en cada una de las lecturas. Nótese que al ser nucleótidos terminadores sólo es incorporado un nucleótido en cada clúster.

Después de que ocurre la incorporación de un dNTP terminal, el fluoróforo es escindido y el nucleótido re-activado, permitiendo así que el siguiente nucleótido se pueda incorporar a la cadena en el siguiente ciclo de síntesis e identificación. El proceso se repite unas 100 - 150 veces (lo que da longitud final de la lectura).

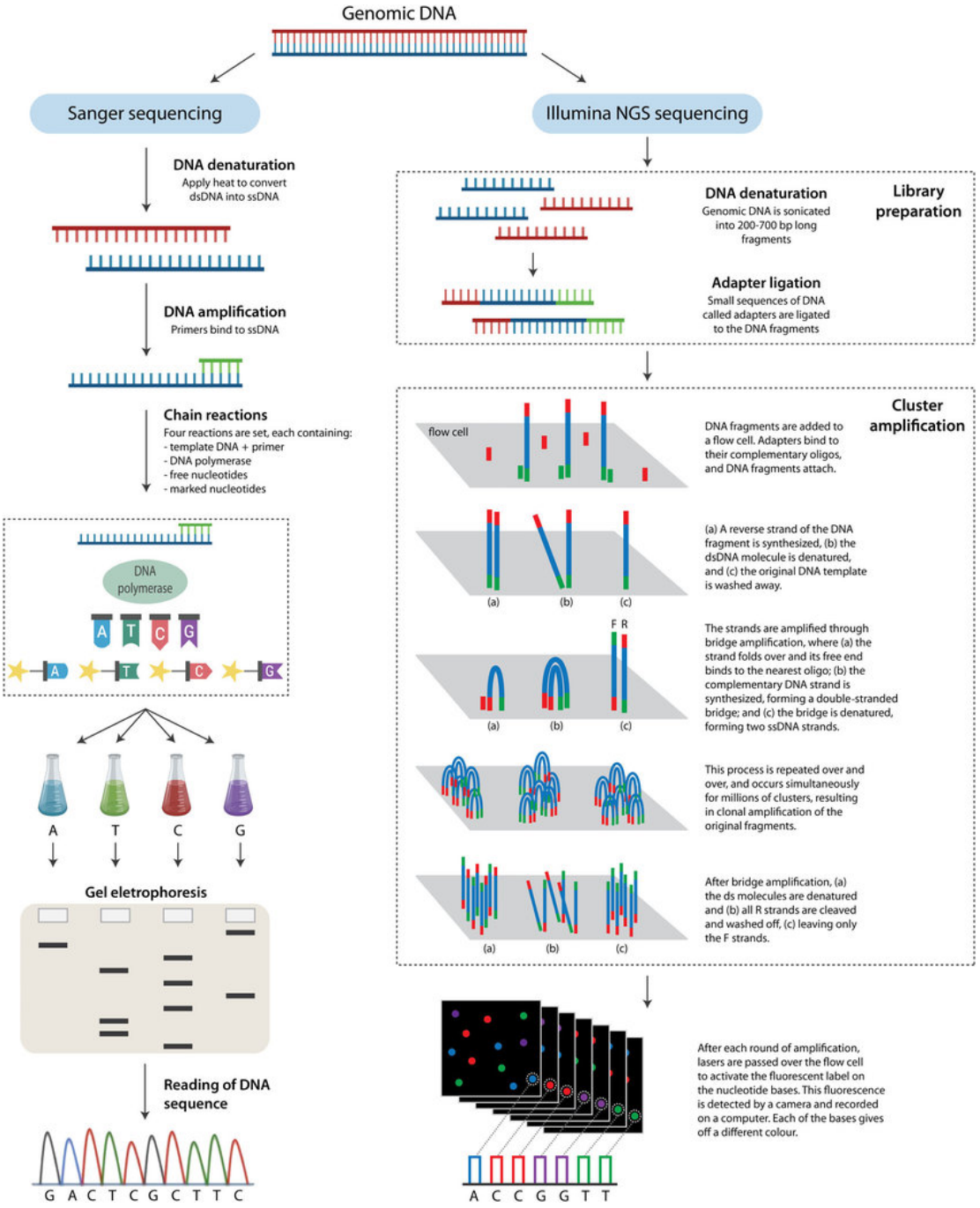

Figura 2: Comparación de a) la secuenciación Sanger (izquierda) y b) la secuenciación de próxima generación Illumina (derecha). [85]

Tecnologías de tercera generación

Las tecnologías de segunda generación en general requieren un paso de amplificación de PCR que es un procedimiento costoso tanto en tiempo como en insumos. Además, se hizo evidente que muchos genomas - especialmente los de plantas - son complejos, con muchas áreas repetitivas en las que las tecnologías de segunda generación debido a que leen lecturas cortas, son incapaces de resolver. La tercera generación de secuenciadores busca resolver ambos problemas.

Hay principalmente dos enfoques comerciales disponibles hoy en día que caracterizan a las tecnologías de tercera generación: el primero, se basa en la secuenciación por síntesis, utilizando detección fluorescente, pero utilizando como sustrato una molécula única de ADN. Esta técnica, es comercializada por Pacific Biosciences y permite obtener lecturas largas (de miles de pb) a una velocidad significativamente superior a la de las técnicas de 2da generación. En segundo lugar, la secuenciación por Nanoporos de Oxford nanopore, se basa en la lectura directa de una hebra de ADN a medida que la misma pasa a través de un poro.

Las tecnologìas de tercera generación, si bien poseen enorme capacidad en cuanto a su volumen de secuenciación y particularmente su capacidad de producir lecturas largas, aún se encuentran subutilizadas en la clínica, dado que su tasa de error ronda entre el 1-5% lo que es significativamente superior a los equipos Illumina. Sin embargo, con el correr del tiempo, la misma disminuye de manera sostenida, lo que sugiere un futuro más que promisorio para las mismas.

Genoma - Exoma - Paneles - Gen individual. Esa es la cuestión.

Independientemente de la tecnología de secuenciación elegida (aunque no tanto) uno de los puntos clave en relación con la aplicación de las mismas en salud humana, consiste en determinar qué porción del genoma es la que se va a secuenciar. Con las tecnologías de NGS, se puede obtener la secuencia completa del ADN de una persona, es decir, todo el genoma, a esto se lo denomina ***Whole Genome Sequencing (WGS),*** o se puede solo obtener la secuencia de ADN de “todos” los genes, es decir, solo secuenciar la regiones que codifican proteínas, a este procedimiento se los llama ***Whole Exome Sequencing (WES)***. Así como se puede secuenciar todos los genes de un individuo, se puede seleccionar para secuenciar un determinado grupo de genes de interés, usualmente asociado a la patología de interés, a esto se los llama secuenciación de un ***Panel de Genes***. Finalmente, siempre es posible secuenciar un sólo gen, o incluso un solo exón o región de interés.

La elección de qué secuenciar estará determinada por el tipo de análisis o investigación que se esté realizando, aunque también por el presupuesto con el que se cuente, ya que la secuenciación por NGS se paga por letra secuenciada. La elección del estudio a realizarse finalmente es una decisión del médico o profesional de la salud, que pesará los distintos factores que entran en juego.

Es importante destacar, que la selección de qué se va a secuenciar (WGS/WES/Panel) se realiza en el paso previo a la secuenciación, en lo que se conoce como “preparación de biblioteca”. En este paso, el ADN purificado primero es fragmentado y luego se selecciona y/o amplifica, qué es lo que se desea secuenciar. Las técnicas de captura, usualmente se basan en la utilización de sondas que se encuentran acopladas a una “bola” magnética y que “capturan” los fragmentos que contengan las secuencias de interés. En el caso de que la biblioteca se prepare por amplificación, se utilizan primers específicos y la amplificación se realiza por PCR. Una vez preparada la biblioteca, la misma es incorporada al secuenciador, dónde los fragmentos son amplificados clonalmente y leídos obteniéndose las lecturas correspondientes.

Aplicaciones de las tecnologías de secuenciación masiva en Genómica Clínica

Fruto del *Proyecto Genoma Humano* [3] y de los avances tecnológicos mencionados en los párrafos anteriores, hoy en día somos capaces de determinar la secuencia del genoma completo (o parte) de un individuo a un costo más que accesible. En este contexto, el desafío para los investigadores consiste ahora en aplicar las mismas en la práctica clínica. El análisis personalizado basado en el genoma de cada persona (definido a partir de sus variantes) promete una medicina personalizada y de precisión, que como describiremos a continuación se espera produzca un impacto significativo en el diagnóstico, el tratamiento y la prevención de numerosas enfermedades.

Diagnóstico. El primer impacto de las tecnologías de secuenciación masiva en la salud, comprende posiblemente el diagnóstico de las Enfermedades Poco Frecuentes (EPOFs) Prácticamente todas las enfermedades humanas poco frecuentes tienen alguna base en nuestros genes. Hasta hace poco, los médicos podían tener en cuenta el estudio de los genes, o la genética, sólo en casos de defectos congénitos y un conjunto limitado de otras enfermedades. Éstas eran condiciones que tienen patrones hereditarios muy simples y predecibles porque cada una es causada, en principio, por una variante en un solo gen. A estas enfermedades, usualmente EPOFs, crónicas y altamente discapacitantes, se las conoce como *mendelianas*, y el efecto predictivo de la presencia de variantes patogénicas en relación con el desarrollo de las mismas, es conocido y determinante. Entonces, si un individuo posee una (o más) variantes patogénicas asociadas a una enfermedad mendeliana, tenemos un alto grado de confianza en que desarrollará la enfermedad. El corolario de esta propiedad, es que cuando nos enfrentamos a un paciente con un diagnóstico clínico de una enfermedad mendeliana, determinar cuál (o cuáles) son las variantes potencialmente responsables del desarrollo de la misma, es lo que se considera un diagnóstico molecular preciso.
En este contexto, y en directa relación con la presente tesis, uno de los principales desafíos de la genómica clínica, consiste en una vez obtenido el genoma de un individuo que posee una sospecha clínica de una enfermedad “presuntamente” mendeliana, determinar cuál (o cuáles) de todas las variantes es la responsable del fenotipo observado.

Tratamiento. Una vez obtenido un diagnóstico, ya sea molecular, clínico o ambos, se suele pasar al tratamiento, si es que existe uno disponible. Los tratamientos farmacológicos, se basan en la utilización de fármacos, tradicionalmente moléculas orgánicas que interfieren con la función de alguna biomolécula (su blanco molecular), típicamente una proteína, y mediante esta interacción, ejercen su acción terapéutica. Dentro del cuerpo, los fármacos no interactúan solamente con su blanco, sino que muchas veces poseen interacciones con otras biomoléculas (las denominadas interacciones off-target), dando lugar a diferentes efectos, muchas veces no deseados, más aún en ocasiones los fármacos deben ser procesados por enzimas del organismo antes de alcanzar su estructura activa, y la gran

mayoría de estos son metabolizados para luego ser excretados. Los fármacos, en general, son metabolizados por las enzimas hepáticas, los citocromos de tipo p450, que los oxidan, haciéndolos más solubles para facilitar su excreción por parte del riñón.

Tanto las proteínas blanco, las off-target y los citocromos p450 son productos de diversos genes, y como tal, poseen variabilidad poblacional, que afecta como cada organismo metaboliza y responde al tratamiento farmacológico correspondiente. Esta disciplina que estudia cómo las variantes genéticas influyen en nuestra respuesta a los medicamentos se la conoce como ***farmacogenómica***, y lo que busca es determinar cuáles y cómo las variantes del genoma de un individuo afectan su capacidad de respuesta a un fármaco determinado. En este sentido su aplicación en la clínica es directa, cuando el profesional médico se encuentra frente a uno o más posibles tratamientos, el genoma del individuo le podrá decir al médico cómo será la respuesta del paciente a cada uno y por ende ayudará a decir cuál es el más adecuado, dando lugar a un tratamiento personalizado y preciso.

Prevención. Si bien la relación no es directa como en el diagnóstico de enfermedades mendelianas, la genómica también contribuye al potencial -o riesgo- de desarrollar enfermedades complejas de alta prevalencia en el adulto, como ser las cardiovasculares, la diabetes tipo 2, la obesidad, el alzheimer o el cáncer entre otras. Los estudios que relacionan el riesgo de desarrollar este tipo de patologías con las variantes del genoma se denominan estudios de asociación genómica (GWAS, del inglés Genome Wide Association Studies). En los GWAS lo que se busca es asociar la presencia de una variante dada, con un riesgo incrementado de desarrollar -en el futuro- cierta enfermedad. Una vez determinada esta asociación y su potencia, -comúnmente denominada tamaño del efecto (OR del inglés Odds Ratio)- esta información puede ser utilizada en el contexto de un genoma personal, para evaluar la contribución de la genética al riesgo del individuo de desarrollar estas patologías en el futuro. Esta información, por supuesto, debe ser integrada con diferentes aspectos relacionados con la dieta, el estilo de vida y el ambiente en general del individuo.

En resumen, el genoma humano, que ha dado lugar a las tecnologías de secuenciación masiva, nos ha dejado a las puertas de una revolución en la práctica médica, basada en la capacidad de determinación del genoma de cada individuo, descrito técnicamente como un listado de variantes respecto del genoma humano de referencia, siendo el desafìo de la genómica clínica determinar cuál de éstas variantes posee relevancia para nuestra salud en alguno de sus tres aspectos fundamentales, prevención, diagnóstico y tratamiento.

Bioinformática como necesidad para NGS

Desafíos bioinformáticos en la era genómica

Datos y más datos

Con el auge de las técnicas de secuenciación masiva, la velocidad y capacidad de generación de datos se ha disparado exponencialmente. El uso cada vez más extendido de

estas tecnologías conlleva la generación y el manejo de grandes volúmenes de datos que implican, no solo soluciones algorítmicas de análisis y visualización de datos, sino problemáticas de big data, storage, data transfer y seguridad que requieren soluciones. En los últimos 10 años la cantidad de datos a procesar crece a un ritmo superior que la velocidad a la cae el costo de secuenciación, ya que no solo secuenciar es más barato en cada secuenciador, sino que además hay cada vez más secuenciadores disponibles.

Para ilustrar la naturaleza de la reducción en los costos de secuenciación del ADN, el gráfico de la figura 3 muestra, por un lado (trazo azul) datos hipotéticos que reflejan la Ley de Moore, que describe una tendencia a largo plazo en la industria del hardware informático que señala la duplicación de la potencia de cálculo cada dos años. Los avances tecnológicos en los métodos de secuenciación tienen un comportamiento similar a la ley de Moore, lo que permite hacer la comparación; y por otro los costos de secuenciación (usando como métrica la secuenciación de un genoma humano) de los últimos 20 años (en verde). El gráfico se realiza en escala semilogarítmica.

La caída abrupta del costo de secuenciación evidente a partir de enero de 2008 -que sobrepasa con creces lo esperado por la ley de moore y es sobre exponencial- representa el momento en que los centros de secuenciación pasaron de las tecnologías de secuenciación de ADN basadas en Sanger (primera generación) a las tecnologías de secuenciación de ADN de segunda generación o de próxima generación.

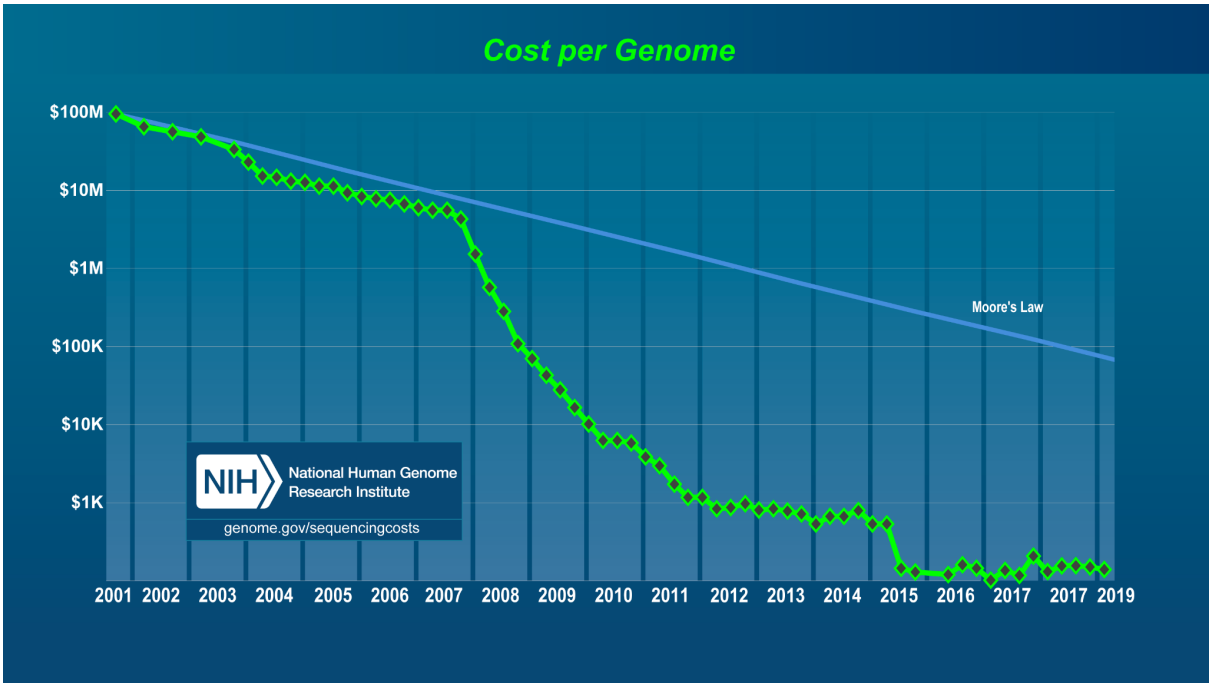

Figura 3: Costo de secuenciación por genoma - 2019

El resultado inevitable de esta fuerte caída en los costos, es un crecimiento en el acceso y capacidad de secuenciación de la humanidad lo que trae como consecuencia un inevitable crecimiento en el número de genomas humanos secuenciados .

Sin embargo, la secuenciación de genomas para generar datos es sólo una parte del trabajo. El control de calidad, el preprocesamiento de las lecturas secuenciadas y el mapeo a un genoma de referencia todavía requieren potentes instalaciones informáticas, algoritmos eficientes y, obviamente, personal experimentado. Es un proceso que requiere mucho tiempo. Además, WGS genera enormes cantidades de datos, lo que supone un reto para el almacenamiento de datos.

En este contexto el cuello de botella está en el análisis de toda esta información. Recordemos que más allá de la tecnología utilizada para preparar la biblioteca y para secuenciar, el resultado de un experimento de secuenciación es un conjunto de lecturas, o mejor dicho una serie de A, C, T y Gs. Es por esto que, por un lado, se necesita contar con herramientas que nos permitan acceder, almacenar, procesar y analizar estos grandes volúmenes de información, para aprovechar todo el potencial que nos brindan los avances tecnológicos mencionados. En esta situación en particular, la bioinformática se torna esencial para la genómica, y de modo más general todas las tecnologías de la información y comunicación (TICs) en el contexto del manejo de datos biológicos.

Genómica y Big Data

En los últimos años se comenzó a ver el cuerpo humano como una fuente de datos, a esta visión aportaron, entre otras, las tecnologías NGS que hicieron ingresar a la genómica al campo de big data, es por ello, que los desarrollos bioinformáticos son esenciales en la genómica y en la salud del siglo 21 en general. Como se sabe big data es un patrón de problema, el desafío de manejar datos estructurados, semiestructurados o no estructurados y crudos en diferentes formas y soportes. La big data se expande en 3 dimensiones (figura 4): volumen, velocidad y variedad.

Figura 4: Las dimensiones en las que se desarrolla la big data.

En particular, en la genómica (y la humana principalmente) se dan especialmente dos dimensiones del concepto Big Data: volumen y variedad. Estos enormes volúmenes de datos necesitan ser almacenados, procesados y analizados. Necesitan ser transformados en conocimiento y relacionarlos con el contexto que les da origen, o sea el individuo y su entorno (o ambiente) que también es una fuente enorme de datos. Esta tarea sólo se puede realizar con herramientas bioinformáticas y tecnologías de la información.

En cuanto a la dimensión variedad en genómica se trabaja con una gran variedad de datos, tanto en la semántica como en la estructura técnica. Algunos datos están estructurados y otros no tanto. Además no solo se trabaja con datos propios, es decir, con la base de conocimiento que se va generando con cada caso analizado, sino que son fundamentales las iniciativas de open Data de organismos internacionales como el NIH (National Institute of Health)[11], EBI (European Bioinformatics Institute) [12], etc., para poder complementar la información obtenida. Estas bases de datos son muy confiables, gozan de alta confiabilidad en la comunidad científica, su veracidad es alta, ya que disponen de recursos que están continuamente curando los datos.

En relación al volumen y tasa de reducción, el procesamiento de los grandes archivos que contienen los datos NGS, consiste en una serie de pasos sucesivos de reducción del volumen. El proceso de análisis es muy iterativo e interactivo, se calcula que la tasa de reducción en cada paso es de 5x, partiendo del archivo de imágenes de datos crudos (TIFF), pasando por los archivos de Reads (FASTQ) hasta llegar al archivo de variantes(VCF), en estas transformaciones para cada genoma se pasa de cientos de GB o TB a entre 3,2GB y 20 GB para el caso de genoma completo, a entre 50 MB y 200 MB en el caso de exomas. Esto dependerá del equipo de secuenciación y de la calidad o cobertura de la misma.

Los archivos VCF (variantes del individuo), son los que se enriquecen con las bases de datos públicas, se guardan y quedan disponibles para el análisis. Luego, dependiendo de las políticas de storage que se definan se podrán almacenar los archivos intermedios, como los fastq, BAM, etc.

Por último respecto a la dimensión velocidad, hay que verla desde dos perspectivas, desde el procesamiento de los datos crudos hasta llegar al listado de variantes, este trabajo se realiza en modo batch y puede tardar varias horas por lo que la dimensión velocidad se ve relegada. Pero una vez que se tiene la información del paciente secuenciada, para realizar el análisis y priorización de variantes, es primordial poder recuperar (visualizar y analizar) la información en poco tiempo.

Objetivos de la tesis

El objetivo general de la tesis fue diseñar un marco de referencia tanto tecnológico como conceptual para la gestión, procesamiento y análisis de información genómica humana derivada de experimentos de secuenciación de próxima generación NGS [13, 14]. Estos desarrollos comprenden dos grandes áreas:

i) Por un lado en el contexto de diagnósticos de pacientes con enfermedades mendelianas, se busca desarrollar una plataforma que permita analizar con detalle y profundidad las variantes de un individuo para, mediante su evaluación, determinar el potencial de las mismas como diagnóstico molecular definitivo.

ii) Por otro en el contexto de la prevención, se busca desarrollar la metodología para poder, a partir de la información genómica, evaluar los riesgos de un individuo, de desarrollar enfermedades complejas de alta prevalencia, como ser las cardiovasculares, el alzheimer, la diabetes tipo 2 o el cáncer entre otras.

iii) Por último en el contexto del tratamiento, se busca desarrollar la metodología para asociar la información genómica de un individuo con la potencial respuesta a los medicamentos, en lo que se conoce como asociaciones farmacogenómicas.

El objetivo comprende por un lado el desarrollo de los procesos bioinformáticos subyacentes que permiten a partir de los datos de NGS (lecturas) obtener la información necesaria para su análisis, y por otro desarrollar una plataforma de software amigable que permita a los profesionales de la salud realizar los análisis mencionados.

Finalmente nos proponemos utilizar los desarrollos en casos concretos de aplicación como prueba de concepto de su funcionalidad y utilidad en la práctica clínica.

Materiales y Métodos

Flujo de procesamiento de los datos NGS

Procesamiento de exomas y paneles de genes

Recordemos que el resultado de un experimento de NGS es la obtención de un conjunto grande (del orden de millones) de fragmentos de ADN de entre 100-150 pb denominadas lecturas. El primer paso obligado de procesamiento de las mismas, en el caso de muestras humanas consiste en lo que se denomina Mapeo-Alineamiento. Este proceso consiste en ubicar las lecturas y alinearlas contra el genoma de referencia *y es necesario* ***para*** **realizar luego el llamado o** *identificación de las variantes presentes en la muestras* ***(llamado de variantes),*** y finalmente la ***anotación funcional*** de las mismas. Como base para este protocolo de tres fases se siguieron las buenas prácticas de *Genome Analysis Toolkit* (GATK) [15][16][17] recomendadas por el Broad Institute [83].

A continuación se describe en detalle el trabajo realizado en cada una de estas tres fases de procesamiento de datos:

Fase 1: Preprocesamiento de datos para la detección de variantes

Conceptualmente en la fase 1 (Figura 5), se toman los datos crudos que produce el secuenciador (lecturas), a través de diferentes algoritmos se los procesa, se los filtra por calidad y finalmente se mapean y se alinean contra el genoma de referencia.

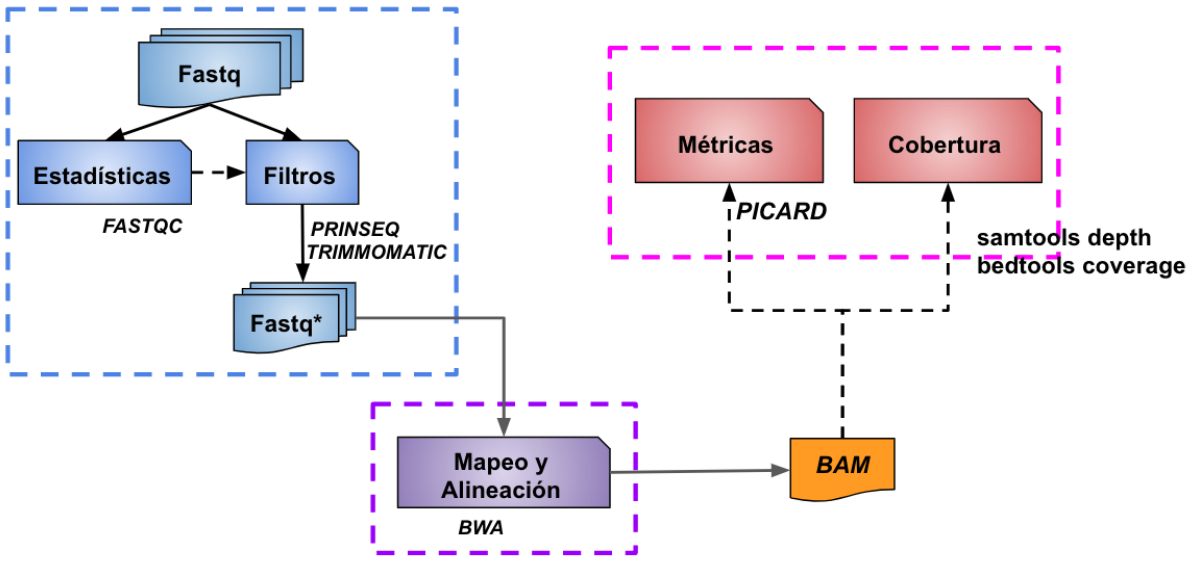

Figura 5. Workflow de Fase 1-preprocesamiento de los datos.

El proceso se divide en 4 etapas. Comenzamos ejecutando una serie de programas para obtener estadísticas de calidad de las lecturas recibidas (fastq) y así saber si hay que realizar alguna corrección o limpieza (trimming) a los datos, y/o eliminar algunas lecturas de baja calidad. A continuación se mapean los reads al genoma de referencia para producir un archivo en formato SAM/BAM (Sequence Alignment Map/Binary Alignment Map)[22] ordenado por coordenadas. Luego, marcamos los duplicados para reducir los sesgos introducidos por la amplificación por PCR durante el proceso de secuenciación. Finalmente, recalibramos los puntajes de calidad de las bases, ya que los algoritmos para el llamado de variantes se basan fuertemente en los puntajes de calidad asignados a las bases.

Específicamente se ejecutaron las siguientes herramientas:

Etapa 1: Análisis de calidad y filtrado de las lecturas

PRINSEQ (prinseq-lite-0.20.4): Prinseq (PReprocessing and INformation of SEQuences)[18] es una herramienta para generar estadísticas y métricas de calidad de los datos de los resultados del experimento de NGS (lecturas). Esta herramienta se puede utilizar para filtrar, reformatear y recortar reads. En particular nosotros la utilizamos para filtrar las lecturas con una calidad media mínima de 20. (-min_qual_mean=20)

FastQC (FastQC v0.11.8): FastQC [19] Proporciona una herramienta para tener una rápida impresión de si los datos obtenidos poseen algún problema antes de poder seguir con el procesamiento de los mismos.

FastqToSam (Picard [20] ***-GATK*** [15][16][17]***):*** FastqToSam es un comando incluido dentro de la suite Picard que convierte un archivo FASTQ en un archivo BAM o SAM no alineado. A partir de un Fastq en un paso posterior se obtiene un uBAM (unaligned Bam).

MarkIlluminaAdapters (Picard [20] ***-GATK*** [15][16][17]***):*** MarkIlluminaAdapters es un comando que lee un archivo BAM/SAM y lo reescribe con nuevas etiquetas para recortar adaptadores. Esta herramienta borra cualquier adaptador existente (XT:i:) en la región de etiquetas opcional de un archivo BAM/SAM.

Etapa 2: Mapeo y Alineamiento contra genoma de referencia

SamToFastq (Picard [20] ***-GATK*** [15][16][17])***:*** SamToFastq es un comando que convierte un archivo SAM/BAM a Fastq.

BWA (bwa-0.7.16a): Burrows Wheeler Aligner (BWA) [21] Se utiliza para mapear cada lectura al genoma de referencia.

MergeBamAlignment (Picard [20] ***-GATK*** [15][16][17]***):*** El uBAM (unaligned Bam) puede contener información útil que se pierde en la conversión a fastq. Esta herramienta toma un bam no alineado(uBAM) con meta-datos, y el bam alineado producto de los comandos SamToFastq y BWA, produce un nuevo archivo BAM que incluye todas las lecturas alineadas, no alineadas, y también los atributos adicionales del uBAM (atributos que se pierden en el proceso de conversión a fastq).

Etapa 3: marcación de duplicados

Esta tercera etapa de procesamiento consiste en identificar las lecturas que probablemente se hayan originado a partir de duplicados de los mismos fragmentos originales de ADN a través de algunos procesos artificiales. Éstas se consideran observaciones no

independientes, por lo que el programa marca todos los pares de lectura duplicados, lo que hace que sean ignorados por defecto durante el proceso de llamado de variantes.

MarkDuplicates (Picard [20] ***-GATK*** [15][16][17]***):*** MarkDuplicates es una herramienta que se utiliza para marcar las lecturas duplicadas.
SortSam (Picard [20] ***-GATK*** [15][16][17]***):*** SortSam es un comando incluido dentro de la suite Picard que ordena los registros por coordenadas genómicas usando la opción SORT_ORDER="coordinate".
SetNmMdAndUqTags (Picard [20] ***-GATK*** [15][16][17])***:*** SetNmMdAndUqTags es un comando incluido en Picard que toma un SAM o BAM ordenado por coordenadas y calcula las etiquetas NM, MD y UQ comparándolas con la referencia.[23]

Etapa 4: recalibración de puntajes de calidad por base

Esta última etapa de preprocesamiento de los datos consiste en aplicar técnicas de machine learning para detectar y corregir patrones de errores sistemáticos en los puntajes de calidad de las bases, que son los parámetros de confianza emitidos por el secuenciador para cada base. Los puntajes de calidad por base juegan un papel importante en la ponderación de la evidencia a favor o en contra de posibles variantes alélicas durante el proceso de descubrimiento de variantes, por lo que es importante corregir cualquier sesgo sistemático observado en los datos. Los sesgos pueden provenir de procesos bioquímicos durante la preparación y secuenciación de la biblioteca, de defectos de fabricación en los chips o de defectos de instrumentación en el secuenciador. El procedimiento de recalibración implica la recopilación de estadísticas de co-variables de todas las bases del conjunto de datos, la construcción de un modelo a partir de dichas estadísticas y la aplicación de ajustes de calidad de base al conjunto de datos basados en el modelo resultante. Finalmente, las reglas de recalibración derivadas del modelo se aplican al conjunto de datos original para producir un conjunto de datos recalibrados. Luego se realiza una operación final de fusión de archivos para producir un único archivo listo para realizar el llamado de variantes. Específicamente se utilizaron los siguientes pasos:

BaseRecalibrator (GATK [15][16][17]***):*** BaseRecalibrator detecta errores cometidos por el secuenciador al estimar los score de calidad de cada base. El comando BaseRecalibrator construye un modelo de covariación a partir de desajustes en los datos de alineación mientras excluye sitios de variantes conocidas y crea un informe de recalibración para su uso en el siguiente paso.
ApplyBQSR (GATK [15][16][17]***):*** ApplyBQSR y el informe de re-calibración se utilizan para corregir los valores de calidad de cada base en el BAM. Aquí es importante mencionar que se realizan dos pasadas de BaseRecalibrator y ApplyBQSR

Por último, en esta cuarta etapa del preprocesamiento de datos se calculan las métricas de calidad y cobertura horizontal y profundidad:

CollectQualityYieldMetrics*(Picard* [20] ***-GATK*** [15][16][17]***):*** se calculan un conjunto de métricas que son utilizadas para describir la calidad general de un archivo BAM. Para consultar el listado de métricas [24].

CollectHsMetrics *(Picard* [20] ***-GATK*** [15][16][17]***):*** Las métricas generadas por CollectHsMetrics se utilizan para el análisis de experimentos de secuenciación target-capture. Las métricas en esta clase se dividen en tres categorías:

- Métricas básicas de secuenciación que se generan como línea de base para evaluar otras métricas o porque se utilizan en el cálculo de otras métricas. Esto incluye medidas como el tamaño del genoma, el número de lecturas, el número de lecturas alineadas, etc.
- Métricas destinadas a evaluar el rendimiento del experimento que generó los datos. Estas métricas se calculan antes de que se apliquen algunos de los filtros.
- Métricas para evaluar la cobertura obtenida en la secuencia para determinar qué tan bien se desempeñarán los datos en lo que sigue del pipeline, como por ejemplo la llamada de variantes. Este grupo incluye métricas como la cobertura media, el porcentaje de bases que alcanzan distintos niveles de cobertura y el porcentaje de bases excluidas por distintos filtros.

Para consultar el listado de métricas [25].

samtools [26]: Se utiliza el comando *depth* de la herramienta samtools para calcular la profundidad en cada posición o región del bed (Browser Extensible Data) utilizado en el experimento de secuenciación.

Bedtools [27]***:*** Se utiliza el comando coverage de la herramienta bedtools que calcula la amplitud de la cobertura observada para cada intervalo indicado en el bed (Browser Extensible Data) utilizado en el experimento de secuenciación.

El resultado de la fase 1 entonces comprende un conjunto de archivos (BAM y SAM) que contienen las lecturas (que superan los filtros de calidad) mapeadas y alineadas cada una contra el genoma de referencia y con un puntaje de calidad recalibrado para cada una de las bases que contiene. El paso que sigue, consiste entonces en realizar el llamado de variantes, pero primeramente se definirá el concepto de profundidad y amplitud de la cobertura.

En un experimento de secuenciación la profundidad (figura 6) no es uniforme en todo el genoma, sino que hay regiones que tienen más cobertura que otras, por eso la cobertura se mide en promedio. La profundidad de cobertura (mapping depth) por base es el número medio de veces que una base de un genoma es secuenciada.
Para un exoma la profundidad de cobertura puede ir de 100x a 300x, para secuenciar paneles de genes se suele utilizar mayor cobertura(más de 500x), y para un genoma completo, lo normal es alrededor de 30x, que es una calidad aceptable. Cuanto más profundidad se requiera más tiempo se tarda en secuenciar y el costo es mayor.

La amplitud de la cobertura (breadth coverage) es el porcentaje de bases de un genoma de referencia que están cubiertas con una determinada profundidad. Puede haber regiones que no estén cubiertas, ni siquiera por una sola lectura.

.

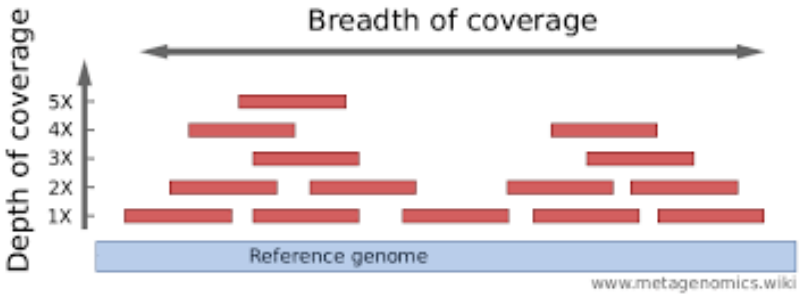

Figura 6: Se visualiza la profundidad y amplitud de cobertura obtenida en el experimento de secuenciación luego de la fase 1 - preprocesamiento de los datos, es decir, luego del mapeo y alineación de las lecturas contra el genoma de referencia.
https://www.metagenomics.wiki/pdf/definition/coverage-read-depth

Fase 2: Llamado de variantes (SNPs + Indels)

Técnicamente las variantes se determinan cuando las lecturas obtenidas en el proceso de secuenciación muestran diferencias respecto del genoma de referencia. En la fase 2 (Figura 7) básicamente se comparan las bases que contienen las lecturas mapeadas contra la referencia para detectar variantes particulares, a este proceso se los llama ***variant calling.***

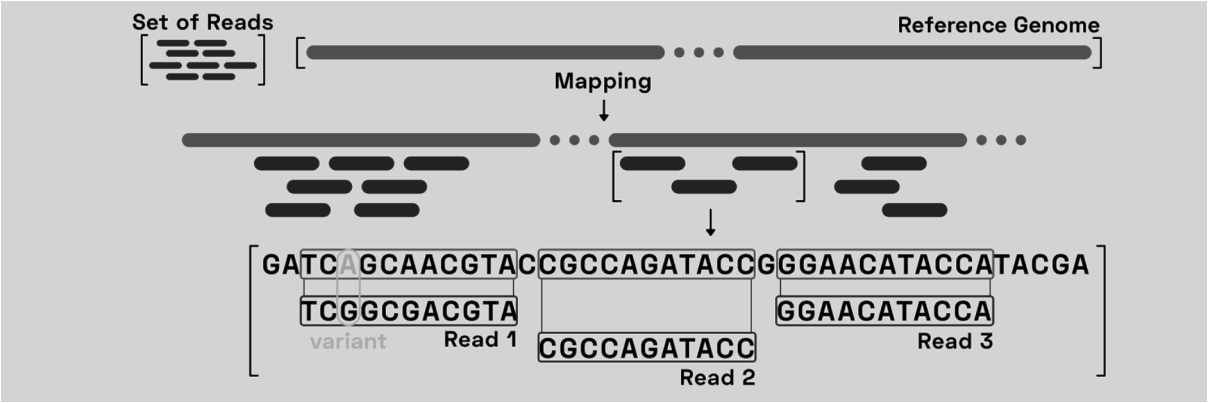

Figura 7: Se visualiza en el alineamiento de las lecturas la variante que produce el cambio de una A por una G.

Cuando se identifica una variante es fundamental reportar su profundidad. Es decir, la cantidad de lecturas(reads) en donde se leyó el alelo alternativo sobre la cantidad total de lecturas en esa posición. En el ejemplo la figura y, la cantidad de lecturas donde se encontró una G (alternativo) sobre la cantidad total de lecturas en esa posición.

El archivo con las variantes generado en esta fase se llama **VCF** (Variant Calling File) y representa en cierto modo la información genómica específica de la muestra. En otras palabras, como ya mencionamos, el VCF es el archivo que define el genoma personal del individuo.

Una vez que los datos han sido preprocesados como se ha descrito anteriormente, son sometidos al proceso de llamado de variantes, es decir, la identificación de los sitios en los que los datos muestran variaciones relativas al genoma de referencia, y el cálculo de genotipos para cada muestra en ese sitio. Debido a que algunas de las variantes observadas son causadas por errores (artefactos) de mapeo y secuenciación, el mayor desafío aquí es equilibrar la necesidad de *sensibilidad* (para minimizar los falsos negativos, es decir, no identificar las variantes reales) vs. *especificidad* (para minimizar los falsos positivos, es decir, no tomar como variantes los artefactos del proceso previo).

Para alcanzar estos objetivos, el proceso de descubrimiento de variantes se descompone en dos pasos (Figura 8) separados: **llamado de variantes** y **filtrado (por calidad) de variantes**. El primer paso está diseñado para maximizar la sensibilidad, mientras que el paso de filtrado tiene como objetivo proporcionar un nivel de especificidad.

Llamado de variantes: Realizamos el llamado de variantes en cada archivo BAM para crear los VCFs que contengan llamadas SNP e indel. Para realizar este paso, en nuestro pipeline hemos utilizado distintas herramientas, entre ellas HaplotypeCaller de GATK [17], Platypus [28], Freebayes [29] y Samtools mpileup [26]. Luego se realiza una combinación (un merge) con el resultado de cada uno de los programas mencionados para generar un archivo de variantes unificado.

Filtrado de variantes: para el filtrado de variantes se realizan dos etapas. La primera de ellas, la que se denomina “Hard Filtering”, que consiste en elegir umbrales específicos para una o más anotaciones y descartar (marcar) cualquier variante que tenga valores de anotación por encima o por debajo de los umbrales establecidos. Por anotación, nos referimos a las diferentes propiedades que describen a cada variante derivadas del proceso de llamado, como por ejemplo, cuántas lecturas la cubrieron, cuántas lecturas cubrieron cada alelo, qué proporción de lecturas estaban en orientación hacia adelante y hacia atrás, etc.

La segunda fase comprende el Variant Quality Score Recalibration (VQSR) y lo que hace es calibrar con un modelo estadístico normal multivariado cada uno de los valores de anotación mencionados. Para ello se toma como conjunto de entrenamiento (o parametrización) todas la variantes presentes en la muestra que se encuentran en bases de datos que superan cierto umbral de frecuencia (o sea son variantes comunes, polimorfismos, que uno espera encontrar en cualquier muestra), y luego se evalúa la posibilidad de error tipo II para cada una de las variantes nuevas en función del mismo. En otras palabras se obtiene para cada variante un parámetro que se interpreta como la probabilidad de que esa variante aparezca en la muestra por error. Usualmente se considera que una variante es de buena calidad si este valor es inferior a 1x10-4 (o menor).

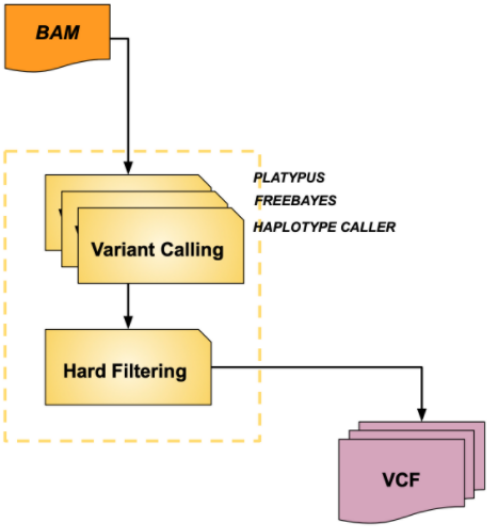

Figura 8. Fase 2 - llamado de variantes.

Fase 3: Anotación estructural y funcional de variantes

Una vez detectadas las variantes (en la fase 2) a cada una de ellas se las vincula y se las enriquece con información biológicamente relevante (que se encuentra a disposición en bases de datos públicas). A este proceso se lo llama **anotación** de variantes (Figura 9).

Se realizan básicamente dos tipos de anotaciones, una estructural y otra funcional:

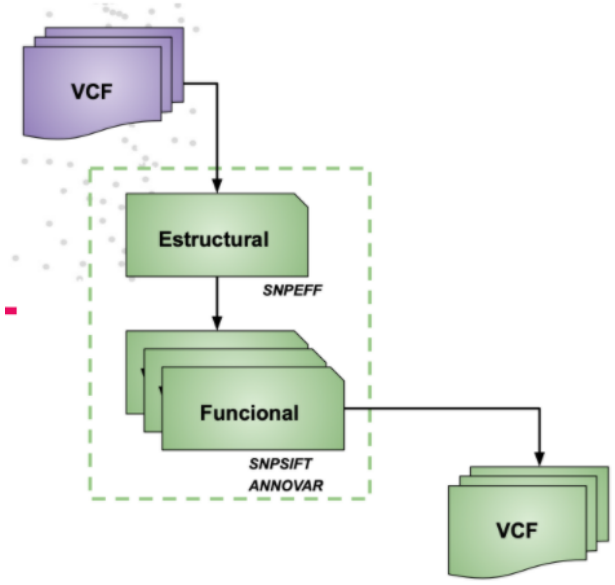

Figura 9. Fase 3 - Anotación funcional de variantes.

En la **anotación estructural** lo que se busca es poder encontrar genes y otros sitios (o posiciones) del genoma con importancia biológica. Cada gen/sitio está asociado a una posición en el genoma y se busca determinar, por ejemplo, si la variante se encuentra en una región codificante, si es una variantes intrónica, si se ubica en un sitio de splicing y el tipo de variante que es, si es un start, un stop prematuro, un frameshift, si es un cambio de

missense o uno sinónimo, etc. Realizamos la anotación estructural con snpEFF/snpSift [30][31].

En la **anotación funcional** lo que se busca es asociar una función biológica conocida a la variante y/o gen/sitio donde se encuentra la misma. De los diferentes criterios utilizados para anotar funcionalmente las variantes destacamos los siguientes:

- **Basados en la información propia del gen o la patología:** se identifica si una variante provoca o no un cambio aminoacídico en la proteína que codifica, o si estas aparecen en las regiones de splicing, y/o en regiones intrónicas etc. Si se conoce que el gen produce alguna patología, o si se sabe el efecto sobre la función proteica. Para este fin se pueden utilizar bases de datos como OMIM [32], ENSEMBL[33], Uniprot[34], entre otras.
- **Basados en información ya documentada de la variante**. Se coteja bases de datos específicas para identificar cambios ya documentados, como dbSNP[35], ClinVar [38] (proporciona información clínica de la variante si es que algún grupo ya la ha clasificado); bases de datos para obtener frecuencias poblacionales en diferentes grupos étnicos como 1000 Genomas [37] y ExAC (Exome Aggregation Consortium)[52].
- **Basadas en predicciones de patogenicidad y conservación**: se agregan score de predicción de la patogenicidad del cambio, utilizando herramientas al efecto como Polyphen [39], SIFT [40] y MutationTaster [41], etc. También se complementa la anotación con los score de conservación: GERP++[44], PhyloP [43], PhastCons [42].

En la próxima sección hablaremos con mayor detalle de las principales bases de datos utilizadas para anotar funcionalmente las variantes halladas.

Para anotar funcionalmente las variantes halladas utilizamos dos herramientas, SnpSift[31] y Annovar[45]. La primera la utilizamos para agregar a las variantes información de dbSNP[35], de ExAC (Exome Aggregation Consortium)[52], de 1000 Genomas [37] y de ClinVar [38].

La segunda herramienta utilizada es Annovar[45]. Entre las anotaciones que agrega podemos detallar los siguientes predictores de patogenicidad: Polyphen [39], SIFT [40] y MutationTaster [41], etc., entre los score de conservación: GERP++[44], PhyloP [43], PhastCons [42].

Finalmente, como último paso del pipeline de procesamiento utilizamos InterVar[46], que es una herramienta bioinformática para la interpretación/anotación clínica de las variantes genéticas según las directrices de ACMG/AMP 2015, detalladas en otra sección.

La entrada de InterVar es un archivo anotado generado a partir de Annovar, y como resultado obtenemos la clasificación ACMG de las variantes, junto con los criterios evaluados para llegar a esa clasificación.

InterVar puede tomar un archivo pre-anotado con Annovar o un archivo VCF como entrada, y generar una clasificación de variantes según los criterios de ACMG. Intervar evalúa de forma automática 18 de los 28 criterios establecidos. El resto de los criterios se puede especificar manualmente y así obtener una clasificación completa.

Bases de datos

Para las ciencias biológicas, y más específicamente dentro del área de la genómica humana, existe una gran variedad de bases de datos públicas, disponibles gratuitamente que pueden ser consultadas libremente. Como se explicó anteriormente para la anotación funcional de las variantes se utiliza información de múltiples base de datos, ya sean las que contienen información asociadas a genes, como aquellas que se centran en información particulares de las variantes. Veamos cuales son las principales.

Basadas en la información propia del gen o la patología

Antes de avanzar con el compendio de bases de datos utilizadas para la anotación funcional de los genes es conveniente recordar que son los genes en relación al genoma de referencia desde la perspectiva de la bioinformática.

En el contexto de la anotación del genoma, un gen tiene al menos un nombre de referencia, una posición en el genoma (comienzo y fin), y además su colección de secuencias de transcripción de ARN relacionadas ("isoformas" o transcriptos). La denominación de los genes y la asignación de las secuencias de los transcriptos más importantes se realiza a menudo de forma manual por un grupo de conservadores de la literatura biológica. En el caso de los humanos, los nombres de los genes son creados por el Comité de Nomenclatura Genética Humana (HGNC[47], anteriormente HUGO), es la única autoridad mundial que asigna una nomenclatura estandarizada a los genes humanos.

Los transcriptos, que usualmente son propuestos por técnicas bioinformáticas y luego verificados experimentalmente mediante secuenciación del RNA (RNAseq) se encuentran disponibles en diversas bases de datos. El NCBI de la Biblioteca Nacional de Medicina [48] o el Instituto Europeo de Bioinformática (EBI) [12] recogen y clasifican estas secuencias de transcripción de parte de los investigadores que trabajan en cada gen. Cada transcripto tiene un identificador único (accession), un gen al que se le asigna, una secuencia y una lista de coordenadas de inicio y fin del cromosoma.

Existen diferentes proyectos de distintas instituciones que tienen como objetivo recolectar los diferentes transcriptos de los genes, por ejemplo, NCBI (RefSeq)[49], UCSC [50], ENSEMBL[33], etc.

Cada institución tiene diferentes reglas sobre cómo anotar los genes. Por ejemplo, los criterios de RefSeq son más estrictos, por lo que hay menos transcriptos en RefSeq que en ENSEMBL/GENCODE.

ENSEMBL

En este trabajo utilizamos principalmente ENSEMBL para asignar y analizar los transcriptos, es por ello que a continuación describimos brevemente la misma.
El proyecto ENSEMBL [33] se inició en 1999, unos años antes de que se completara el genoma humano. Este proyecto es un trabajo en colaboración entre el European Bioinformatic Institute (EMBL-EBI)[12] y el Wellcome Trust Sanger Institute [51], con el objetivo de anotar automáticamente genomas de vertebrados y otras especies eucariotas, integrar esta anotación con otros datos biológicos, centralizar la información disponible y poner todo esto a disposición del público a través de la web.

Los conjuntos de genes de ENSEMBL están basados en la evidencia de las bases de datos de secuencias incluyendo Uniprot[34] y RefSeq[49]. Cuando un transcripto de ENSEMBL es coincidente con un transcripto en RefSeq, los dos transcriptos se vinculan.
Además de enlazar la anotación ENSEMBL con la correspondiente anotación RefSeq, el conjunto completo de modelos RefSeq se importa en ENSEMBL para el humano y el ratón.

Mientras que los modelos de genes ENSEMBL se anotan directamente en el genoma de referencia, RefSeq anota en las secuencias de ARNm. Debido a las diferencias de la secuencia entre los genomas de referencia y mRNAs individuales, algunos de los mRNAs de RefSeq no pueden mapear perfectamente al genoma de referencia.

OMIM, online Mendelian Inheritance in Man

Para ver el espectro completo de los trastornos mendelianos, uno de los mejores lugares para buscar es Online Mendelian Inheritance in Man (OMIM) [32]. OMIM es una base de datos pública de información bibliográfica sobre genes humanos y trastornos genéticos, en particular trastornos mendelianos, denominados “MIM” (del inglés Mendelian Inheritance in Man). Contiene información de más de 4.000 trastornos mendelianos conocidos y sus fenotipos característicos asociados, además de contener información de un total de más de 15 mil genes.

El objetivo principal de OMIM es catalogar todas las enfermedades que posean un componente genético mendeliano (MIM) , y generar relaciones bibliográficas con otras bases de datos con datos genómicos. OMIM se centra en la relación entre fenotipo y genotipo. Se actualiza diariamente, y las entradas contienen abundantes enlaces a otros recursos genéticos.

Es un amplio compendio de genes humanos y fenotipos genéticos. Cada entrada de OMIM tiene un resumen de texto completo de un fenotipo y/o gen, tiene abundantes enlaces a otros recursos genéticos como el ADN y la secuencia de proteínas, referencias de PubMed, bases de datos de mutaciones, nomenclatura de genes aprobada y más.

En este trabajo de tesis utilizamos OMIM principalmente para vincular el fenotipo de los pacientes estudiados, en relación con el fenotipo descrito en OMIM para los genes donde encontramos variantes de interés.

Uniprot

Uniprot (de Universal Protein) [34] es una fuente integral de datos de secuencias proteicas y anotaciones funcionales relacionadas a ellas. Las bases de datos que componen Uniprot son UniprotKB (Uniprot Knowledgebase), UniRef (Uniprot Reference Cluster) y Uniprot Archive (UniParc).

UniprotKB es el eje principal para la recolección de información funcional en proteínas, a partir de un vasto número de anotaciones. A su vez, esta base de datos está dividida en dos partes, según sus entradas hayan sido manualmente curadas (Swiss-Prot) o subidas de manera automática (TrEMBL). Por ejemplo, el proteoma de Homo Sapiens tiene 173.324 proteínas, donde aproximadamente el 12% pertenece a Swiss-Prot y el 88% a TrEMBL. Mientras que las primeras disponen de evidencia experimental, el segundo grupo tiene poca o ninguna evidencia de que se expresan en el organismo.

UniParc es la recopilación de secuencias proteicas de todas las bases de datos más conocidas y agrupadas bajo un ID único por proteína. Esto permite eliminar la redundancia de secuencias debido a múltiples fuentes de datos. Allí también se guarda el “historial” de las proteínas, es decir, cómo van cambiando sus anotaciones, secuencia o si son eliminadas por nuevas entradas.

Por último, Uniref agrupa las secuencias de UniprotKB en clusters según su porcentaje de identidad de secuencia. Por ejemplo, P62258 está en humano, gallina y en otras especies, comparten el 100% de la secuencia, pero en gallina tiene el ID Q5ZMT0. Tanto P62258 como Q5ZMT0 (y otras 14 proteínas de distintas especies) pertenecen al cluster UniRef100_P62258.

A la hora de priorizar las variantes, y de asignarles significancia clínica, la información de la proteína - expresada por el gen afectado - contenida en UNIPROT es sumamente importante, porque nos permite aventurar y comprender el efecto de la variante sobre la función proteica.

La información a la que podemos acceder en cada entrada de UNIPROT, está dividida en secciones. Describiremos las más relevantes al momento de la priorización:

La primera sección proporciona información útil sobre la función de la proteína, principalmente conocimientos biológicos. Por ejemplo, se informa sobre la actividad catalítica de la enzima, es decir, la reacción que la enzima cataliza, propiedades biofísicas y físico-químicas, vías metabólicas asociadas, los sitios activos directamente involucrados en la actividad indicando los residuos involucrados en la catálisis. También se informan los sitios de unión para cualquier grupo químico (coenzima, grupo prostético , etc.), se describe la interacción entre un aminoácido y otra entidad química, entre otras. De este modo es posible conocer si la variante analizada afecta a un aminoácido involucrado en los sitios importantes de la proteína afectando por ende su función.

Por otra parte, una sección interesante es la proporciona información sobre las enfermedades y fenotipos conocidos asociados a la proteína. Esta información se extrae de la literatura científica y las enfermedades que también se describen en la base de datos de OMIM. Aquí se listan las variantes conocidas de la proteína. Se busca que haya correspondencia entre la historia clínica del paciente analizado, y las enfermedades y fenotipos asociados a la proteína.

En Uniprot se puede encontrar información de expresión de la proteína, la especificidad tisular, por ejemplo que las isoformas 3, 7 y 8 se expresan en el músculo cardíaco. La isoforma 4 se expresa en el músculo esquelético vertebrado y la isoforma 6 se expresa en el músculo esquelético. Esta información se puede relacionar, en algunos casos, con la clínica del paciente.

Además se encontrará información de interacción de la proteína afectada con otras proteínas (o complejos de proteínas), muchas veces identificando los residuos involucrados, lo que permitirá saber qué otras funciones se pueden ver afectadas.

También proporciona información sobre la estructura terciaria y secundaria de la proteína. Esto permite saber si el cambio de aminoácido producido por la variante se encuentra en una alfa hélice, en una hoja beta o en un loop. Conociendo la estructura de la proteína se pueden además realizar estudios de interacción entre los aminoácidos afectados por la variante. A continuación podemos conocer la familia y dominios de la proteína que da información sobre las similitudes de secuencia con otras proteínas y los dominios presentes en la proteína.

Por último, tenemos la sección de secuencias donde se muestra la secuencia canónica de la proteína y todas las isoformas descritas en la literatura. También incluye información pertinente a las secuencias, incluyendo la longitud y su peso molecular. Aquí es importante saber la longitud de la proteína porque por ejemplo, si se está analizando una variante que produce un stop prematuro es importante saberlo para determinar en cuanto a su efecto qué parte de la proteína se trunca, si es al comienzo, a la mitad o al final.

Basadas en información ya documentada de la variante.

CLINVAR

ClinVar [38] es una base de datos pública sobre relaciones entre variantes y fenotipos humanos, validadas, con evidencia. Esta base de datos se construye a partir de informes de las relaciones entre las variantes presentes en pacientes y/o individuos que forman parte de protocolos de investigación, y sus fenotipos clínicos. ClinVar facilita así el acceso y la comunicación sobre las relaciones informadas entre las variantes presentes en un individuo y el estado de salud observado, y la historia de esa interpretación. ClinVar procesa y almacena los informes de las variantes encontradas en las muestras de los pacientes, las afirmaciones hechas con respecto a su significado clínico, la información sobre el remitente, y otros datos que soportan la interpretación de la misma. Los alelos descritos en los envíos se asignan a secuencias de referencia y se informan de acuerdo con el estándar HGVS. ClinVar entonces presenta los datos para los usuarios tanto de manera interactiva (vía web),

como en “batch” (se accede a toda la base de datos) para aquellos que deseen usar ClinVar en los flujos de trabajo diarios y otras aplicaciones locales. ClinVar trabaja en colaboración con organizaciones interesadas para satisfacer las necesidades de la comunidad de genética médica de la manera más eficiente y eficaz posible.

ClinVar soporta presentaciones de diferentes niveles de complejidad. La presentación puede ser tan simple como la representación de un alelo (variante) y su interpretación (a veces denominada presentación de nivel de variante), o tan detallada como la provisión de múltiples tipos de evidencia observacional estructurada (a nivel de caso) o experimental sobre el efecto de la variante en el fenotipo.

Uno de los principales objetivos de ClinVar es apoyar la evaluación computacional, tanto de los genotipos como de las aseveraciones, y permitir la evolución y el desarrollo continuo del conocimiento sobre las variaciones y los fenotipos asociados. ClinVar es un socio activo del proyecto ClinGen[82], proporcionando datos para la evaluación y archivando los resultados de la interpretación de reconocidos paneles de expertos y proveedores de guías de práctica. ClinVar archiva y presenta versiones, lo que significa que cuando los remitentes actualizan sus registros, la versión anterior se conserva para su revisión.

Nuevamente al momento del análisis de variantes el nivel de confianza en la precisión del llamado de variantes y las afirmaciones de significación clínica depende en gran medida de las evidencias de apoyo, por lo que esta información, cuando está disponible, se recopila y es visible(y accesible para descarga) para los usuarios a través de ClinVar.

ClinVar utiliza términos estándar para clasificar las variantes reportadas según su significancia clínica ***(Clinical significance)*** de acuerdo a los cinco criterios recomendados por ACMG/AMP[30], estas son ***"pathogenic", "likely pathogenic", "uncertain significance", "likely benign", y "benign"*** -, y se utilizan para describir las variantes identificadas en los genes que causan los trastornos mendelianos.

Adicionalmente se utilizan otros términos no estándares para clasificar algunas variantes particulares que no caen en las categorías anteriores, como por ejemplo:

Drug response: Un término general para una variante que afecta la respuesta del individuo al uso de algún fármaco .
Association: Para las variantes identificadas en estudios de GWAS.
Risk factor: Para las variantes cuya interpretación determina que no es causante de un trastorno, sino que aumenta el riesgo de contraer o desarrollar alguna patología.
Protective: Aquellas variantes que disminuyen el riesgo de desarrollar ciertas patologías, incluyendo las infecciones.
Affects: Para las variantes que causan un fenotipo particular pero no una enfermedad, como por ejemplo la intolerancia a la lactosa.
Not provided: No se posee información adicional para la variante.

Además ClinVar informa el nivel de evidencia que sustenta la significancia clínica asociada a la variante reportada. Informa el “estado de revisión” y a cada una le asigna una cantidad de estrellas para poder representar gráficamente la confianza en la anotación reportada.
ClinVar ha desarrollado un sistema de clasificación de cuatro estrellas, que representa el

"Estado de revisión" de cada entrada/variante. Por defecto, las variantes reportadas en ClinVar tienen el estado de revisión ***single submitter - criteria not provided***. Sin embargo, las entradas pueden luego ser clasificadas con los estados de ***single submitter - criteria provided***, ***expert panel*** y ***practice guidelines*** según las descripciones que figuran a continuación.

Tabla 1: Lista de estados de revisión de ClinVar

#Stars	Review status	Description
4	practice guideline	La información de la variante fue revisada por el Comité de ClinGen. ClinGen es una entidad financiada por recursos del Instituto Nacional de Salud (NIH) [11] dedicada a construir un ente central autorizado que defina la relevancia clínica de los genes y variantes, para su uso en medicina de precisión e investigación.
3	reviewed by expert panel	La clasificación fue revisada por un grupo de expertos.
2	criteria provided, multiple submitters, no conflicts	Cuando dos o más remitentes reportan que la variante cumple con los criterios requeridos para la clasificación otorgada, y presentan evidencia que sustenta la misma interpretación.
1	criteria provided, conflicting interpretations	Múltiples remitentes reportan que la variante cumple con los criterios requeridos para la clasificación otorgada y presentan pruebas, pero hay interpretaciones conflictivas. Los valores independientes se enumeran por su importancia clínica.
	criteria provided, single submitter	Un remitente reporta que la variante cumple con los criterios requeridos para la clasificación otorgada y presenta evidencia.
0	no assertion for the individual variant	La variante no fue interpretada directamente en ninguna entrada; fue subida a ClinVar sólo como un componente de un haplotipo o un genotipo.
	no assertion criteria provided	La variante se incluyó en una entrada con una interpretación pero sin los criterios requeridos para la clasificación otorgada.
	no assertion provided	La variante se incluyó en una entrada que no proporcionó una interpretación.

Otros datos importante que aparecen en ClinVar para cada entrada son:

- el campo ***Last evaluated,*** que es la fecha en que el solicitante evaluó por última vez la importancia de la variante.
- El campo ***Last Updated***, que indica la fecha en que ClinVar actualizó el registro por última vez. Esto incluye las actualizaciones de los envíos de la variante, así como las

actualizaciones de los datos que ClinVar proporciona, como los enlaces a recursos relacionados.

Entonces resumiendo, al momento de evaluar una variante reportada en ClinVar es importante prestar atención al menos a estos tres campos que se mencionaron anteriormente, Clinical significance, Review status y Last Updated, ya que si una variante es patogénica, probablemente patogénica o de significado incierto es importante conocer cuál es el estado de su revisión y cuándo fue la última vez que se actualizó debido a que muchas veces las variantes se reclasifican.

gnomAD

gnomAD (Genome Aggregation Database) es una base de datos poblacional con información de las frecuencias alélicas de las variantes en diferentes poblaciones. Es un proyecto en el que los investigadores tratan de agregar y normalizar los datos de secuencias de exomas y genomas de una gran variedad de proyectos de secuenciación a gran escala, para poner a disposición de la comunidad científica los resultados de la genómica poblacional resultante. En su primera publicación [52], que contenía exclusivamente datos de exomas, se la conoció como el Consorcio de Agregación del Exoma (ExAC).

En la versión más reciente, gnomAD [53] v3, abarca el análisis de 71.702 genomas de individuos no relacionados, secuenciados como parte de diversos estudios de enfermedades específicas y de genética de poblaciones, todos alineados contra el genoma de referencia GRCh38. La versión v2 también disponible incluye 125.748 exomas y 15.708 genomas de individuos, que al igual que la versión v3, no están relacionados entre sí, y fueron secuenciados como parte de varios estudios de genética poblacional y específica de enfermedades, con un total de 141.456 individuos, en este caso alineados contra el genoma de referencia GRCh37.

gnomAD es una base de datos curada, ya que, entre otras cosas, se han filtrado aquellos individuos afectados por enfermedades pediátricas severas (y todos sus parientes en primer grado), por lo que las frecuencias calculadas se consideran de individuos “sanos”. Sin embargo, hay que tener en cuenta que algunos individuos con enfermedades severas pueden seguir estando incluidos en los conjuntos de datos debido a alguna enfermedad de aparición tardía, aunque probablemente con una frecuencia equivalente o inferior a la observada en la población general.

El conjunto de datos de la gnomAD contiene individuos secuenciados utilizando múltiples métodos de secuenciación, por lo que la cobertura varía entre los individuos y entre los distintos sitios. Esta variación de la cobertura gnomAD la incorpora en los cálculos de la frecuencia de las variantes para cada variante.

Tener información de la frecuencia poblacional de las variantes nos permite clasificarlas al momento de la priorización. Entonces basándonos en la frecuencia del alelo minoritario, podemos clasificar una variante como rara (poco frecuente) si ocurre en menos del 0,5% de la población (este valor es uno de los puntos de corte para uno de los criterios de evidencia

de la clasificación ACMG). Variantes de baja frecuencia, son aquellas que tienen una frecuencia entre 0,5% y el 5% de la población. Y las variantes más comunes son aquellas que ocurren en más del 5% de la población.

1000 Genomas

El Proyecto 1000 Genomas[37] se desarrolló entre 2008 y 2015, creando el mayor catálogo público de variantes humanas y datos de su genotipo. El objetivo del Proyecto era encontrar la mayoría de las variantes genéticas con frecuencias de al menos el 1% en todas las poblaciones estudiadas.

El proyecto comenzó utilizando tecnología de genotipificación por microarreglos, pero durante su desarrollo se aprovecharon los avances en la tecnología de la secuenciación, lo que redujo drásticamente el costo y expandió de manera drástica la información generada. Fue el primer proyecto que secuenció los genomas de un gran número de personas, para proporcionar un recurso completo sobre la variación genética humana. Los datos del Proyecto de “los 1000 Genomas” se pusieron rápidamente a disposición de la comunidad científica mundial mediante bases de datos públicas de libre acceso. Las muestras del Proyecto son anónimas y no tienen datos médicos o de fenotipo asociados. El proyecto contiene datos sobre el origen étnico y el género declarados por los propios interesados. Todos los participantes declararon estar sanos en el momento en que se recogieron las muestras. Actualmente, el mismo permite determinar cuándo uno encuentra una variante, si la misma ya fue encontrada en 1000 Genomas y se conoce su frecuencia.

Recomendaciones del ACMG/AMP para la clasificación de variantes

Retomando el tema señalado en la introducción en la sección Clasificación de las variantes en la genómica clínica, la determinación precisa de la patogenicidad de una variante es una de las tareas más complejas e importantes, dado que define un diagnóstico preciso y puede ser el primer determinante de una decisión terapéutica.

En este contexto, luego de que varios trabajos científicos reportan inconsistencias en la priorización de variantes en diferentes laboratorios internacionales, y se observara que en ClinVar alrededor de un 17% de las variantes - con más de una entrada - tenían conflictos de interpretación[6], en el año 2015, el American College of Medical genetics (ACMG) y la Asociación de Patología Molecular (AMP), convocaron a un conjunto de expertos para que trabajaran en la revisión de las pautas vigentes para la interpretación y clasificación de variantes y, en un posterior desarrollo de un enfoque sistemático y transparente para la clasificación de variantes, principalmente centrado en genes asociados a enfermedades mendelianas.

El enfoque de la ACMG recomienda el uso de terminología específica: "Patogénica", "Probablemente Patogénica", "Significado Incierto", "Probablemente Benigna" y "Benigna" para describir a las variantes identificadas. Por otra parte, propone un proceso de dos etapas para clasificar variantes (en estas cinco categorías) teniendo en cuenta la evidencia que aportan diferentes características de la variante, por ejemplo, datos poblacionales, computacionales, funcionales, de segregación familiar, etc.

La primera etapa consiste entonces en la recolección de las evidencias que presenta la variante, y la segunda en la clasificación dentro de una de las cinco categorías, consolidando la sumatoria de las evidencias presentes. A este enfoque sistemático se lo conoce usualmente como l*as recomendaciones de la ACMG/AMP* [6].

Las recomendaciones de la ACMG/AMP evalúan 28 tipos diferentes de evidencia asociados a características inherentes a cada variante, como pueden ser, la frecuencia alélica, análisis funcionales, predicciones in-silico, análisis de segregación, relación genotipo-fenotipo, asignándole a cada una un código alfanumérico que valora su contribución al nivel de patogenicidad.

Específicamente hay un código que asigna un puntaje alfanumérico a cada nivel (o tipo) de evidencia. Estos son: *Pathogenic Very Strong (PVS)*, *Pathogenic Strong (PS1,PS2,PS3,PS4)*, *Pathogenic Moderate (PM1,PM2,PM3,PM4,PM5,PM6), Pathogenic Supporting (PP1,PP2,PP3,PP4)*, *Benign Stand Alone (BA)*, *Benign Strong (BS1,BS2,BS3,BS4)* y *Benign Supporting (BP1,BP2,BP3,BP4,BP5,BP6,BP7)*. La numeración dentro de cada categoría permite diferenciar los distintos criterios y NO implica una diferencia de relevancia. Estos códigos, detallados más adelante en las ***Tabla 2: Evidencia para variantes patogénicas y Tabla 3: Evidencia para variantes Benignas***, funcionan luego como un puntaje, que combinándolos dan lugar a la clasificación final de cada variante en una de las 5 categorías: Pathogenic (P), Likely Pathogenic(LP), Variant of Uncertain Significance (VUS), Likely Benign (LB), o Benign(B).

A modo de ejemplo, para que una variante sea patogénica se necesitan 2 PS, o 1 PS y más de 2 PM, y para que una variante sea benigna 1 BA, o más de 2 BS. Cuándo el nivel de evidencia (o sea la suma de los puntajes) no alcanza para que una variante sea P/LP o B/LB, o existe información contradictoria, la misma es catalogada como de Significado Incierto (o VUS).

Recolección de evidencias

El primer punto importante es reconocer que los “tipos” de evidencia se pueden agrupar de acuerdo al proceso biológico en el que se sustentan. En la siguiente Tabla se resumen los 26 niveles de evidencia, agrupados por categoría (o puntaje alfanumérico)

Tabla 2: Evidencia para variantes patogénicas

Muy fuerte	
PVS1	Variante nula (sin sentido, desplazamiento de marco, sitios de corte y empalme canónico ± 1 o 2, codón de iniciación, deleción simple o multi exón) en un gen donde “loss-of-function” (LOF) es un mecanismo conocido de la enfermedad. ● Cuidado cuando LOF no es un mecanismo seguro que produzca enfermedades en ciertos genes. ● Cuidado con variantes cercanas al extremo final del gen. ● Cuidado con la presencia de varios transcriptos.
Fuerte	
PS1	Mismo cambio de aminoácido que una variante patogénica previamente reportada.
PS2	De novo (confirmado por maternidad y paternidad) en un paciente con la enfermedad y sin antecedentes familiares.
PS3	Estudios funcionales in vitro o in vivo que apoyan un efecto dañino sobre el gen o el producto génico resultante.
PS4	La prevalencia de la variante en las personas afectadas aumenta significativamente en comparación con la prevalencia en los controles (riesgo relativo o OR> 5,0).
Moderada	
PM1	Localizada en un dominio funcional bien establecido (por ejemplo, sitio activo de una enzima) donde todos las variantes missense en estos dominios identificados hasta la fecha han demostrado ser patogénicas. Estos dominios también deben carecer de variantes benignas.
PM2	Ausente en los controles (o a una frecuencia extremadamente baja si es recesivo, menor de lo esperado).
PM3	Para trastornos recesivos, variantes detectadas en trans con una variante patogénica.
PM4	La longitud de la proteína cambia como resultado de deleciones o inserciones dentro del marco de lectura (in-frame) en una región no repetitiva, o variantes de pérdida de codón stop (stop-loss)
PM5	Nueva variante missense en un residuo en el que se ha observado una mutación missense diferente reportada como patogénica. Ejemplo: Arg156His es patogénica; ahora observa Arg156Cys.
PM6	Presunta de novo, pero sin confirmación en padre y madre.

Secundarias/Apoyo	
PP1	Co-segregación de la variante con la enfermedad en múltiples miembros de la familia afectada y en un gen candidato principal.
PP2	Variante Missense en un gen donde la mayoría de las variantes de cambio de aminoácido son patogénicas (no benignas).
PP3	Varias líneas/predictores de evidencia computacional respaldan un efecto nocivo sobre el gen o la proteína (criterios de conservación, evolución, impacto de corte y empalme, etc.).
PP4	El fenotipo o la historia familiar del paciente es altamente específica para una enfermedad con una sola etiología genética.
PP5	Una base de datos curada la reportó recientemente como patogénica.

Tabla 3: Evidencia para variantes benignas

Independiente/Única	
BA1	La frecuencia del alelo es mayor al 5%.
Fuerte	
BS1	La frecuencia del alelo es mayor de lo esperado para el trastorno
BS2	La variante se observó en un individuo adulto sano con la misma cigosidad para una enfermedad totalmente penetrante.
BS3	Los estudios funcionales in vitro o in vivo no muestran ningún efecto perjudicial sobre la función proteica o el plegamiento de la proteína resultante.
BS4	Falta de segregación en los miembros afectados de una familia.
Secundarias/Apoyo	
BP1	Missense en un gen para el cual se sabe que la enfermedad es producida por variantes que causan truncamientos en la proteína.
BP2	Observada en trans (en distintos alelos) con una variante patogénica para un trastorno dominante completamente penetrante u observado en cis (mismo alelo) con una variante patogénica en cualquier patrón de herencia.
BP3	Inserciones o deleciones dentro del marco de lectura (in-frame) en una región repetitiva sin función conocida.

BP4	Múltiples líneas/predictores de evidencia computacional sugieren que no hay impacto en el producto genético (criterios de conservación, impacto evolutivo, plegado, etc.)
BP5	Encontrada previamente en un caso con una base molecular alternativa para una enfermedad.
BP6	Reportada como benigna recientemente por una base de datos curada.
BP7	Una variante sinónima (silenciosa) para la que los algoritmos de predicción no predicen ningún impacto en la secuencia consenso de empalme ni la creación de un nuevo sitio de splicing. Además, el nucleótido no está muy conservado.

Clasificación de las variantes

Una vez establecidas las evidencias que la variante presenta, la clasificación dentro de las cinco categorías se consolida como una sumatoria de las evidencias presentes que se combinan para dar como resultado la clasificación de la misma.

Es interesante destacar que si bien las posibles combinaciones parecen independientes, siguen cierto patrón. Si tomamos, por ejemplo, solo valoraciones que indican potencial patogenicidad (PVS, PS, PM, PP) en líneas generales cada nivel de puntaje puede ser reemplazado por 2 o más del nivel inmediatamente inferior. Así una variante será patogénica si posee 2 PS, ó 1 PS y 2 PM, o 1 PS, 1 PM y 2 PP. Al existir menos códigos asignables, la clasificación de una variante como “benigna” es más simple y se logra con 1 BA ó 2 BS, mientras que una variante es catalogada como Probablemente benigna con 1BS y 1BP ó 2 BP.

Las combinaciones de puntaje que dan lugar a cada clasificaciòn se resumen en la siguiente tabla.

Tabla 4: Combinación de los criterios de acuerdo con las reglas de puntuación

Patogénica	i. 1 Muy fuerte (PVS1) Y a. ≥1 Fuerte (PS1–PS4) ó b. ≥2 Moderadas (PM1-PM6) ó c. 1 Moderada (PM1-PM6) y 1 de Apoyo (PP1-PP5) ó d. ≥2 de Apoyo (PP1–PP5) ii. ≥ 2 Fuertes (PS1-PS4) iii. 1 Fuerte (PS1-PS4) Y a. ≥3 Moderadas (PM1–PM6) ó b. 2 Moderadas (PM1–PM6) Y ≥2 de Apoyo (PP1–PP5) ó c. 1 Moderada (PM1–PM6) Y ≥4 de Apoyo(PP1–PP5)
Probablement	i. 1 Muy fuerte (PVS1) Y 1 Moderada (PM1–PM6)

e patogénica	ii. 1 Fuerte (PS1–PS4) Y 1-2 Moderadas (PM1–PM6) iii. 1 Fuerte (PS1–PS4) Y ≥2 de Apoyo (PP1–PP5) iv. ≥3 Moderadas (PM1–PM6) v. 2 Moderadas (PM1–PM6) Y ≥2 de Apoyo (PP1–PP5) vi. 1 Moderada (PM1–PM6) Y ≥ 4 de Apoyo (PP1–PP5)
Benigna	i. 1 Independiente/Única (BA1) ii. ≥2 Fuertes (BS1–BS4)
Posiblemente Benigna	i. 1 Fuerte (BS1–BS4) and 1 de Apoyo (BP1–BP7) ii ≥2 de Apoyo (BP1–BP7)
Significado incierto	i. No cumple con los criterios mencionados anteriormente. ii. Se contradice la evidencia de patogenicidad con reportes de carácter benigno de la variante.

Si bien la tarea de evaluar toda la evidencia para un número significativo de variantes por caso puede parecer una tarea ciclópea, el implementarlo de manera ordenada, siguiendo un camino lógico y sistemático para la recolección de información, como el que se propone la ACMG/AMP, puede facilitar enormemente la tarea.

Finalmente, es importante mencionar que para facilitar el trabajo en la interpretación y valoración de las variantes, se agregó al protocolo de procesamiento de datos, la clasificación de variantes según los criterios de ACMG de forma semiautomática fijando los distintos niveles de evidencia y realizando el cálculo subsiguiente para una primera clasificación. Además se han implementado en la plataforma los filtros necesarios para poder identificar rápidamente las variantes significativas según esta clasificación.

HPO (Human Phenotype Ontology)

Como ya se mencionó múltiples veces, a pesar de los avances en las tecnologías de secuenciación, sigue siendo difícil hacer un diagnóstico clínico preciso, debido, entre otras cosas, a las complejas y aún no comprendidas relaciones entre las variantes genéticas y los fenotipos clínicos que se observan en los pacientes. En la última década se difundieron ampliamente distintas herramientas utilizadas para mejorar la eficiencia de los diagnóstico de enfermedades mediante el análisis comparativo de los fenotipos observados en un paciente para el que se busca diagnóstico molecular, y los fenotipos predominantes observados en pacientes con variantes patogénicas confirmadas.

En este contexto y para facilitar el establecimiento de relaciones entre el fenotipo observado en el paciente y los potenciales genes responsables se incorporó a la plataforma, como herramienta de trabajo, ontología del fenotipo humano (HPO, Human Phenotype Ontology) que proporciona un vocabulario estandarizado de los rasgos fenotípicos encontrados en las enfermedades humanas [54].

Cada término en la ontología HPO describe un rasgo fenotípico. HPO contiene múltiples tipos de información para cada fenotipo, como el modificador de frecuencia y las

definiciones de los términos. Algunos ejemplos (en inglés del original):

- ***Memory impairment (HP:0002354)**: "An impairment of memory as manifested by a reduced ability to remember things such as dates and names, and increased forgetfulness."*
- ***High Palate (HP:0000218)**: "Height of the palate more than 2 SD above the mean (objective) or palatal height at the level of the first permanent molar more than twice the height of the teeth (subjective)."*
- ***Macrocephaly (HP:0000256)**: "Occipitofrontal (head) circumference greater than 97th centile compared to appropriate, age matched, sex-matched normal standards. Alternatively, an apparently increased size of the cranium."*
- ***Hirsutism (HP:0001007):** "Abnormally increased hair growth referring to a male pattern of body hair (androgenic hair)."*

El proyecto HPO está en continua evolución y esta ontología de términos es utilizada cada vez más en la literatura médica y por diferentes bases de datos de enfermedades como Orphanet [55], DECIPHER [56], y OMIM[32]. HPO contiene actualmente más de 13.000 términos y más de 156.000 anotaciones de enfermedades hereditarias.

Una de los elementos importantes de HPO, es que los términos se encuentran organizados como un gráfico acíclico dirigido (DAG) para describir las características fenotípicas y sus relaciones (la Figura 10). Esto permite categorizar los términos según su precisión qué es inversamente proporcional al número de genes con los que el término fenotípico se relaciona. Términos muy generales como “fiebre” se asocian a muchos genes, mientras que aquellos más precisos se asocian a uno o unos pocos, permitiendo en el caso de que el médico los observe reducir el número de genes candidatos a analizar de manera significativa.

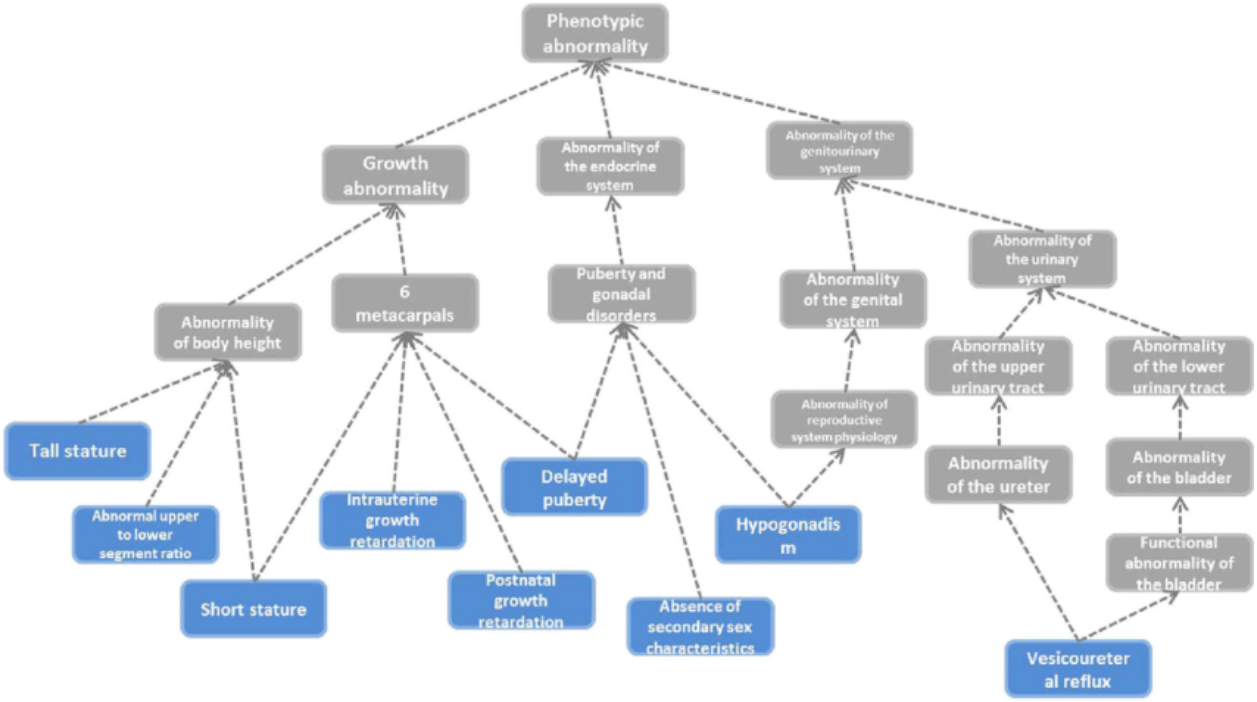

Figura 10: Un ejemplo de rasgo fenotípico (HP:0000118) que forma un gráfico acíclico dirigido (DAG), en el que los nodos representan términos de fenotipo y los bordes representan "subclase de " relaciones entre términos de fenotipo.HPO (Human Phenotype Ontology)[54].

Recientemente se han desarrollado diferentes herramientas que utilizan HPO (como phenomizer [59]) y que permiten a partir de un conjunto de síntomas observado determinar un conjunto de genes candidatos. La clave de estos métodos radica en su capacidad para comparar la historia clínica del paciente con el fenotipo esperado, y es en este contexto donde los estudios y desarrollos de similitud semántica del fenotipo se ha convertido en un área de investigación en auge [57, 58].

En la plataforma, la incorporación de HPO permite aplicar filtros a las variantes (genes) de acuerdo a los síntomas de la HC que representen términos contenidos en HPO. Esta restricción, de la búsqueda a aquellos genes directamente asociados con el cuadro clínico del paciente es muy útil para el médico genetista para identificar los potenciales genes causales de la enfermedad y ayudar al diagnóstico del paciente.

Evaluación de un perfil de riesgo

Comparación entre las enfermedades mendelianas y las complejas

Para comenzar esta sección hablaremos brevemente de las diferencias entres las enfermedades mendelianas y las enfermedades que denominaremos complejas, con el objetivo de explicar las bases del cálculo de un perfil riesgo. Dentro de estas enfermedades complejas, englobamos las enfermedades cardíacas, la diabetes -particularmente la de tipo 2-, las neurodegenerativas (alzheimer, parkinson), algunas autoimmunes/autoinflamatorias como la enfermedad de Crohn, y diferentes tipos de cáncer.

Ya se ha mencionado la base genética de las enfermedades mendelianas. Estas son enfermedades en las que, usualmente, un solo gen defectuoso es suficiente para que la misma se desarrolle. En las enfermedades mendelianas, además, prácticamente todas las personas que tienen la enfermedad tienen una mutación en un gen específico, y no hay nadie que tenga la enfermedad sin la mutación, o viceversa. Se dice que estos genes son altamente penetrantes.

Las enfermedades mendelianas son típicamente poco frecuentes, tiene baja prevalencia en la población debido a que están sometidas a una gran presión de selección. La prevalencia suele ser inferior al 1% en la población.

En contraposición, en las enfermedades complejas, *hay un componente tanto genético (en el que múltiples genes están involucrados), cómo ambiental*. Por ejemplo, la dieta, el estilo de vida, la exposición a químicos ambientales, o el uso/abuso de ciertas medicinas. Entonces en el caso de enfermedades complejas, algunas personas que tienen una dada mutación pueden desarrollar la enfermedad, pero hay otras que tienen la enfermedad sin poseer la mutación. También puede ocurrir que personas tengan la mutación y no desarrollen la enfermedad.

Una característica de una enfermedad compleja es que usualmente a nivel clínico la persona presenta el fenotipo de la enfermedad, pero no necesariamente se encuentra una razón o correspondencia con el genotipo. Es decir, donde a nivel genético no existe la predisposición y aún así presentan la enfermedad, tal vez por otra razón. Por ejemplo, podría ser cáncer de mamá, celiaquía, etc.

Otra característica de los desórdenes complejos es lo que se denomina *penetrancia incompleta*, esto implica que no todas las personas genéticamente susceptibles (que poseen genotipo patogénico) desarrollarán la enfermedad. Por ejemplo, una persona que tiene una mutación en un gen puede tener una predisposición genética pero no desarrollar nunca la enfermedad.

Las enfermedades complejas también poseen *expresividad variable*. Los individuos con el mismo genotipo también pueden mostrar diferentes grados del mismo fenotipo. La expresividad es el grado en que la expresión de los rasgos difiere entre los individuos. A diferencia de la penetrancia, la expresividad describe la variabilidad individual, no la variabilidad estadística entre una población de genotipos.

Los síndromes complejos también poseen una amplia heterogeneidad genética. Es decir, que las mutaciones en diferentes genes pueden conducir a la misma enfermedad. Así que, de nuevo, a nivel clínico estas enfermedades pueden parecer indistinguibles. Pero a nivel genético pueden deberse a mutaciones genéticas en diferentes genes. Ya sabemos que para el cáncer de mama existe el gen BRCA1, el gen BRCA2, el TP53 y muchos otros en los que las mutaciones pueden conducir al cáncer de mama.

Finalmente algunos de los signos (o características) que podrían sugerir una condición genética subyacente para una enfermedad compleja son:

La edad de aparición es más temprana de lo esperado. Un ejemplo aquí es un infarto de miocardio o un ataque al corazón que ocurre en una persona de 25 años contra uno que ocurre en una persona de 80 años, que podría ser más típico.

Cuándo una condición que ocurre en el sexo menos afectado. Por ejemplo, para el cáncer de mama, por supuesto, las mujeres son las principales afectadas, pero si se tiene un hombre afectado, podría ser un indicio de que hay una etiología genética.

La historia familiar, si existe una historia familiar de una enfermedad, en particular una enfermedad que ocurre en múltiples generaciones, tal vez a una edad más temprana de aparición que lo habitual. Eso podría ser un indicio de que hay una base genética.

Y por último, cuando la enfermedad ocurre en ausencia de factores de riesgo conocidos. Por ejemplo, una enfermedad como la diabetes, donde la mayoría de las personas con diabetes también tienen obesidad, es una comorbilidad, y quizás una causa de la diabetes. Pero entonces si una persona delgada tiene diabetes se podría pensar que tal vez hay alguna etiología genética.

Cuando hablamos de enfermedades complejas comunes y de pruebas genéticas para estas enfermedades, normalmente nos referimos a test de predisposición o de cálculo y evaluación de riesgo. Estas son pruebas que se pueden usar para predecir si una persona es propensa a contraer una determinada enfermedad compleja. Para esto se utilizan los estudios de asociación genómica que buscan descubrir relaciones entre variantes alélicas específicas y alguna característica fenotípica de interés.

Estudios de asociación amplia del genoma (GWAS)

El acrónimo GWAS hace referencia a Genome-Wide Association Studies, estudios a nivel de genomas completos. La teoría que subyace a los GWAS, es que las enfermedades complejas se deben al efecto acumulativo de muchas variantes de riesgo comunes (con alta frecuencia) en un individuo. Si bien cada una de ellas son polimorfismos comunes en la población y puede aportar muy poco al desarrollo de una dada patología, su efecto acumulativo puede derivar (en conjunción con otras variables, típicamente ambientales) en el desarrollo de la enfermedad.

Los GWAS son estudios en los que se analizan cientos de miles polimorfismos de nucleótido único (SNPs, Single Nucleotide Polymorphism) en un experimento. Directa o indirectamente se analizan todas las variantes comunes y se busca asociarlas a alguna patología de interés.

En el genoma humano compuesto de unos 3 mil millones de pares de bases, la mayoría de las secuencias de ADN de los individuos coinciden, pero alrededor del 0,1% son SNPs, que es donde radica la diferencia. Tales SNPs causan la diversidad genética de un individuo y “algunos de ellos” la predisposición a diferentes enfermedades.

A la hora de diseñar un GWAS se tiende a usar un enfoque o diseño de estudios epidemiológicos que son de naturaleza observacional.

Estudios de observación comunes son:

Cohorte: en este diseño de estudio se tiene un grupo de individuos sanos a los cuales se les hace seguimiento a lo largo del tiempo. Luego se observa quién desarrolla la enfermedad. Y luego se analiza entre las personas con la enfermedad cuántos de ellos tienen la variante genética estudiada. Este tipo de estudios presentan algunos inconvenientes inherentes: por ejemplo si la enfermedad en estudio es relativamente rara se necesita inscribir un gran número de sujetos en el estudio para poder identificar un número suficiente de casos con la enfermedad. Y si se tiene una enfermedad que tiene un largo período de latencia se tendrá que seguir a los individuos por un tiempo muy largo para que desarrollen la enfermedad.

Casos-Control: son aquellos en los que se tiene un grupo de individuos que ya tienen la enfermedad (Casos). Y luego se consigue un grupo de personas que no tienen la

enfermedad (Controles), y se hacen comparaciones entre estos grupos. Éste es el tipo de estudios preferidos para los GWAS.

Analizar e interpretar GWAS

Una vez que se tiene genotipificada (determinados sus SNPs) a la población en estudio, básicamente lo que se hace es comparar a las personas que tienen la enfermedad contra las que no, y luego se busca asociar la presencia de algún SNP particular con el desarrollo de la misma.

Por ejemplo: si en un tipo de estudio Casos-Controles, tenemos que 50% de los Casos son portadores de la variante de interés posiblemente asociada a la enfermedad pero sólo el 17% de los controles son portadores de dicha variante, entonces podríamos decir que los Casos son más propensos a tener la variante que los Controles. Pero para que esta premisa sea correcta se necesita un tipo de prueba estadística que determine si esta es una diferencia significativa o no.

A esto se agrega la complejidad de que los humanos son diploides, cada persona lleva dos copias de cada gen. Así que aunque se pueda clasificar a los individuos como portadores o no de una variante, se tiene que saber si llevan una o dos copias de esa variante, es decir su genotipo es homocigota para la variante, heterocigota, u homocigota para la referencia (en los casos de que en la posición analizada no haya variante).

Entonces, para analizar los resultados de un GWAS se cuentan cuántas personas son homocigotas, heterocigotas y homocigotas para la referencia, tanto en el grupo de casos como de controles, y lo que se hace es determinar la proporción de individuos enfermos y no enfermos con cero copias, una copia o dos copias del alelo variante. Básicamente se cuentan los genotipos encontrados.

Una vez determinadas las proporciones mencionadas (lo que se resume en una tabla de contingencia) se determina si hay una asociación estadísticamente significativa entre el alelo alternativo y la enfermedad de interés.

Se puede realizar una prueba de chi-cuadrado para determinar si estas diferencias se deben al azar o son significativas (donde el valor p es menor a un umbral determinado). Una prueba de chi-cuadrado es una prueba de hipótesis que compara la distribución observada de los datos con una distribución esperada de los datos.

Si el p-value es menor o igual que el nivel de significancia, normalmente un p-value menor de 0.05, se rechaza la hipótesis nula y se concluye que hay una asociación estadísticamente significativa entre las variables. Indicar que hay menos del 5% de probabilidad de que se haya observado esa distribución sólo por azar, esto significa que hay una pequeña probabilidad, alrededor del 5%, de que sea un falso positivo. Es más que probable que sea un verdadero positivo, y llamamos a esta *asociación estadísticamente significativa*.

Este análisis se realiza para todos los SNPs determinados, buscando entonces cuáles son aquellos que se encuentran asociados significativamente con la enfermedad.

Una vez determinado que existe una asociación estadísticamente significativa, lo que se intenta determinar es, si una persona tiene una variante genética, ¿cuál sería su riesgo de desarrollar la enfermedad?

Para ello a partir ***del estudio de Casos y Controles*** se calcula una probabilidad relativa, denominada "Odds Ratio (OR)", que en castellano podríamos traducir como “cociente de chances”.

El Odds Ratio (OR) es un ***estimador del riesgo relativo*** de poseer la enfermedad si uno posee algún alelo alternativo, respecto de la población de referencia. El OR se calcula como el cociente entre la probabilidad (odds) de desarrollar la enfermedad si uno posee el alelo alternativo, respecto de la chance de desarrollar la enfermedad para el alelo de referencia, que sería la probabilidad que tiene la población de referencia. El odds ratio se define como la relación entre el número de casos y controles para cada condición.

$$OR = \frac{Casos\ Alt\ /\ Controles\ Alt}{Casos\ Ref\ /\ Controles\ Ref}$$

La interpretación del OR es que es una medida de cuántas veces más probable es que el individuo que posee el SNP desarrolle la enfermedad, respecto de la población general.

Entonces si se utiliza el siguiente ejemplo tenemos que:

Tabla 5: Cálculo de OR

/	Casos	Controles	Odds	Odds Ratio	Interpretacion
Alleles	100	100	--	--	--
TT	55	25	2.2	2.2/0.3=7.3	tiene 7 veces más chances de desarrollar la enfermedad
TC	35	40	0.8	0.9/0.3=3	tiene 3 veces más chances de desarrollar la enfermedad
CC	10	35	0.3	referencia	normal

Entonces para evaluar la contribución genética a las enfermedades complejas, debemos buscar en el genoma aquellos SNPs que mediante los GWAS han sido asociados a estas patologías (alelos de riesgo) y luego mediante su OR evaluar el impacto de los mismos en el riesgo de desarrollarla. Los datos de los SNPs asociados se obtienen directamente de GWAS Catalog [60] y son asociados a los SNPs encontrados en el VCF.

GWAS Catalog

El acrónimo GWAS hace referencia a Genome-Wide Association Studies, estudios a nivel de genomas completos. Los estudios de asociación genómica buscan descubrir relaciones entre variantes alélicas específicas y alguna característica fenotípica de interés. Más adelante en la sección *Evaluación de un perfil de riesgo* se explicará con mayor detalle este tipo de estudios.

Estos estudios son particularmente relevantes para estudiar características con patrones de herencia no mendelianos, como las enfermedades complejas mencionadas en el apartado anterior.

GWAS Catalog [60] es una base de datos de libre acceso de las asociaciones entre traits (muchos de ellos enfermedades) y variantes (principalmente SNP). Las asociaciones de genotipo-fenotipo, junto con la información del estudio GWAS, se extraen manualmente de la literatura y se introducen en el catálogo. Esta información es curada por expertos, y luego se pone a disposición gratuitamente y se puede buscar en el sitio web GWAS Catalog , para permitir a los científicos interpretar los datos con precisión.

GWAS Catalog fue fundado por el Instituto Nacional de Investigación del Genoma Humano (NHGRI) en 2008, y desde 2010 ha sido una colaboración entre el EBI y el NHGRI, en respuesta al rápido aumento del número de estudios GWAS publicados.

Es importante destacar que los estudios GWAS se incluyen en el catálogo si cumplen ciertos requisitos, por ejemplo:
Si se realizó la genotipificación de las muestras de una gran cantidad de genomas y si en el estudio se analizaron más de 100.000 SNPs. Esto incluye los GWAS publicados anteriormente que se incorporan a nuevos análisis (meta-análisis).
Para cada uno de estos estudios, se incluyen las asociaciones de caracteres (traits) - variantes sí:

- Tienen un p-value <1,0 x 10-5 en la población total (GWAS inicial + replicación)
- Se extrae la variante más significativa de cada locus independiente

GWAS Catalog contiene una gran cantidad de datos y está diseñado para ser fácilmente accesible a los científicos que deseen utilizar los datos. Se puede buscar por SNP, por gen, por característica, etc, y contiene además visualizaciones útiles de la relación entre traits y variantes. Todos los traits y enfermedades están mapeados en una ontología para mejorar la capacidad de búsqueda.

En GWAS Catalog los resultados de todos los trabajos de asociación son subidos en un formato estandarizado, permitiendo su análisis en conjunto. Para cada entrada se cuenta con los siguientes campos, en negrita, los que campos que vamos que usamos para seleccionar las variantes que usaremos en el cálculo de riesgo:

- DATE_ADDED_TO_CATALOG

- PUBMEDID
- FIRST_AUTHOR
- **DATE (fecha de publicación del estudio)**
- JOURNAL
- LINK
- STUDY
- DISEASE_TRAIT
- INITIAL_SAMPLE_SIZE
- REPLICATION_SAMPLE_SIZE
- REGION
- CHR_ID
- CHR_POS
- REPORTED GENE(S)
- MAPPED_GENE
- UPSTREAM_GENE_ID
- DOWNSTREAM_GENE_ID
- SNP_GENE_ID
- UPSTREAM_GENE_DISTANCE
- DOWNSTREAM_GENE_DISTANCE
- **STRONGEST_SNP-RISK_ALLELE (alelo de riesgo reportado).**
- **SNPS (identificador de la variante en cuestión).**
- MERGED
- SNP_ID_CURRENT
- CONTEXT
- INTERGENIC
- **RISK_ALLELE_FREQUENCY (frecuencia del alelo de riesgo en la población en cuestión).**
- **P-VALUE (valor de significancia de la asociación en cuestión).**
- PVALUE_MLOG
- **P-VALUE (TEXT) (indica si el p-value reportado está asociado a algún subgrupo particular -género, fumadores, etc.).**
- **OR or BETA.**
- **95% CI (TEXT) (intervalo de confianza asociado al p-value. Indica además si el valor en el campo anterior es un OR o un beta).**
- PLATFORM_SNPS_PASSING_QC
- CNV
- **MAPPED_TRAIT (nombre del rasgo asociado con vocabulario estándar de EFO (experimental factor ontology). Permite recolectar estudios asociados a los rasgos de interés en forma precisa).**
- MAPPED_TRAIT_URI
- STUDY_ACCESSION
- GENOTYPING_TECHNOLOGY

Estos estudios brindan una oportunidad sin precedentes para investigar el impacto de las variantes comunes en las enfermedades complejas; sin embargo, la búsqueda e identificación de los GWAS publicados puede resultar difícil, y el gran potencial que contienen los datos de estas publicaciones puede resultar inaccesible para los investigadores, si estos no son catalogados y resumidos sistemáticamente.

Evaluación de un perfil farmacogenómico

La farmacogenómica se puede definir básicamente como el uso de la información genómica de una persona para mejorar la eficacia, o reducir los efectos secundarios, de los fármacos en su organismo. Se puede pensar la respuesta de un individuo a un fármaco como si se tratara de una enfermedad compleja, ya que esta depende de factores genéticos y no genéticos (que en este caso son principalmente las condiciones fisiológicas del individuo). Una de las principales diferencias es que la asociación entre genes y enfermedades complejas tienden a tener efectos muy pequeños. Mientras que los genes relacionados con la farmacogenómica tienden a tener efectos algo mayores.

Los fármacos no siempre funcionan de igual manera en las personas, se sabe que las tasas de respuesta a las drogas varían entre el 25% y el 80%. Por ejemplo, algunos analgésicos tienen, en general, muy buena eficacia, mientras que muchos, o la mayoría, de los medicamentos oncológicos no, y la respuesta depende extremadamente del genoma del individuo y del tumor.

Por otro lado, muchos medicamentos provocan reacciones adversas Una reacción adversa a un medicamento o ADR, es por definición, algo que no es intencional, es una consecuencia no deseada y usualmente es nociva, pudiendo llegar en casos extremos a la muerte del paciente. Se estima, que del 1,5 al 2% de las personas que toman un medicamento pueden desarrollar una reacción adversa. Así que, aunque individualmente son bastante raras, colectivamente, pueden ser bastante comunes.

Comencemos con algunos antecedentes sobre los factores genéticos que afectan a la farmacocinética y la farmacodinámica, dos parámetros fundamentales para comprender cómo un medicamento interactúa con un organismo. La farmacocinética, analiza cómo cuando se toma una droga, de qué manera la concentración de la misma cambia a medida que se mueve a través del organismo. La farmacocinética, se divide en los siguientes pasos resumido en el acrónimo ADME de A*bsorción, Distribución, Metabolismo y Eliminación*. Entonces cuando una droga entra en el cuerpo, es absorbida. Entra ya sea por vía oral, intravenosa o vía aérea. Una vez dentro del organismo la droga necesita ser distribuida por todo el cuerpo. Lo hace a través del sistema circulatorio, con la ayuda de proteínas transportadoras, por ejemplo, hay transportadores de solutos como la proteína SLCO1B1.

La droga es entonces metabolizada o biotransformada. Este proceso se lleva a cabo en el hígado con la ayuda de ciertas enzimas. La familia más grande de estas enzimas son los citocromos de tipo P450 (CYPs). Finalmente el producto de la droga, una vez que es metabolizada, es eliminado. Se excreta del cuerpo, a través de los riñones, de las heces o del aliento. En cada uno de estos pasos intervienen proteínas y por ende existe un potencial efecto farmacogenómico. Actualmente, la misma se concentra en los pasos de distribución y metabolismos, dependientes principalmente de los CYPs.

Casos de aplicación de Farmacogenómica

La mayoría de los medicamentos que se utilizan en la actualidad son metabolizados por los citocromos P450. Los más relevantes son los citocromos P450, CYP3A4, CYP3A5 y los CYP2D6, CYP2C8, CYP2C9 y CYP1A2. Las enzimas codificadas por estos genes son polimórficas, son muy variables, por lo que, diferentes individuos tendrán diferente actividad enzimática, y por tanto diferente respuesta a los fármacos.
Para estudiar la farmacogenómica en relación con los CYPs, se suele clasificar a las personas según su actividad enzimática en:

- Metabolismo ultrarrápido.
- Metabolismo extensivo. Esto es el metabolismo normal.
- Metabolismo intermedio, es decir, que tienen un metabolismo algo disminuido.
- Metabolismo bajo, no es bueno metabolizando drogas.

Se sabe que hay cientos de variantes genéticas que pueden llevar a la reducción o a la pérdida completa de función de estos genes del citocromo P450, para comprender mejor el concepto, a continuación, se plantean algunos ejemplos:

Codeína
La codeína es metabolizada por la enzima CYP2D6. La codeína en sí misma es en realidad una pro-droga. El metabolito activo es la morfina. La enzima CYP2D6 es la que metaboliza la codeína en morfina. La morfina se une a su objetivo (target o blanco molecular) y produce su efecto analgésico.

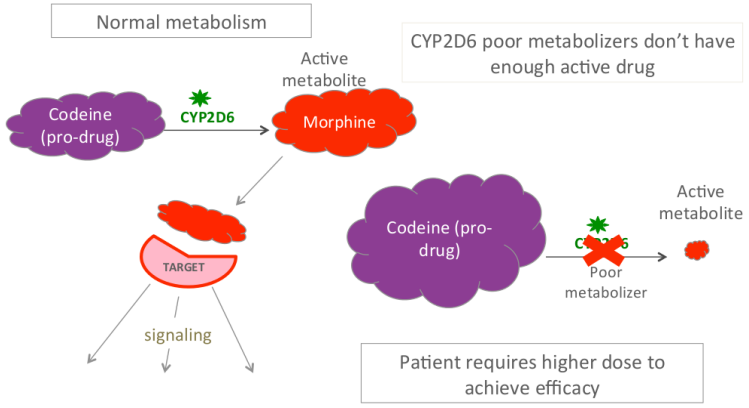

Figura 11: Metabolismo de la Codeína. [84] Genomic and Precision Medicine.

Entonces si un individuo tiene una mutación en el gen CYP2D6 que provoca su pérdida de función total o parcial, la persona se convierte en un mal metabolizador de la codeína. La enzima no se expresa lo suficiente y ya no es posible metabolizar la codeína. Esto aumenta su concentración pero disminuye la cantidad del metabolito activo, la morfina, reduciendo la eficacia del fármaco. Esto hace que la persona requiera una dosis mayor para mejorar su eficacia.

Warfarina

La propia warfarina es un compuesto activo. Esta droga es un anticoagulante que se une al receptor y produce una alteración en la coagulación. Esa es la función de la warfarina y la misma es metabolizada por la enzima CYP2C9. El producto de la catálisis es un metabolito inerte, que luego se elimina.

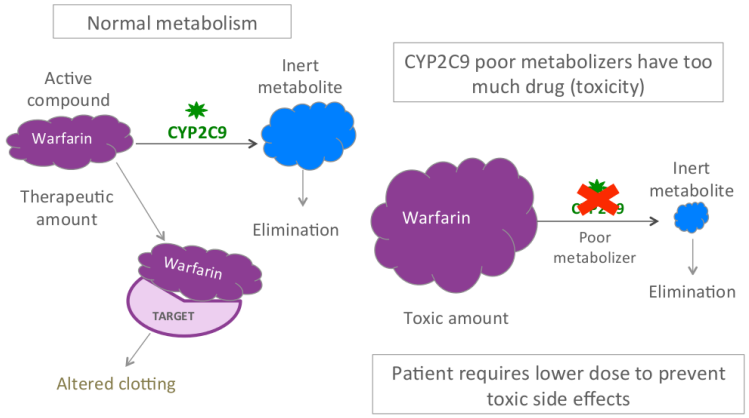

Figura 12: Metabolismo de la Warfarina. [84] Genomic and Precision Medicine.

Si se tiene una mutación en el gen CYP2C9 que produce la pérdida de función del gen, la cantidad de warfarina se acumulará en el sistema y debido a que este es un compuesto activo, se convertirá en tóxico. Así las personas que son malos metabolizadores de la warfarina debido a una pérdida de función del gen CYP2C9 requerirán una dosis menor para prevenir los efectos secundarios tóxicos.

Ivacaftor

Un último ejemplo es el caso del fármaco Ivacaftor, que se utiliza para tratar la fibrosis quística (CF). Un individuo sano tiene canales de cloruro en sus células epiteliales. Este canal está regulado por una proteína de membrana expresada por el gen CFTR. En su estado normal el canal de cloruro permite el transporte normal de iones a la célula.

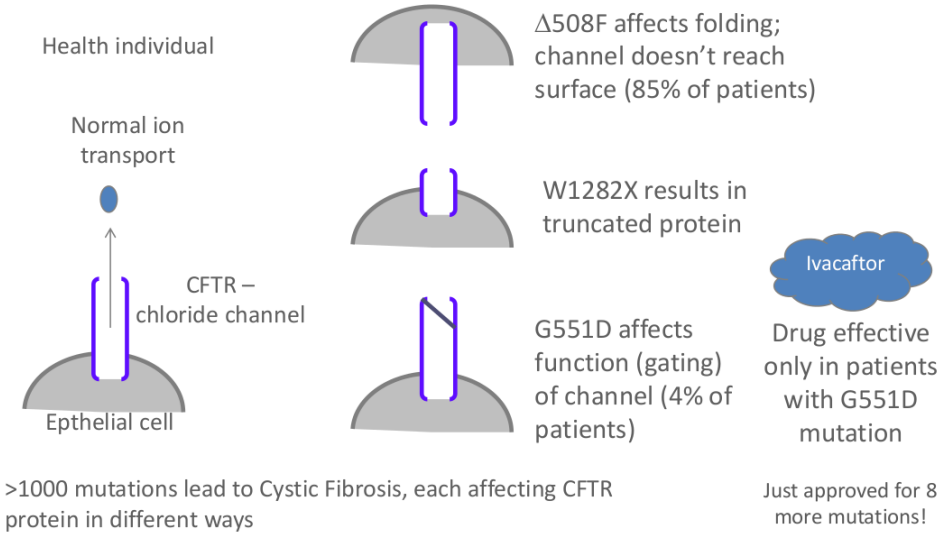

Figura 13: Metabolismo de Ivacaftor. [84] Genomic and Precision Medicine.

Se conocen múltiples mutaciones del gen CFTR [61] que pueden conducir al desarrollo de la enfermedad de la fibrosis quística. Cada una de ellas afecta a esta proteína de una forma ligeramente diferente. La variante más común asociada a CF es la llamada Delta F508 o F508del-CFTR (rs113993960). Esta variante afecta al plegamiento de la proteína de modo tal que el canal ni siquiera llega a la superficie de la célula. Esta es la variante responsable de alrededor del 85% de los pacientes con fibrosis quística.

Hay otras variantes menos comunes pero igualmente deletéreas. Por ejemplo,la W1282X, que produce el truncamiento de la proteína, por lo que se puede ver que el canal de cloruro es un poco más corto.

La variante llamada G551D, que es una sustitución de aminoácidos, afecta la apertura de este canal. Esta variante es responsable de alrededor del 4% de los pacientes con CF. El Ivacaftor es una droga que fue desarrollada específicamente para atacar este defecto. Por lo que la droga sólo funciona en personas con este tipo de defecto. No funciona en el tipo más común, en el que el canal de cloruro no llega a la superficie o en otros tipos.

Entonces este es un ejemplo en donde sería conveniente primero hacer un test de farmacogenómica para asegurarse de que la droga funcione, para que sea eficaz.

PharmGKB

PharmGKB [62] es un proyecto que recopila y organiza información clínica relacionada con la farmacogenómica, o sea cómo los genes/variantes se relacionan o influyen en la respuesta a los fármacos. PharmGKB incluye guías de dosaje y asociaciones gen(variante)-fármaco con potencial relevancia clínica. Además, contiene información de cuáles son tanto las vías metabólicas como las de acción (incluyendo blanco principal y secundarios u *off-target*) de cada uno de los fármacos aprobados por la FDA.

Las anotaciones de PharmGKB presentan un breve resumen de las recomendaciones de dosificación basadas en el genotipo. Estas pautas/recomendaciones de dosificación de los medicamentos están basadas en las guías publicadas por Clinical Pharmacogenetics Implementation Consortium (CPIC) [63].

El CPIC es un consorcio internacional dedicado a facilitar el uso de pruebas farmacogenéticas para el cuidado de los pacientes. Una de las barreras para la implementación de las pruebas farmacogenéticas en la clínica, es la dificultad de traducir los resultados de las pruebas genéticas de laboratorio en decisiones de prescripción accionables para los medicamentos afectados. El objetivo del CPIC es justamente abordar esta problemática, mediante la creación, el curado y la publicación de directrices de práctica clínica de genes/drogas detalladas, actualizadas, revisadas por pares basándose en la evidencia disponible.

Las recomendaciones terapéuticas del CPIC se basan en la ponderación de la evidencia de una combinación de datos funcionales y clínicos, así como en algunas directrices de consenso específicas de la enfermedad.

Las **principales anotaciones de PharmGKB** presentan un breve resumen de las recomendaciones de dosificación basadas en el genotipo. Entre las anotaciones que podemos encontrar se encuentran:

i) La información de prescripción, que incluye pautas clínicas para ajustar el tratamiento de ciertos medicamentos en base a la información genómica de la persona. Algunas directrices específicas pueden ser, por ejemplo, cómo ajustar la dosis de un medicamento, o recomendar la utilización de medicamentos alternativos.

ii) Las etiquetas de los medicamentos de ciertos fármacos contienen consejos sobre cómo ajustar la dosis basándose en la información genética, o información sobre la posibilidad de efectos adversos en personas con ciertas variantes genéticas, o información sobre las proteínas implicadas en la metabolización del fármaco en el cuerpo. Además se les asigna un nivel de información farmacogenómico (PGx) que incluye si recomienda o requiere la realización de análisis genómicos del paciente antes de suministrar el medicamento.

iii) También se puede encontrar información sobre las vías metabólicas, es decir, cómo se metaboliza un medicamento en el cuerpo, cómo funciona un medicamento en el cuerpo, o ambas cosas. La información de PharmGKB con respecto a esto se centran en la farmacocinética o en la farmacodinámica de un medicamento, y describen los genes clave que codifican las proteínas implicadas en estos procesos.

iv) Los VIPs (Very Important Pharmacogenes) son descripciones de genes que son particularmente importantes en el campo de la farmacogenómica. La información disponible para cada VIP incluye información de las proteínas que expresa, las relaciones que tiene con distintas enfermedades, e información detallada sobre las variantes genéticas (o haplotipos) particularmente importantes para la farmacogenómica del gen en cuestión.

v) Las anotaciones clínicas, finalmente, resumen todas las anotaciones de PharmGKB de las pruebas publicadas de la relación entre una variante genética particular y un medicamento. Esta información es la esencia de PharmGKB. A estas anotaciones, se les da una calificación por parte de PharmGKB dependiendo de cuánta evidencia publicada haya de la relación y la calidad de esa evidencia. En otras palabras, cuán probada e importante es la asociación variante-fármaco.

Estos niveles de evidencia son:

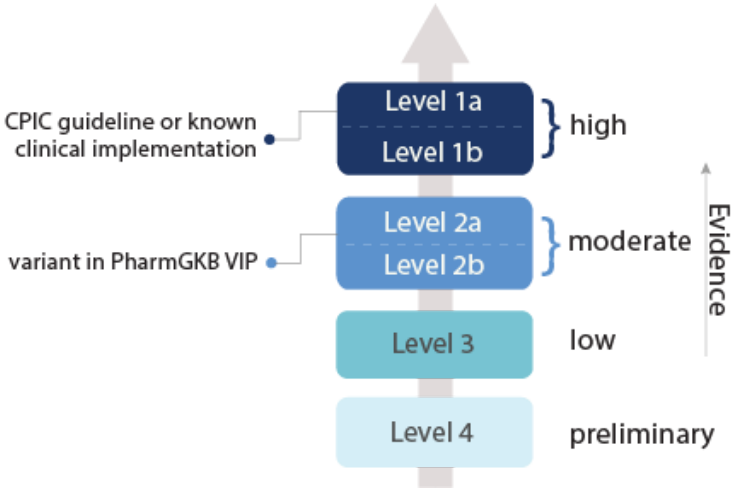

Figura 14: Niveles de evidencia. PharmGKB [62]

Nivel 1A: Este nivel de evidencia se le asigna a la relación variante - droga validada por el CPIC.
Nivel 1B**:** Este nivel de evidencia se le asigna a la relación variante - droga en la que la preponderancia de las pruebas muestra una asociación. La asociación debe ser replicada en más de una cohorte con valores p significativos, y preferiblemente tendrá un fuerte tamaño de efecto.
Nivel 2A: Este nivel de evidencia se asigna a la relación variante - droga cuando la variante se encuentra en algunos de los genes VIPs (Very Important Pharmacogene). Las variantes del nivel 2A se encuentran en fármacogenes conocidos, por lo que es más probable que tengan importancia funcional.
Nivel 2B: la clasificación 2B se da cuando las pruebas de asociación son moderadas. La asociación debe ser replicada, pero puede haber algunos estudios que no muestran significancia estadística.
Nivel 3: en este nivel se indica que la relación variante - droga está basada en un único estudio significativo (aún no replicado) o se corresponde a un conjunto de estudios pero sin pruebas claras de una asociación.
Nivel 4: esta clasificación se asigna cuando la relación variante droga se debe a un estudio no significativo o pruebas de ensayo in vitro, molecular o funcional solamente.

Aquí es pertinente que mencionar en nuestro protocolo de procesamiento de los datos de NGS, en la sección de anotación funcional, anotamos cada variante con el nivel evidencia de asociación variante-droga, y luego en el proceso de priorización solo nos quedamos con las de nivel de evidencia 1A,1B, 2A y 2B.

Para **trabajar con PharmGKB** se pueden tomar dos puntos de partida

- Tener un fármaco de interés y buscar/analizar cuáles son los genes/variantes asociados (o relevantes) para el mismo.

- A partir de las variantes de un genoma/exoma/panel dado analizar cuál(es) poseen información y efecto farmacogenómico.

En el protocolo definido para este trabajo de tesis se definió la anotación funcional de las variantes en estudio con PharmGKB, tomando como punto de partida la segunda opción mencionada.
Esto comprende lo que sería una "**anotación farmacogenómica de un conjunto de variantes**". La anotación comprende el resumen de la asociación entre una sola variante genética y una respuesta a la droga.

Para finalizar, se presenta un ejemplo. Para la variante rs28399504 cuya posición genómica es chr10:96522463, sus alelos pueden ser A>G / A>T y se encuentra en el gen CYP2C19, la anotación clínica de PharmGKB dice que existe una asociación entre esta variante y el fármaco clopidogrel que se utiliza en el tratamiento de enfermedades cardiovasculares, como el síndrome coronario agudo. Con un nivel de evidencia 1A, PharmGKB informa que:

- Los pacientes con el genotipo AA 1) pueden presentar un mayor metabolismo de clopidogrel y 2) pueden tener un menor riesgo de eventos cardiovasculares secundarios cuando son tratados con clopidogrel en comparación con los pacientes con el genotipo GG y AG.
- Los pacientes con el genotipo AG 1) pueden tener un metabolismo deficiente de clopidogrel y 2) pueden tener un mayor riesgo de eventos cardiovasculares secundarios cuando son tratados con clopidogrel en comparación con los pacientes con el genotipo AA.
- Los pacientes con el genotipo GG 1) pueden tener un metabolismo deficiente de clopidogrel y 2) pueden tener un mayor riesgo de eventos cardiovasculares secundarios cuando son tratados con clopidogrel en comparación con los pacientes con el genotipo AA.

Resultados

B-Platform

Como parte de este trabajo de tesis se desarrolló un software, que denominamos *B-Platform*, para el análisis, la clasificación y priorización de variantes. El mismo es una plataforma de análisis genómico que permite analizar datos de secuenciación, principalmente de tipo de NGS, de muestras humanas haciendo fácil la priorización y análisis de variantes. La plataforma permite analizar estudios simples, considerando un solo paciente, y/o múltiples, por ejemplo en el caso de trabajar con tríos (padre, madre e hijo), y otros tipos de diseños experimentales.
Una vez que los datos (VCF anotado) se encuentran cargados en la plataforma, la misma permite (entre otras cosas):

- Realizar un análisis estadístico de los datos derivados de los procesos de secuenciación, mapeo y el llamado de variantes;
- Realizar consultas predeterminadas y programar filtros avanzados para la priorización de genes y variantes de interés;

- Buscar genes por su relación con síntomas, de acuerdo a la nomenclatura estandarizada establecida por el Human Phenotype Ontology [54], restringiendo la búsqueda a aquellos genes directamente asociados con el cuadro clínico del paciente.
- Tener acceso directo a links de bases de datos externas comúnmente utilizados por la comunidad científica (OMIM[32], ClinVar[38], GnomAD[53], ENSEMBL[33], etc.).
- Trabajar con paneles de genes personalizados (y definir los mismos).
- Exportar los datos para la realización de un reporte.

B-Platform como herramienta para ayudar en el Diagnóstico

Introducción

Poder llegar a un diagnóstico ayuda a proporcionar una explicación y terminar con la comúnmente llamada "odisea diagnóstica". Ésta situación se da cuando un paciente que parece tener un trastorno mendeliano, se le realizan análisis bioquímicos, radiológicos, y muchas otras pruebas estándar para tratar de llegar o confirmar un diagnóstico, y no se tiene éxito o se llega a un diagnóstico erróneo. Esto resulta, muchas veces, en un tratamiento ineficaz, la progresión de la enfermedad, la frustración en la familia, un posible diagnóstico tardío, y la acumulación de gastos médicos sustanciales. El paciente se encuentra entonces en una migración constante de médico a médico y de prueba en prueba en el intento de llegar a un diagnóstico.

En el contexto de la secuenciación de próxima generación (NGS) cuando se hace mención al dilema del diagnóstico, se refiere a la posibilidad de secuenciar al paciente para intentar encontrar una variante en un gen asociado al trastorno que padece, siendo la misma potencialmente diagnóstica y de este modo terminando con la odisea del paciente.

Como ya se mencionó anteriormente en la introducción, con las tecnología de NGS, se puede obtener la secuencia completa del ADN de una persona, es decir, todo el genoma, a esto se lo denomina Whole Genome Sequencing (WGS), o se puede solo obtener la secuencia de ADN de “todos” los genes, es decir, solo secuenciar la regiones que codifican proteínas, a este procedimiento se los llama Whole Exome Sequencing (WES). También se puede seleccionar para secuenciar un determinado grupo de genes de interés, usualmente asociado a la patología de interés, a esto se los llama secuenciación de un Panel de Genes.

Se podría comenzar con la secuenciación del genoma completo, en cuyo caso se encontrarán aproximadamente 4,5 millones de variantes. O se puede empezar con la secuenciación del exoma completo, en cuyo caso se tendrá hasta 100.000 variantes. O quizás secuenciando un panel de genes candidatos en el cual, dependiendo del tamaño del panel, se podrá tener entre cientos y miles de variantes. En todos los casos el desafío es encontrar las variantes responsables de la patología y por ende diagnósticas.

En la actualidad, en general, se están secuenciando exomas antes que genomas y paneles. Esto se debe a que las mutaciones que encontramos en los exones son más fáciles de entender debido al código genético, su efecto es más predecible. No es que no se sepa

nada sobre la variantes en regiones no codificantes, sólo que se sabe mucho menos y la predicción de cuál será el efecto de una variante tendrá mucho menos éxito que si fuera sobre una variante que está dentro de un exón.

Además se sabe que más del 85% de las mutaciones conocidas para los desórdenes mendelianos raros ocurren en los exones [64], así que es una fuente rica de variantes, si se está tratando de entender los cambios genéticos que causan enfermedades. Por último, resulta más barato, más rápido y más fácil de analizar sólo el 2% en lugar de todo el genoma.

El exoma es útil para encontrar pequeñas variantes, como snp, o pequeños indels. Además algunos CNVs también pueden ser detectados confiablemente dependiendo de dónde están y de su tamaño. Sin embargo, las tecnologías actuales no son muy buenas para detectar indels más grandes o para las repeticiones de trinucleótidos similares a las que se producen, por ejemplo, en el síndrome del x frágil o la enfermedad de Huntington.

En el estudio de genomas o exomas humanos la identificación y caracterización del conjunto de variantes presentes en una muestra es uno de los procesos clave y el más sensible para la correcta aplicación de la secuenciación masiva en el diagnóstico genético [65]. *Para poder realizar esta tarea se necesitan herramientas bioinformáticas y tecnologías de la información. El correcto y eficiente procesamiento, análisis, almacenamiento, integración e interpretación de los datos genómicos, es el principal desafío y escollo que encuentra hoy la genómica para llegar a la clínica. La herramienta desarrollada en esta tesis, y el protocolo que describiremos a continuación busca dar una solución a este problema.*

Análisis y priorización de variantes

Protocolo de priorización

Luego del procesamiento de los datos de secuenciación se obtiene, como ya se mencionó, un enorme número de variantes que necesitan ser evaluadas.

Analizar cada variante individualmente resultaría extremadamente laborioso, por lo que es necesaria la aplicación de filtros para la priorización de las mismas. Para esto primero se cargan las variantes en la *B-Platform*.

A partir de la información anotada en las variantes, producto del paso de anotación funcional del flujo de procesamiento definido en la sección de métodos de esta tesis, se deben seleccionar aquellas con mayor probabilidad de ser causantes del fenotipo observado, en un proceso denominado priorización de variantes.

A partir de la sinopsis clínica del paciente, y de acuerdo con el profesional médico responsable del caso, se realiza usualmente primero una búsqueda en la bibliografía y base

de datos (por ejemplo OMIM) de aquellos “genes candidatos”, que son aquellos que puedan albergar variantes patogénicas que pudieran explicar la patología clínica observada.

Una vez definida una lista de genes candidatos se puede utilizar la misma para reducir el universo de búsqueda de variantes solo a aquellas que se encuentren en estos genes. Entonces se puede crear en la plataforma este tipo de panel de genes virtual para que pueda ser utilizado como filtro inicial en las búsquedas.

Creación y uso de paneles de genes

La plataforma cuenta con un módulo para la administración de listas de genes candidatos, que luego se pueden utilizar en los filtros para priorizar las variantes. Estas listas de genes funcionan como paneles virtuales. Para realizar este paso, una vez dentro de la plataforma se puede acceder desde el menú principal a la solapa “Paneles” (figura 15), y se podrá ver una lista de paneles de genes previamente creados (ya sea por nosotros o por otros colegas):

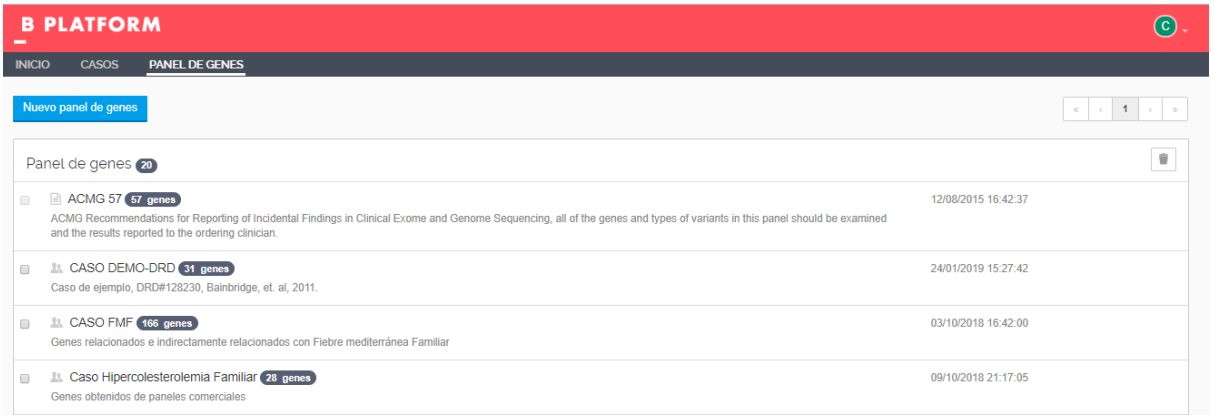

Figura 15. *B-Platform* - Vista de la pestaña “Panel de genes”.

Desde esta pantalla para crear un panel de genes, se ingresa a “Nuevo panel de genes”. *Aquí se desplegará un formulario de alta del panel para completar todos los datos, entre ellos la lista de genes candidatos del caso (ver* Figura 16*)*.

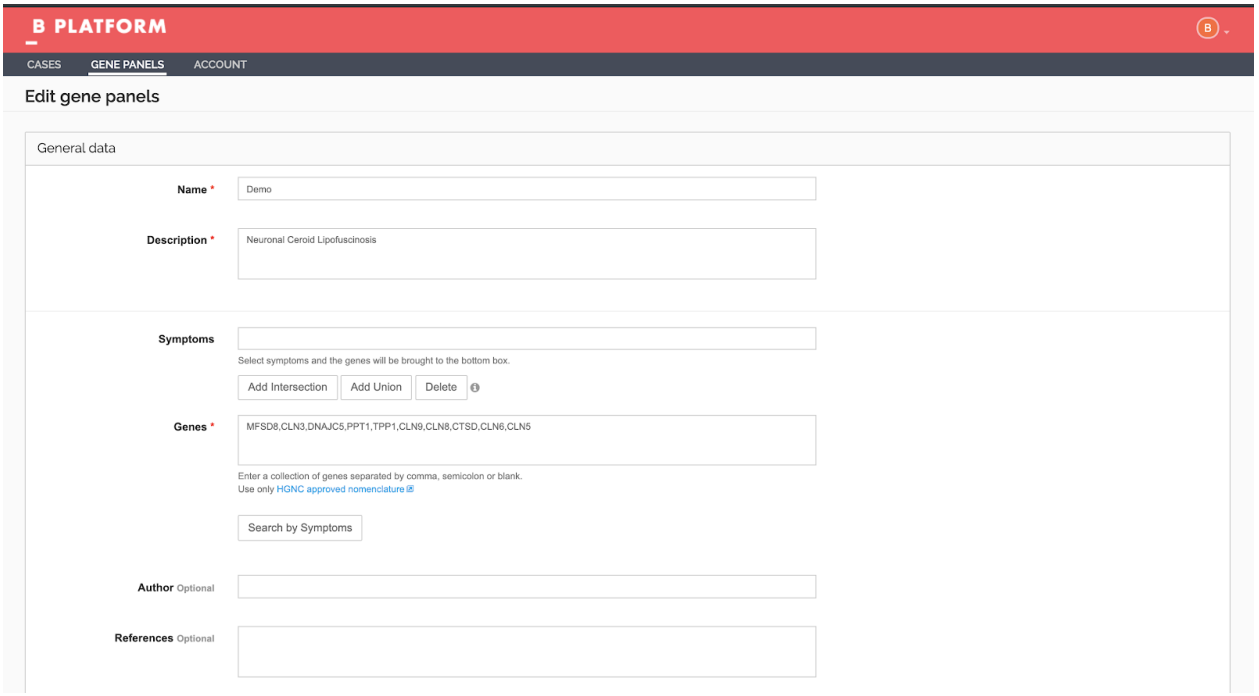

Figura 16. *B-Platform* - Vista de la ventana para la creación de un nuevo panel de genes.

Para facilitar la selección de los genes en la creación del panel, presionando sobre el botón “Buscar por síntoma”, se podrán agregar genes que estén directamente relacionados con los síntomas principales. Para ello, se debe colocar cada síntoma individualmente de forma manual, y seleccionar de la lista desplegable aquella opción que se corresponda, o más se asemeje, al síntoma buscado. La nomenclatura estándar utilizada para la descripción de los síntomas es HPO (Human Phenotype Ontology)[54].

Finalmente, podrá buscar y seleccionar todos los síntomas que desee y luego pedir la intersección “génica” de los mismos (es decir, que la plataforma agregue a la lista únicamente aquellos genes para los que existe evidencia de correlación genotipo-fenotipo para **todos** los síntomas elegidos). Alternativamente, puede armar un panel con la unión de los genes relacionados con cada síntoma por separado (se sumarán a la lista todos los genes asociados a cada entrada). La asociación síntomas-genes (relación fenotipo-genotipo) se hace a partir de los síntomas listados en OMIM para cada gen.

Es importante destacar que en el alta de un panel se valida también que los genes ingresados estén registrados en HUGO Gene Nomenclature Committee (HGNC)[47].
Una vez creado el panel, como ya se mencionó, se lo puede utilizar como filtro para cualquier muestra que se desee analizar en el futuro.

La restricción de la búsqueda al panel de genes candidatos, generalmente si bien reducirá de manera significativa el número de variantes a analizar con mayor detalle, aún resulta en un elevado número de variantes que no son relevantes para el caso, variantes benignas, sinónimas etc, lo importante, cómo veremos a continuación, es que las variantes en estos genes deben ser variantes patogénicas o probablemente patogénicas que puedan afectar al o los genes candidatos que pueden dar lugar a la enfermedad del paciente. Saber que la variante está presente mejorará el cuidado del paciente.

Entonces, para progresar en el análisis del caso, debemos continuar aplicando filtros para reducir al mínimo la cantidad de variantes a analizar, para finalmente llegar a la variante o variantes candidatas que puedan explicar el fenotipo del paciente.

Aplicación de filtros: Simples y por Modelo de enfermedad

Filtros simples

Para comenzar con la priorización de variantes, se debe acceder al panel de filtros de búsqueda como se muestra en la figura 17. Esto abrirá un menú sobre el cual se podrá ir configurando diferentes filtros de acuerdo a diferentes categorías en las que agrupamos sus propiedades (variantes, genes, predicción del impacto y grado de evidencia). Es importante destacar que los filtros son iterativos, lo que permite que los mismos se puedan ir anidando para refinar las búsquedas (o sea aplicar un filtro a las variantes ya filtradas por un filtro anterior).

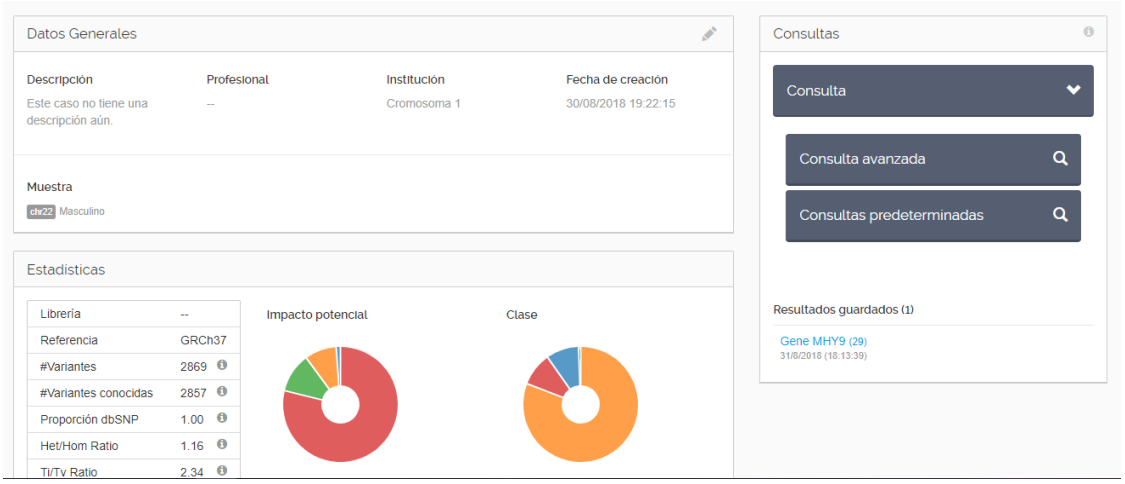

Figura 17. *B-Platform:* Estadísticas del caso. En los gráficos de tortas señalan la cantidad de variantes clasificadas por impacto funcional y clase funcional. Además sobre el margen derecho se pueden ver los botones de acceso para ingresar a la sección de filtros.

Como se mencionó anteriormente, un primer paso importante consiste en buscar variantes conocidas en los genes candidatos que sean patogénicas, o probablemente patogénicas, y que por ende se tenga evidencia de que afectan a la función del gen. Entonces primeramente se suelen buscar variantes con reportes previos de patogenicidad en bases de datos de asociaciones clínicas (por ejemplo, ClinVar). Usualmente, encontrar una variante de este tipo en alguno de los genes candidatos, posee altas chances de ser la variante diagnóstica.

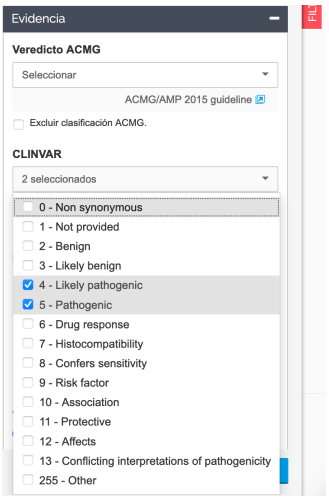

Figura 18. *B-Platform* - Filtro por clasificación según ClinVar

Otro filtro importante a aplicar, es el de frecuencia poblacional de las variantes. Si una variante es demasiado frecuente en la población es clasificada como benigna (ver por ejemplo los criterios de la ACMG). Recordemos, que el término polimorfismo técnicamente corresponde tradicionalmente a una mutación que ha alcanzado una frecuencia alélica de más del 1% en la población. Para analizar la frecuencia, se consultan bases de datos poblacionales como Gnomad y 1000 Genomas.

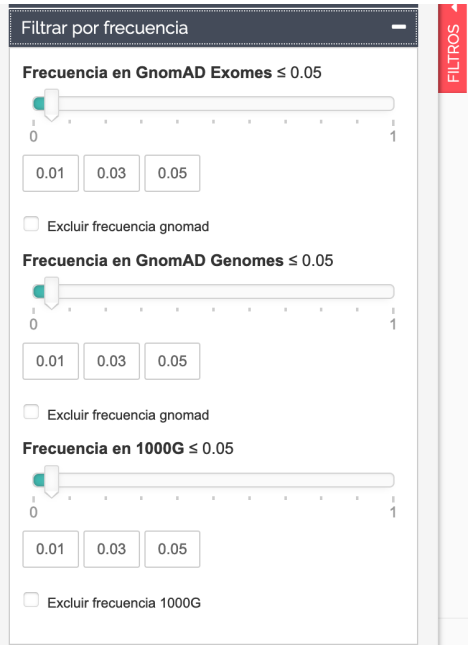

Figura 19. *B-Platform* - Filtro por frecuencia poblacional

Otro filtro muy relevante, particularmente en el contexto de la práctica clínica, es el de ACMG - basado en la asignación de códigos en función de la evidencia disponible de cada variante, siguiendo la guía establecida por el American College of Medical Genetics. Este filtro permite seleccionar variantes de acuerdo a las categorías de la ACMG (patogénicas, probablemente patogénicas, significado incierto, probablemente benignas, benignas), y también por cada una de los niveles de evidencia (PVS, PM, etc).

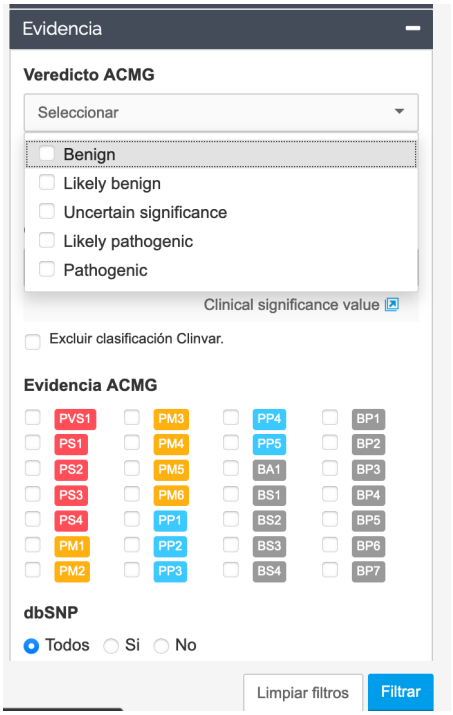

Figura 20. *B-Platform* - Filtro por criterios de ACMG

Además de los filtros mencionados, que son los principales en el protocolo de priorización definido, hay disponibles muchos filtros más que se complementan a los ya mencionados anteriormente, pero que para no extendernos demasiado en esta sección, no se describen en detalle y solo se mencionan brevemente.

Entre los filtros disponibles se pueden mencionar los que están agrupados en la sección “Consecuencias”. Aquí se puede elegir el efecto molecular de la variante: como cambio de marco de lectura, terminación prematura, involucrada en splicing, distintos tipos de inserción / deleción; clase funcional: silenciosa, sinónima/no sinónima; y grado de impacto potencial, alto, moderado, etc.

También se podrá filtrar por algunos predictores bioinformáticos de patogenicidad, en particular M-CAP, DAMN, Mutation taster, polyphen-2 y Sift, seleccionando el score a partir del cual los mismos consideran a una variante como (probablemente) patogénicas. Las predicciones automáticas de este tipo son esenciales para interpretar grandes conjuntos de datos que incluyan variantes genéticas nuevas o poco frecuentes, ayudando a dirigir el análisis en profundidad de las variantes candidatas más prometedoras.

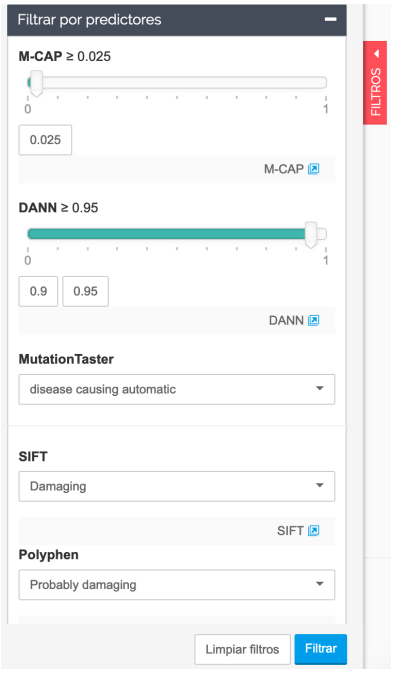

Figura 21. *B-Platform* - Filtro por predictores de patogenicidad

Filtros por modelo de enfermedad

Cuando para el análisis de un caso se tiene la posibilidad de secuenciar un trío, probando, padre y madre (u otro tipo de grupo familiar), un filtro de mucha utilidad para reducir la cantidad de variantes es aplicar un enfoque basado en el modelo de herencia de la enfermedad de la que se sospecha y/o es la determinada para el gen que posee las variantes de interés.

Este enfoque consiste en suponer por ejemplo un modelo de herencia autosómica recesiva, dominante o ligada al cromosoma X, y poder luego aplicar estos modelos como filtro utilizando la información de la cigosidad de las variantes, y su presencia ausencia en cada uno de los miembros de la familia. Por ejemplo, para una enfermedad que sigue un modelo de herencia dominante podemos, analizar si una variante en el paciente no está presente en ninguno de los padres si son sanos (lo que correspondería a una mutación de novo), o en heterocigosis en el padre afectado (patrón dominante típico), o en el caso de una patología recesiva, podemos verificar si ambos padres son portadores sanos de una variante en heterocigosis, que se la han heredado por igual a su hijo.

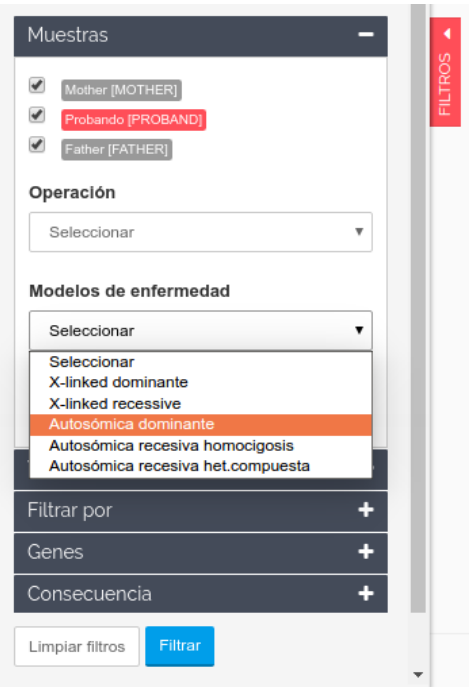

Figura 22. *B-Platform* - Filtro por predictores de patogenicidad

Entonces, como se muestra en la figura 22, también se puede filtrar directamente aplicando un modelo de herencia de la enfermedad en estudio, pudiendo elegir entre dominante o recesivo ligado al X.

Análisis de resultados

El objetivo de la plataforma, desde el punto de vista del diagnóstico, es reducir el número de variantes a analizar de manera manual y detallada en el caso en estudio y luego facilitar el análisis de las variantes candidatas. La meta es llegar a un conjunto reducido de variantes que puedan potencialmente explicar el fenotipo del paciente y que estas puedan ser analizadas en detalle.

Una vez realizado el proceso de filtrado con los parámetros deseados, la plataforma muestra un listado con los genes/variantes que cumplen con los criterios establecidos.

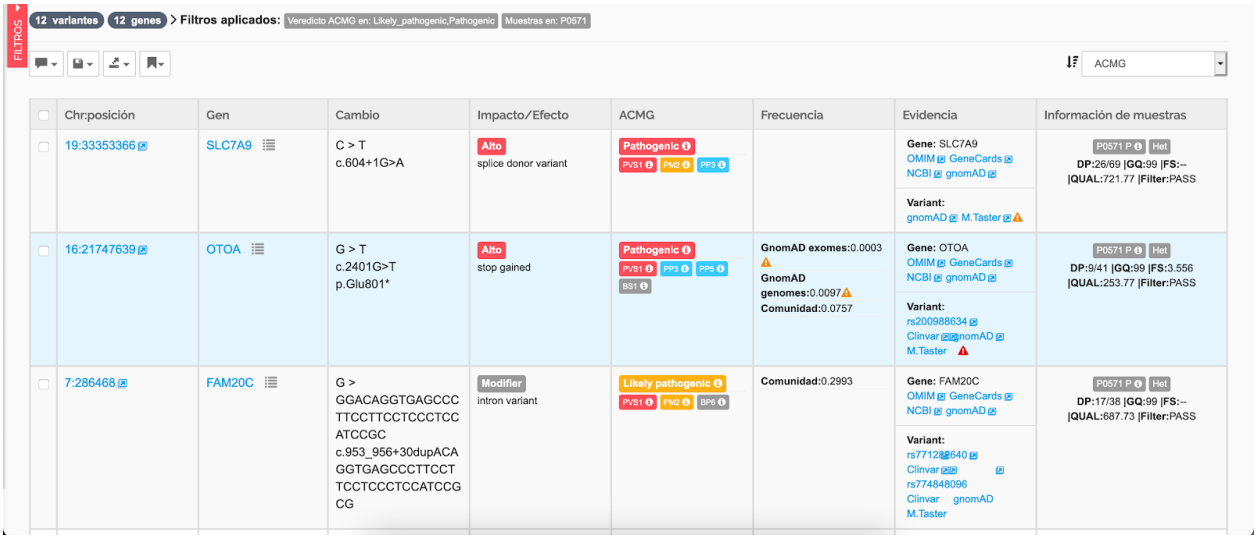

Figura 23. *B-Platform* - Tabla de resultados: lista de variantes

A simple vista se puede verificar en cada fila (que corresponde a una variante) el resultado de la información obtenida, como cromosoma y posición genómica de la variante hallada y acceso al transcripto canónico en la base de datos ENSEMBL. También se encontrará información acerca del fenotipo asociado al gen en OMIM (con su acceso directo), y el modelo de herencia correspondiente.

Para identificar mejor a la variante, se presenta el cambio producido en tres niveles: genómico, cDNA y de la proteína. Cada variante está categorizada según el impacto/efecto a nivel molecular y hace referencia a la magnitud de impacto de la variante en la proteína resultante. Se le asigna un impacto alto a aquellas variantes que resultan en ganancia o pérdida de codones de inicio o finalización de la traducción, y/o cambios en el marco de lectura. El impacto se considera moderado cuando las modificaciones a nivel proteico involucraron cambios no sinónimos de un único aminoácido y/o pequeñas inserciones o deleciones que mantuvieron el marco de lectura. Finalmente, el impacto bajo suele hacer referencia a variantes sinónimas y a aquellas variantes modifier, que son las contenidas en la región intrónica o 3´ y 5´ UTR.

Como ya mencionamos, un campo sumamente importante a la hora de la priorización de variantes, es la clasificación ACMG. Desde los resultados se podrá acceder a dicha clasificación, así como a cada uno de los niveles de evidencia cumplidos por la variantes para que sea ubicada en dicha categoría.

Se encontrarán además los datos de frecuencia poblacional (si la hubiera) de cada variante en poblaciones de referencia obtenidas de gnomAD, 1000 Genomas y frecuencias internas.

Dentro de la evidencia disponible, también se dispone de información relacionada con el código único rs asignado a aquellas variantes que poseen una entrada en la base de datos dbSNP, y otros links de interés a bases de datos externas, tales como UniProt, MutationTaster, ClinVar, gnomAD, OMIM y GeneCards.

Finalmente, y no de menor importancia, se tiene la información de la muestra, es decir, de la secuenciación. Se indica el nombre de la muestra en la que se encuentra presente la variante, la cigosidad de la misma y diversos parámetros de calidad (profundidad, GQ, FS, QUAL, Filter).

Cada variante candidata listada en la plataforma se puede, luego, analizar con mayor detalle. Para esto se investiga cada una individualmente a través de la información brindada por la plataforma (como se muestra en la Figura 24).

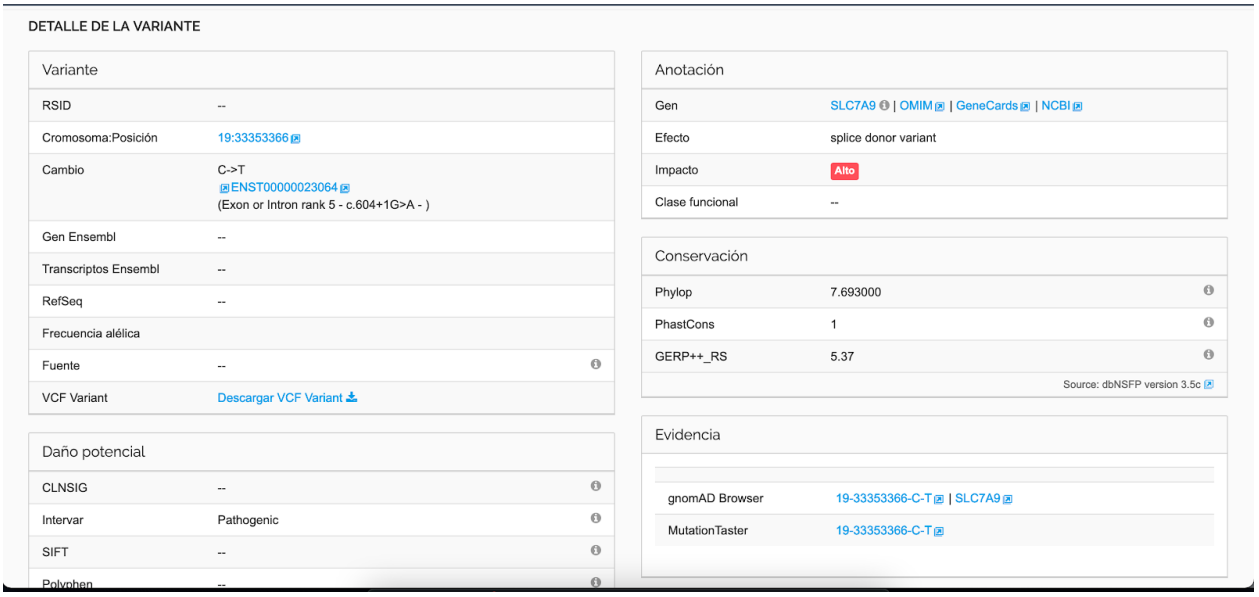

Figura 24. *B-Platform* - Detalle de una variante particular.

Al acceder al detalle de la variante se podrá observar información específica como ser: el gen, efecto, impacto, clase funcional, el identificador rs, frecuencia poblacional, la posición genómica, información sobre el/los transcripto/s y el exón en el que se encuentra. Finalmente, se mostrará información del "Daño potencial", donde figuran scores de conservación y de patogenicidad asignados a la variante por distintos predictores y, en el caso de existir, evidencia en ClinVar, con un significado clínico (ClinSig) asociado.

Entonces, resumiendo esta sección, uno de los objetivos del desarrollo de la plataforma es contar con una herramienta de análisis de datos genómicos, que permite analizar datos de secuenciación, principalmente NGS; de muestras humanas haciendo fácil la priorización, clasificación y análisis de las miles de variantes obtenidas en la secuenciación de paneles, exomas o genomas. Para que luego de seleccionadas las variantes candidatas, a partir del protocolo presentado anteriormente, se pueda realizar la generación de un reporte con aquellas variantes genómicas, posiblemente asociadas con la patología que presente el paciente. Para concluir, este sistema organiza y vincula información biológica derivada de diversas base de datos públicas, y por otro, mediante el análisis de variantes genómicas y su caracterización, permite determinar y predecir el potencial impacto fenotípico de cada una de las variantes encontradas.

B-Platform como herramienta para perfil de riesgo y farmacogenómica

Introducción

En la sección anterior se ha presentado la plataforma como herramienta para el diagnóstico para enfermedades mendelianas. A continuación presentaremos el desarrollo, los principios y aplicaciones de la plataforma para evaluar perfiles de riesgo de enfermedades complejas (poligénicas) y relaciones farmacogenómicas.

Como se mencionó anteriormente, cuando se explico el cálculo y la evaluación de un perfil de riesgo, en el caso de enfermedades poligénicas, usualmente de alta prevalencia en la población adulta, lo que se busca es obtener a partìr de las variantes presentes en el individuo, la predisposición “relativa” a desarrollar diferentes enfermedades que agrupamos en el grupo de complejas (dado su carácter poligénico y su fuerte influencia ambiental).

Hay varios ejemplos de predisposición a enfermedades complejas, las más comunes están relacionadas con el cáncer, como el cáncer de mama y el cáncer colorrectal. Si se tiene una fuerte historia familiar de cáncer de mama, se puede deber a mutaciones en el gen BRCA1 o en el gen BRCA2, genes que tienden a comportarse más como las enfermedades mendelianas, existen una enorme cantidad de variantes en diversos genes que modifican nuestro riesgo de desarrollar cáncer. Otro ejemplo, pueden ser variantes que confieren un mayor (o menor) riesgo de desarrollar enfermedades cardiovasculares, incluyendo el infarto de miocardio o la hipertensión, y/o las variantes genéticas en el locus HLA-DQ asociadas con la enfermedad celíaca.

En los capítulos anteriores también se mencionó la farmacogenómica.Ésta se define como el uso de la información genómica de una persona para mejorar la eficacia o reducir los efectos secundarios de los fármacos en su organismo. Entonces, de manera similar a lo descrito en el módulo para la clasificación y priorización de variantes utilizado para ayudar a llegar a un diagnóstico, como parte de este trabajo de tesis se desarrolló en la plataforma un módulo para calcular el perfil de riesgo de diversas enfermedades complejas predefinidas, y las asociaciones farmacogenómicas de una muestra.

El resultado de este módulo, es que una vez secuenciada la muestra y siguiendo el flujo de procesamiento de los datos NGS propuesto en esta tesis, el archivo de variantes (VCF) se puede subir a la plataforma para el análisis y cálculo del perfil de riesgo para ciertas enfermedades complejas y el perfil farmacogenómico de la misma.

Perfil de Riesgo

Selección de caracteres

Muchos rasgos medibles en nuestro organismo (tanto patológicos como no) no son el simple producto de la acción de un gen, sino que provienen de la interacción de varios loci. Además, el rol del ambiente, no es para nada despreciable en muchos de ellos, por lo que aún teniendo todas las variantes asociadas a un determinado rasgo el poder de predicción sobre el fenotipo es limitado, y de naturaleza probabilística. Un enfoque ampliamente aceptado para la detección de variantes con estas características son los GWAS, como ya explicamos en secciones anteriores.

En este contexto, y como primer paso para agregar el módulo de cálculo de perfil de riesgo a la plataforma se seleccionaron un conjunto de enfermedades complejas (rasgos) de interés a partir de GWAS Catalog[60].

Para cada caracter (enfermedad) seleccionado de GWAS Catalog, se tuvieron en cuenta los siguientes campos:

- **DATE: fecha de publicación del estudio GWAS**
- **STRONGEST_SNP-RISK_ALLELE: el alelo de riesgo reportado**
- **SNPS: identificador de tipo rs de la variante en cuestión.**
- **RISK_ALLELE_FREQUENCY: frecuencia del alelo de riesgo en la población analizada .**
- **P-VALUE: es el valor de significancia de la asociación en cuestión.**
- **P-VALUE (TEXT): indica si el p-value reportado está asociado a algún subgrupo particular - población, género, fumadores, etc.**
- **OR or BETA: el Odds Ratio o BETA**
- **95% CI (TEXT): Es un intervalo de confianza asociado al p-value. Indica además si el valor en el campo anterior es un OR o un BETA.**
- **MAPPED_TRAIT: nombre del carácter (enfermedad) asociado con el vocabulario estándar de EFO [66](experimental factor ontology). Permite recolectar estudios asociados a los rasgos de interés en forma precisa.**

Luego para cada enfermedad (o característica) se tomaron aquellos SNPs asociados al mismo que tuvieran mayor relevancia y confianza. Para ello sobre los campos tenidos en cuenta, se aplicaron los siguientes filtros (por las razones descritas a continuación de cada uno):

- p-valor asociado (campo P-VALUE menor a 10.e-9). Se utilizó este valor de corte para determinar que la variante posee una asociación estadísticamente significativa.
- La variante debe tener un rsID asociado (el campo SNPS no debe ser nulo). O sea, debe haber una variante “conocida” asociada a la patología (trait).
- El alelo de riesgo está correctamente definido (STRONGEST_SNP-RISK_ALLELE debe tener una letra en [A|C|G|T]). Se debe poder identificar de manera correcta el alelo de riesgo, o sea la variante que está asociada al riesgo.

- El OR (o beta en caso de variables cuantitativas) debe estar correctamente especificado (el campo OR or BETA no debe ser nulo).
- La frecuencia del alelo de riesgo en la población de referencia es conocida (el campo RISK_ALLELE_FREQUENCY no debe ser nulo).
- Las variantes han sido reportadas en estudios realizados en poblaciones caucásicas (el campo P-VALUE (TEXT) es nulo, o contiene la secuencia “EA”). Esto se realiza porque a veces los estudios son realizados sobre poblaciones específicas poco relevantes en nuestro país.
- De haber varios trabajos asociados con la misma posición y al mismo trait, nos quedamos con el más reciente (campo DATE más cercano al presente).

Para vincular de manera ordenada las variantes con condiciones de salud, las mismas fueron agrupadas por el campo MAPPED_TRAIT, que refiere a un nombre con vocabulario controlado según la Experimental Factor Ontology - EFO[66]
Entonces siguiendo los criterios anteriores, se seleccionaron un conjunto de patologías, y para cada una de ellas una lista de variantes, de acuerdo a lo que se muestra en la siguiente tabla:

Tabla 6: Lista de patologías seleccionadas.

MAPPED_TRAIT	**SNPS**
Type II diabetes mellitus	rs1496653, rs831571, rs7756992, rs4458523, rs10146997, rs6857, rs10401969, rs10203174, rs312457, rs12010175, rs576674, rs2244020, rs10886471, rs849134, rs849135, rs7177055, rs6467136, rs343092,rs1061810, rs791595, rs2283228, rs12779790, rs9470794, rs6795735, rs9936385, rs2867125, rs3802177, rs2237895, rs4712523
Crohn disease	rs10865331, rs11229030, rs11230563, rs10947261, rs7657746, rs7702331, rs7015630, rs76418789, rs6740462, rs2641348, rs670523, rs2274910, rs2945412, rs2542151, rs1736135, rs11894081, rs13003464, rs17582416, rs17391694, rs7517847,rs9286879, rs6856616, rs3897478, rs4263839, rs6062504, rs11175593, rs10486483, rs1551398, rs10758669, rs11010067, rs11190140, rs35320439, rs4256159, rs3197999, rs7076156, rs174537, rs9292777, rs864745, rs6545946, rs2382817, rs10210302, rs17695092, rs12663356, rs9258260, rs1893217, rs6478109, rs1456893, rs12722515, rs1250546, rs11584383, rs9273363, rs6908425, rs744166, rs3091315, rs2872507, rs4246905, rs1517352, rs12994997, rs17234657, rs11747270, rs9264942, rs3828309
Prostate carcinoma	rs1933488, rs16902094, rs10934853, rs7758229, rs6983561, rs76934034, rs5945619, rs4646284, rs2659124, rs3850699, rs4713266, rs10774740, rs17023900, rs7725218, rs6983267, rs5759167, rs3771570, rs7241993, rs2273669, rs11650494, rs1270884, rs11135910, rs10009409, rs2405942, rs721048, rs7501939, rs684232, rs2735839, rs10009409, rs7584330, rs7929962, rs4245739, rs11263763, rs103294, rs1016343, rs1894292, rs3129859, rs817826, rs7153648, rs8102476, rs17765344, rs188140481, rs10505477, rs4844289, rs8008270, rs2242652, rs8064454, rs4242382, rs1859962, rs11228565, rs7141529, rs11568818, rs12155172, rs10896449, rs71277158, rs9364554,r s5945572

Myocardial infarction	rs3782886, rs10455872, rs10757278, rs2891168, rs646776, rs9349379, rs28451064, rs7528419, rs653178
psoriasis	rs17728338, rs7709212, rs8016947, rs4795067, rs2066807, rs10865331, rs7709212, rs2066807, rs2082412, rs9988642, rs4085613, rs643177, rs2675662, rs20541, rs6677595, rs643177, rs280519, rs12564022, rs34517439, rs4845459, rs458017, rs492602,rs10789285, rs7637230, rs9533962, rs4845459, rs33980500, rs4406273, rs17728338, rs12188300, rs842625, rs34536443, rs2395029
Age-related macular degeneration	Rs1061170, rs920915, rs5749482, rs13278062, rs10801555, rs2230199, rs12153855, rs541862, rs13081855, rs8135665, rs10033900, rs943080, rs6795735, rs4698775, rs8017304, rs3764261, rs1864163, rs334353, rs2071277, rs3130783, rs429608, rs1329424, rs4420638, rs10490924, rs1061147, rs3750848
Parkinson disease	rs393152, rs199347, rs6430538, rs11158026, rs76904798, rs34372695, rs1555399, rs4784227, rs12185268, rs11060180, rs9275326, rs34637584, rs329648, rs2414739, rs10797576, rs34311866, rs10513789, rs6812193, rs12456492, rs12637471, rs823118, rs356182, rs14235, rs11711441, rs11248051, rs356220
Alzheimer's disease	rs11754661, rs75932628, rs9331896, rs6733839, rs679515, rs28834970, rs2373115, rs11771145, rs1476679, rs983392, rs9271192, rs10948363, rs2718058, rs6656401, rs17125944, rs2732703
melanoma	rs1636744, rs4785763, rs1805007, rs13016963, rs258322, rs10739221, rs35407, rs910873, rs498136, rs1393350, rs7412746, rs16953002, rs6059655
glaucoma	rs523096, rs4656461, rs3213787, rs4977756
hypertension	rs13333226, rs11646213
stroke	rs2200733, rs12425791
Ovarian carcinoma	rs7953249, rs11651755, rs7705526, rs9886651, rs7405776

Determinación del riesgo a partir de los datos de riesgo alélico (Odds Ratio)

Los OR reportados en GWAS Catalog [60] para cada patología y cada variante seleccionada se corresponden con la presencia de un alelo para dicha variante. Teniendo en cuenta que somos diploides, y asumiendo independencia entre el riesgo conferido por la presencia de cada alelo, debemos convertir los OR por alelo (derivados de GWAS) a OR genotípicos, que son dados por las siguientes ecuaciones (***donde R refiere al alelo de riesgo y Q refiere al alelo sano***):

- OR(RR) = OR x OR = OR^2
- OR(QR) = OR x 1 = OR
- OR(QQ) = 1 x 1 = 1

Es importante destacar que estos OR crudos se refieren a la magnitud del efecto del genotipo de riesgo relativo ***al compararse contra el genotipo sano (o de referencia) utilizado en el estudio de GWAS***, lo cual no es directamente aplicable a un análisis de perfil de riesgo de un individuo. Para poder aplicarlo al riesgo de un individuo debemos ajustar los valores de los OR para que reflejen la magnitud del efecto ***al compararse contra la prevalencia poblacional,*** que se puede definir como:

$$\textbf{Prevalencia} = OR(QQ)P(QQ) + OR(QR)P(QR) + OR(RR)P(RR)$$

Donde P(QQ), P(QR) y P(RR) refieren a las frecuencias poblacionales **de los genotipos QQ, QR y RR**.

Asumiendo que el polimorfismo en cuestión respeta el ***equilibrio de Hardy-Weinberg***, la prevalencia puede calcularse a partir del OR y la frecuencia poblacional del alelo de riesgo (ambos presentes en GWAS catalog en los campos OR y RISK_ALLELE_FREQUENCY) con la siguiente ecuación:

$$\textbf{Prevalencia} = OR(QQ)x(1-P(R))^2 + OR(QR)x2xP(R)x(1-P(R)) + OR(RR)xP(R)^2$$

Entonces, a partir de los OR por genotipo y la prevalencia, podemos determinar los OR ajustados (marcados **OR***) mediante el simple cociente de los OR derivados de GWAS y la prevalencia calculada previamente.

- OR*(RR) = OR^2 / prevalencia
- OR*(QR) = OR / prevalencia
- OR*(QQ) = 1 / prevalencia

Los OR* corresponden entonces a factores de riesgo relativos, por genotipo, y ajustados por la prevalencia de la enfermedad en la población, por lo que permiten evaluar el riesgo relativo de desarrollar la enfermedad para un individuo que los porte, en relación a la población de referencia.

Recordemos que para una patología (trait) dada puede haber varios alelos que se encuentran asociados, y para cada uno de ellos se determinará su OR*. Para entonces calcular el riesgo del individuo combinando todas las variantes presentes se asume primero que la contribución de cada variante al riesgo es independiente de las demás. Entonces, el factor de riesgo total es igual al producto de los OR ajustados de todas las variantes relacionadas con una enfermedad particular. De este modo, para un *trait* relacionado con **k** variantes, el OR* total (combinando la contribución de todas las variantes encontradas) pueden definirse como:

$$\prod_{i=0}^{k} \textbf{OR*}_{\textbf{genotípico i}} = \textbf{odds(}\textit{trait}\textbf{ | genotipo)}$$

Este valor, entonces, corresponde al riesgo relativo que aporte el genoma del individuo al desarrollo de la enfermedad en cuestión.

Análisis de estadística poblacional por 1000G

Para tener un primer análisis de las distribuciones de riesgo poblacional que se obtienen utilizando el modelo propuesto, y a partir de las mismas definir valores umbrales para estratificar a la población, utilizando los datos de genotipo del proyecto 1000 Genomes [37], se calcularon las distribuciones de los OR* y se tomaron los percentiles 80 y 95. Las distribuciones fueron calculadas determinando el OR* de cada individuo y luego haciendo un histograma para toda la población (ver Figura 25).

Entonces para infarto de miocardio, por ejemplo, se obtuvieron los siguientes valores:

Trait	Min	Max	Mean	20th_perc	5th_perc
myocardial_infarction	0.63082	3.74034	1.69476	1.96872	2.52189

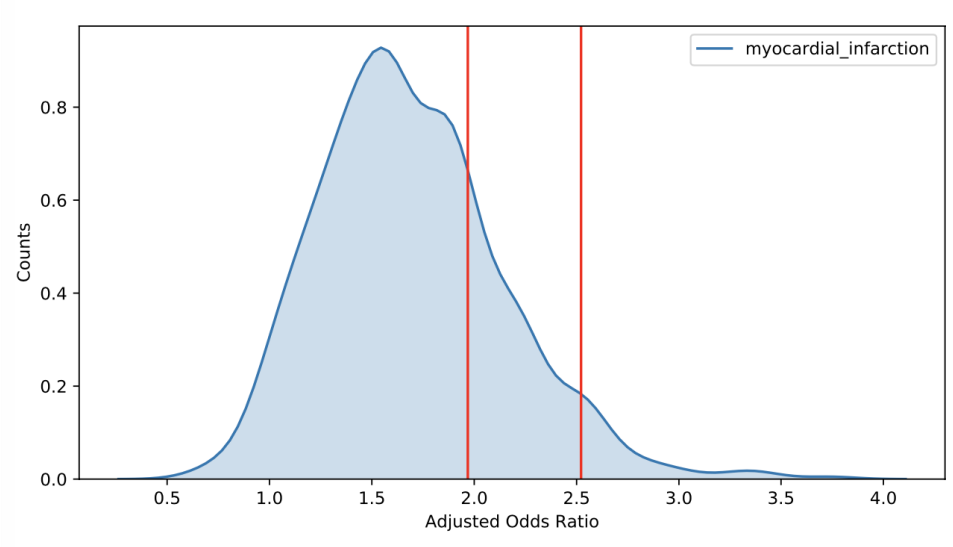

Figura 25: Distribución de los OR* para el trait infarto de miocardio. Las líneas verticales en rojo indican los percentiles 20th y 5th.

Un último ejemplo puede ser la enfermedad de Parkinson

Trait	Min	Max	Mean	20th_perc	5th_perc
Parkinson_disease	0.18215	2.86504	0.77159	1.04416	1.48478

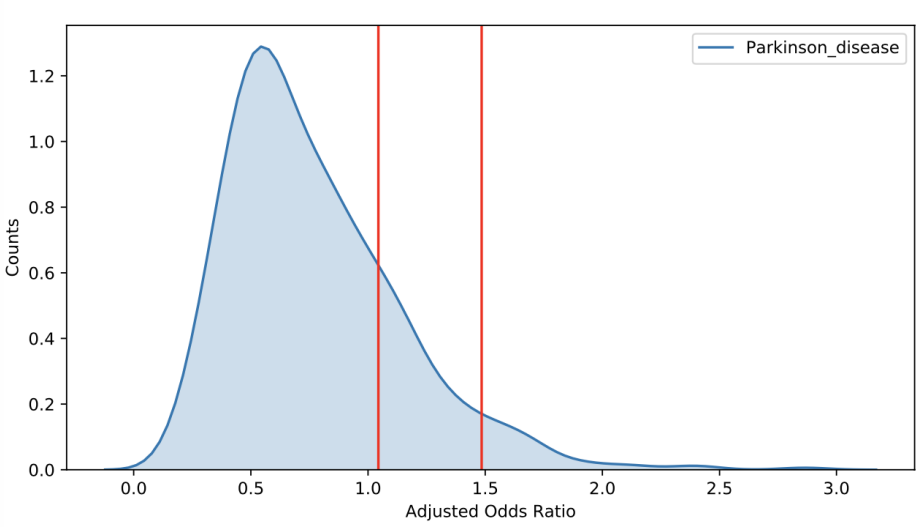

Figura 26: Distribución de los OR* para la enfermedad de Parkinson. Las líneas verticales en rojo indican los percentiles 20th y 5th.

A partir de los histogramas, para cada rasgo se definió una escala tripartita (barras rojas en los gráficos) y se clasificó a cada individuo de acuerdo a las siguientes tres categorías:

- El riesgo **NORMAL** fue asignado a los valores por debajo del percentil 80 en la distribución poblacional obtenida.
- El riesgo **MODERADO** fue asignado a los valores entre los percentiles 80 y 95 en la distribución obtenida.
- El riesgo **ELEVADO** fue asignado a los valores por encima del percentil 95 en la distribución obtenida.

A partir de los gráficos y las categorías definidas, podemos entonces decir que un individuo cuyo análisis genómico lo ubique en riesgo moderado para infarto de miocardio, posee hasta dos veces más chances de desarrollar un infarto que un individuo con riesgo normal, mientras que el individuo con riesgo elevado puede poseer un riesgo hasta 4 veces mayor

De igual manera se calcularon las distribuciones y los valores de corte para cada una de las enfermedades de interés. La tabla resultante con los valores de corte, más el algoritmo para el cálculo de riesgo fueron incluidos en la plataforma para que dada una muestra cargada en la plataforma, poder calcular el perfil de riesgo de ese individuo.

Visualización del Perfil del riesgo en la Plataforma.

Luego de que una muestra es procesada con el protocolo definido anteriormente, se obtiene un archivo de variantes anotadas que contiene la información de GWAS Catalog [60] para cada una de las variantes de riesgo seleccionadas, que luego se carga en la plataforma para el análisis de riesgo.

Después de cargar el archivo de variantes en la plataforma, se accede una grilla donde están agrupadas lógicamente las enfermedades en la sección análisis de riesgo, como se puede apreciar en la figura 27. La grilla permite acceder a las variantes por enfermedad (o fármaco).

Entonces, por ejemplo, bajo la etiqueta “Enfermedades del sistema nervioso”, podemos agrupar la enfermedad de parkinson y la enfermedad de alzheimer.

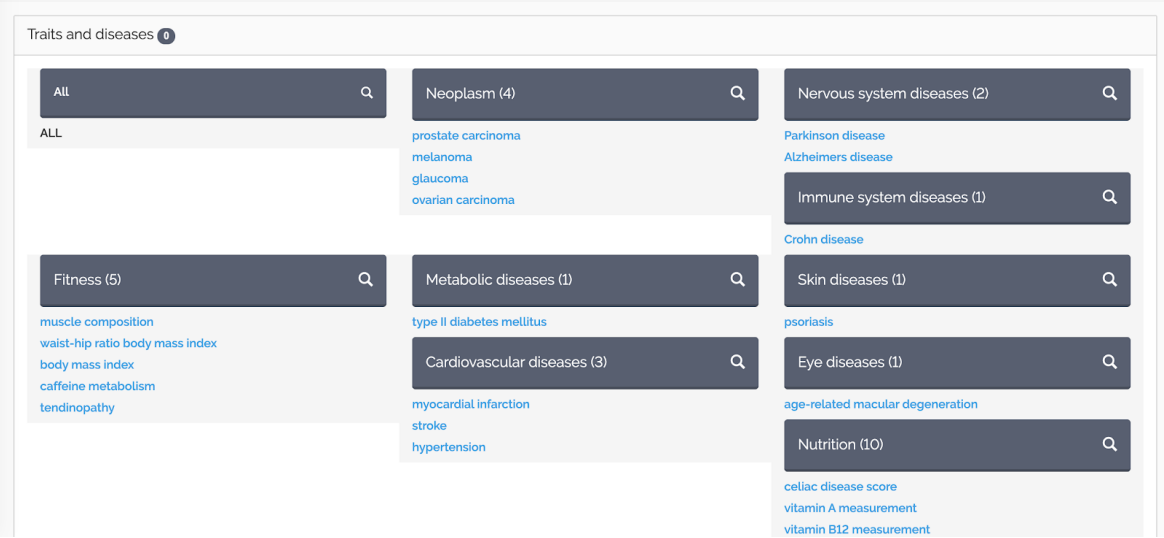

Figura 27. *B-Platform* - Grilla de agrupamiento lógico de traits.

Una vez seleccionado “Alzheimer disease”, se listan todas las variantes elegidas previamente para calcular el riesgo de Alzheimer, como explicamos anteriormente. Para cada variante se indica el genotipo de la muestra, y luego el cálculo de los OR ajustados y el riesgo total. En este caso, como se puede ver en el recuadro rojo de la figura 28, el riesgo para la enfermedad de Alzheimer para esta muestra es ***Aumentado*** (Enhanced).

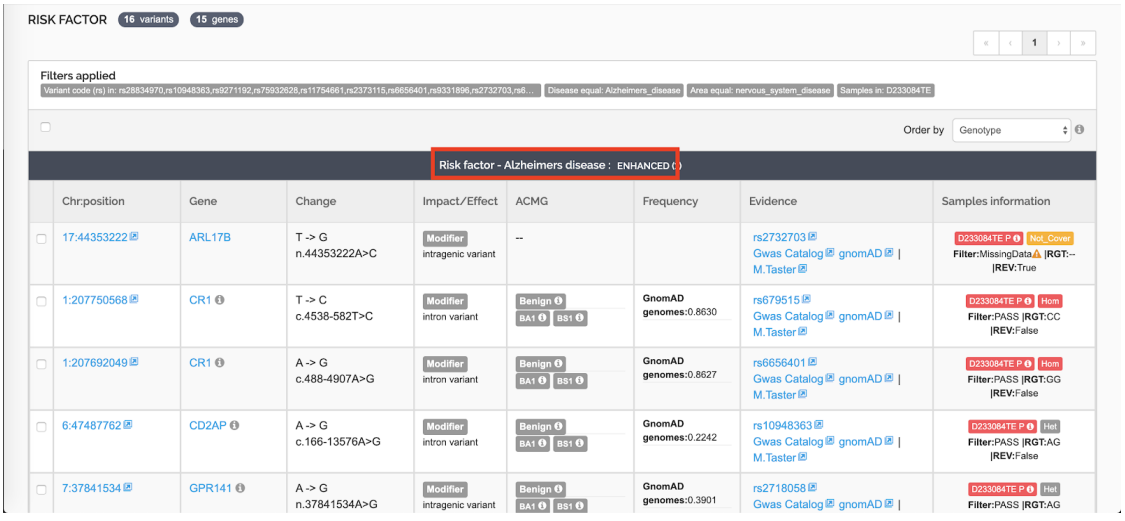

Figura 28. *B-Platform* - Resultado de la evaluación de riesgo elevado para enfermedad de alzheimer.

De igual manera, se puede acceder a las variantes elegidas para calcular el riesgo de desarrollar enfermedad de Parkinson. En este caso el perfil de riesgo para Parkinson es ***Normal***

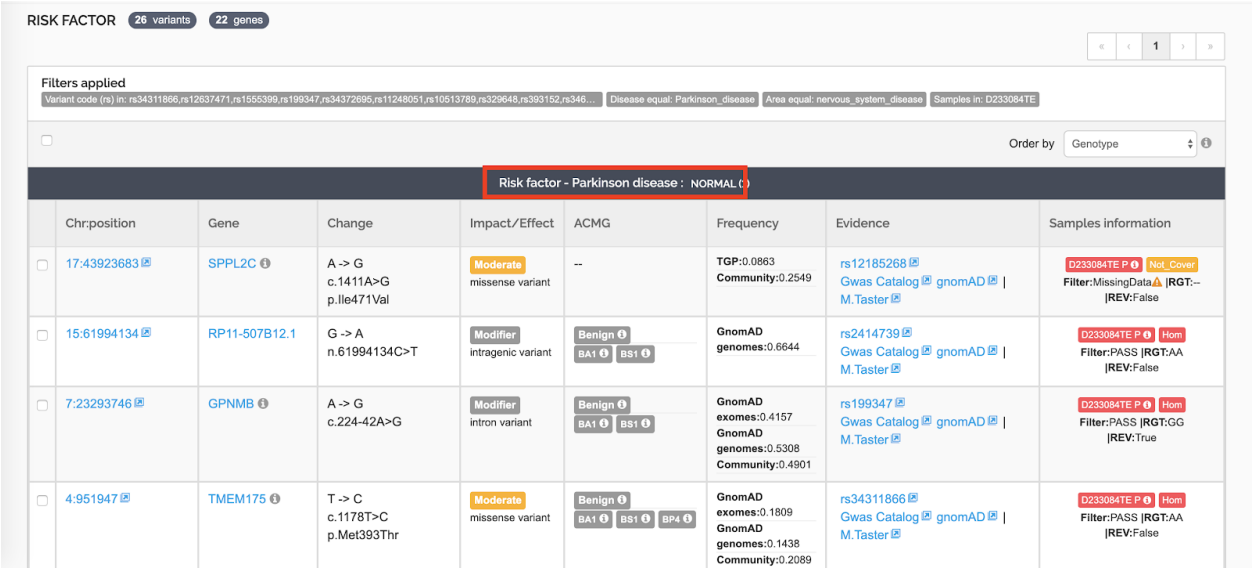

Figura 29. *B-Platform* - Resultado de la evaluación de riesgo normal para Parkinson disease.

También, como ejemplo de riesgo moderado, se puede mostrar el cálculo de riesgo para Psoriasis.

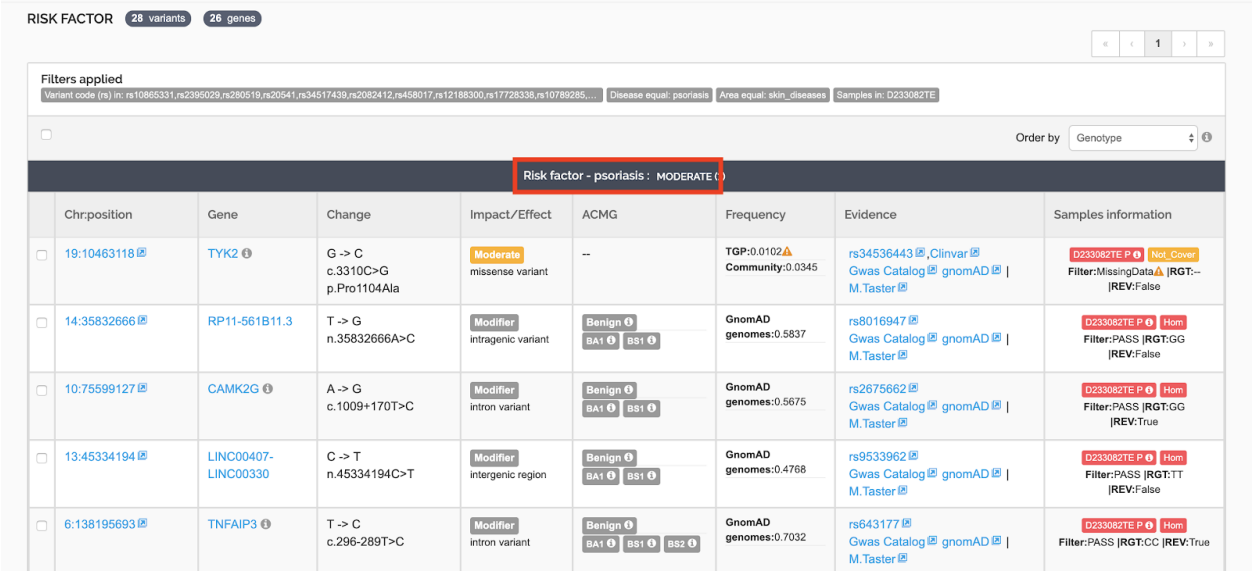

Figura 30. *B-Platform* - Resultado de la evaluación de riesgo moderado para Psoriasis.

Para cerrar esta sección, se muestra un ejemplo más de riesgo moderado asociado a accidente cerebrovascular (Stroke en inglés).

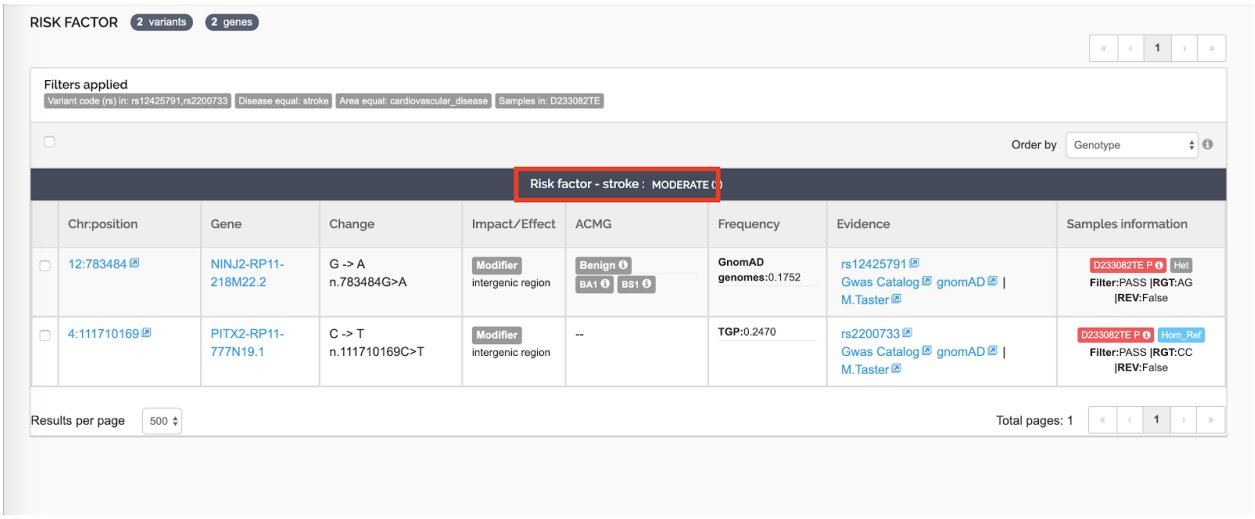

Figura 31. *B-Platform* - Resultado de la evaluación de riesgo moderado para ACV.

En resumen, la plataforma permite calcular y luego visualizar el perfil de riesgo de desarrollar unas 12 enfermedades complejas, entre las que se destacan cardiovasculares, enfermedades neurodegenerativas, metabólicas como la diabetes tipo 2 y cáncer. La categorización del perfil de riesgo del individuo, para cada una de ellas de acuerdo a la distribución poblacional permite tener luego, para cada una de ellas, una valoración relativa de la chance de desarrollarla. De este modo, el individuo puede focalizar los esfuerzos preventivos, ya sea mediante cambio en los hábitos de vida, como en el potencial uso de medicamentos preventivos y/o modificación en la frecuencia y tipo de análisis para la detección temprana, en aquellas patologías donde posea mayor riesgo.

Perfil Farmacogenómico

Como ya se mencionó, la farmacogenómica es la rama de la genómica que estudia la interacción entre los fármacos y nuestro genoma. Más precisamente, cómo las diferentes variantes alélicas de nuestros genes afectan la respuesta a los medicamentos .

Como base para la implementación de este módulo se tomó la información de PharmGKB. De esta base de datos se seleccionaron un conjunto de fármacos de interés, y se agruparon primero por área terapéutica (ver Tabla 7).

Selección de fármacos

Para cada fármaco elegido, se tomó el snp asociado y se extrajo principalmente la información de las denominadas anotaciones clínicas. Estas anotaciones incluyen, entre otras cosas, el nivel de evidencia de la asociación variante-fármaco y el efecto/asociación para cada droga en función del genotipo de la muestra.

Tabla 7: Fármacos seleccionados de PharmGKB

DRUG	Area	SNPS
rosuvastatin	Cardiology	rs2231142,rs4693075
diltiazem	Cardiology	rs12946454
prasugrel	Cardiology	rs12248560, rs4244285
simvastatin	Cardiology	rs4149056
clopidogrel	Cardiology	rs28399504
warfarina	Hematology	rs1057910, rs9923231
ibuprofen	Rheumatology	rs20417
amitriptyline	Psychiatry	rs4244285
escitalopram	Psychiatry	rs12248560
salbutamol	Pneumology	rs1042713
tamoxifen	Oncology	rs4646
ondansetron	Gastroenterology	rs776746

La información asociada a cada uno de estos fármacos comprende lo que sería una “**anotación farmacogenómica para un conjunto de variantes**”. La anotación comprende el resumen de la asociación entre una sola variante genética y una respuesta al fármaco en cuestión.

Por ejemplo, para la rosuvastatina en PharmGKB podemos encontrar para la variante rs2231142 (https://www.pharmgkb.org/variant/PA166156544/clinicalAnnotation/1154221922) la siguiente información:

La variante, que se encuentra en el gen ABCG2, está asociada a la eficacia en el uso del fármaco rosuvastatin en el tratamiento a pacientes con hipercolesterolemia, o infarto de miocardio con un nivel de evidencia 2A (un nivel de alta confianza). Se brinda, además, información esencial de la relación entre el genotipo del paciente para esta variante, y la correspondiente respuesta del fármaco.

Tabla 8: Relación genotipo-fenotipo de la variante rs2231142 para con el fármaco rosuvastatin

Alelos	Fenotipo
GG	Los pacientes con el genotipo GG que son tratados con rosuvastatina: 1) pueden tener concentraciones plasmáticas más bajas de rosuvastatina 2) Pueden tener una respuesta reducida al tratamiento, determinada por una reducción menor de LDL-C en comparación con los pacientes con el genotipo TT.
GT	Los pacientes con el genotipo GT que son tratados con rosuvastatina: 1) pueden tener concentraciones plasmáticas más bajas de rosuvastatina 2) pueden tener una respuesta reducida al tratamiento, determinada por una reducción menor de LDL-C en comparación con los pacientes con el genotipo TT, o a) pueden tener concentraciones plasmáticas más altas de rosuvastatina b) Pueden tener una mejor respuesta al tratamiento, determinada por una reducción mayor de LDL-C en comparación con los pacientes con el genotipo GG.
TT	Los pacientes con el genotipo TT que son tratados con rosuvastatina: 1) pueden tener concentraciones plasmáticas más altas de rosuvastatina 2) pueden tener una mejor respuesta al tratamiento, determinada por una mayor reducción de LDL-C en comparación con los pacientes con el genotipo GG.

Entonces, analizando el genotipo de todas las variantes asociadas a los fármacos de interés seleccionadas se puede ir construyendo el perfil farmacogenético del paciente. Por ejemplo, para este casos, si el paciente tuviera el genotipo GG para la variante rs2231142 se puede decir que este se asocia con una disminución del porcentaje de cambio en los niveles de colesterol LDL cuando se lo trata con rosuvastatina en comparación con individuos con el genotipo TT.

Visualización del Perfil del farmacogenómico en la Plataforma.

De manera similar a como se diseñó el acceso al perfil de riesgo, se pensó el módulo del perfil farmacogenómico. Este presenta una grilla donde aparecen agrupados los distintos fármacos agrupados por áreas terapéuticas.

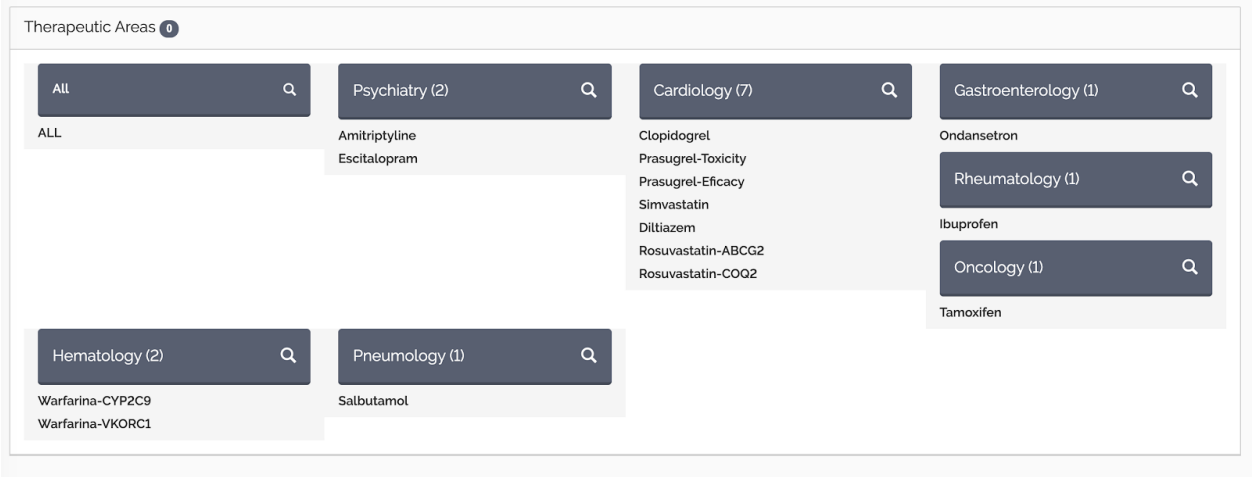

Figura 32. *B-Platform* - Grilla de agrupamiento de fármacos por áreas terapéuticas.

Luego al acceder al área de cardiología, se listan todas las variantes asociadas a fármacos utilizados en esta área.

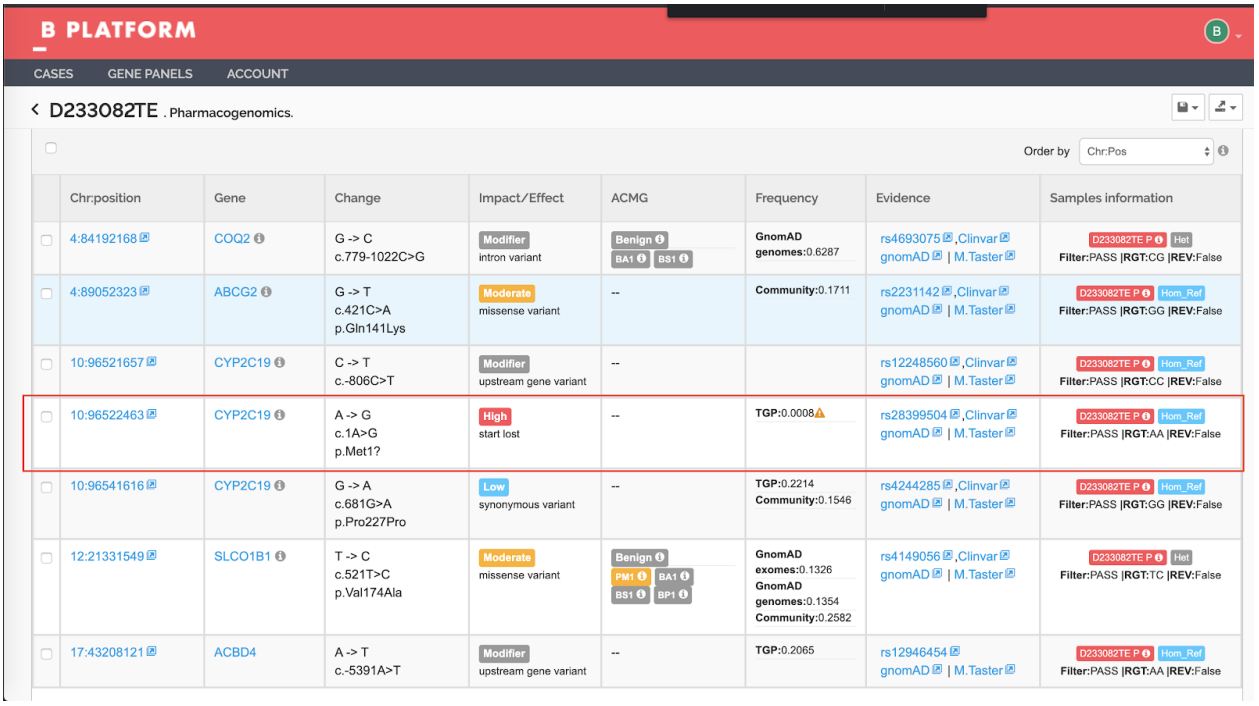

Figura 33. *B-Platform* - Lista de variantes asociadas a fármacos del área de Cardiología.

Si se elige la variante rs28399504 asociada a Clopidogrel, podemos acceder al detalle de la variante seleccionada, y encontraremos información, del gen en donde se encuentra, en este caso en el CYP2C19, el nivel de evidencia de la asociación variante-fármaco, si esta

asociación está relacionada a la eficacia o toxicidad de la droga, y finalmente el fenotipo asociado.

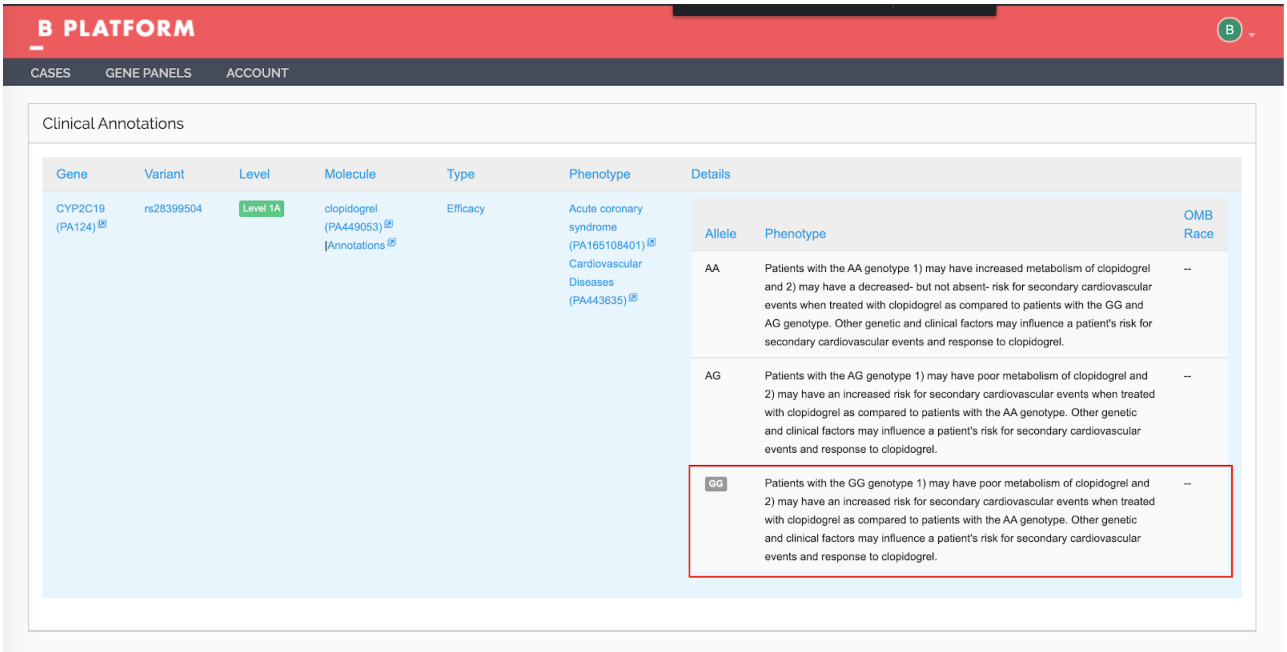

Figura 34. *B-Platform* - En el detalle se ve asociado la explicación al genotipo del paciente.

Como se muestra en la Figura 34, en el detalle se ve asociado al genotipo del paciente las recomendaciones de CPIC (Clinical Pharmacogenomics Implementation Consortium) [63], por ejemplo en este caso el paciente es homocigota GG, la CPIC resalta que: Los pacientes con el genotipo GG:

1) pueden tener un metabolismo deficiente de clopidogrel y

2) pueden tener un mayor riesgo de eventos cardiovasculares secundarios cuando son tratados con clopidogrel en comparación con los pacientes con el genotipo AA.

Casos de Aplicación

Caso I: La campaña 100 exomas

Introducción

Uno de los logros principales de este trabajo de tesis fue la utilización con éxito de la plataforma bioinformática desarrollada en la campaña 100 Exomas, que permitió, tomando como punto de partida los datos obtenidos del secuenciador, llegar a la interpretación biológica y clínica de las variantes encontradas en cada caso. Además, de permitir una validación “a campo” del protocolo diseñado.

En este contexto la plataforma fue utilizada por diferentes profesionales de la Salud (Médicos, Bioquímicos, Genetistas, Biólogos) de más de 30 instituciones de todo el país, para el procesamiento y análisis de los exomas secuenciados para la campaña.

La campaña 100 exomas fue desarrollada bajo el concepto de que la cooperación interinstitucional y la complementariedad disciplinar son esenciales para el éxito de un programa de genómica clínica. En el contexto de Argentina, con 40 millones de habitantes distribuidos heterogéneamente a lo largo y ancho del país con acceso dispar a servicios de salud y con una variedad de especialidades médicas que mostraron interés, se priorizó el contacto directo con los médicos responsables de los casos; limitando el número de casos de cada institución participante. Participaron del proyecto 58 médicos agrupados en 32 instituciones de salud tanto públicas como privadas que incluyen principalmente hospitales y laboratorios de análisis clínicos. El grupo de análisis fue conformado por bioinformáticos y biólogos moleculares con experiencia en análisis de datos genómicos. En total se procesaron y analizaron 129 exomas correspondientes a 100 casos, con diagnósticos presuntivos de enfermedades asociadas a uno, o unos pocos genes -como el síndrome de Sotos, epilepsia mioclónica progresiva - a enfermedades con asociación a decenas (como desórdenes de glicosilación), o incluso centenas de genes - como las Inmunodeficiencias primarias.

El éxito de la campaña, derivó en la entrega del premio CEDIQUIFA y del envío de una publicación a la revista Química Viva [67] mencionados más adelante en la sección correspondiente.

Los resultados de la campaña, como se describen a continuación, junto con esfuerzos previos del mismo grupo de trabajo, han tenido un enorme impacto en la comunidad biomédica, generando conocimiento de gran relevancia que ha trascendido las fronteras nacionales.

Métodos

Protocolo de enrolamiento clínico.

El reclutamiento de casos se llevó a cabo a través de la difusión de la *campaña "100 exomas"* en instituciones de la salud, tanto públicas como privadas, dentro del territorio argentino. Se realizaron entrevistas con el personal médico interesado, durante las cuales se presentaron aquellos casos que se pensara pudieran beneficiarse de una secuenciación exómica. Como criterio de selección se dió prioridad a aquellos casos para los cuales se tuviera un fuerte indicio de enfermedad hereditaria monogénica, con un buen diagnóstico clínico y conocimiento previo de la etiología molecular, es decir, que existieran evidencias previas de mutaciones en genes candidatos que resulten en el fenotipo asociado al diagnóstico. También fue necesario descartar en esta etapa aquellos casos que tuvieran una alta probabilidad de ser causados por modificaciones genéticas no detectables por medio de la secuenciación de exomas, como inserciones y deleciones grandes (mayores a 150 pares de bases), variaciones en el número de copias, rearreglos cromosómicos o mutaciones en regiones no codificantes del genoma.

Una vez preseleccionados los casos, se procedió a realizar para cada uno un reporte de factibilidad. El mismo consistió en analizar: i) la historia clínica y familiar del paciente, ii) las asociaciones genotipo-fenotipo conocidas para la patología, iii) trabajos previos que dieran cuenta del uso de tecnologías NGS para casos similares, y iv) la existencia de paneles de secuenciación comerciales que cubran las enfermedades descritas relacionadas con el diagnóstico presuntivo. A partir de esto se confeccionó una lista de genes candidatos para los cuales se analizó la cobertura horizontal teórica para el kit de captura utilizado y, finalmente, se asignó a cada caso un nivel de factibilidad cualitativo (bajo - medio - alto), junto con una propuesta sobre a quiénes secuenciar del grupo familiar.

De esta manera, cada análisis de factibilidad confeccionado fue entregado a los médicos responsables de los casos clínicos, quienes, en base al nivel de factibilidad asignado y la propuesta de secuenciación, tomaron la decisión final sobre continuar o no con el estudio. En caso afirmativo, en cada caso se procedió a secuenciar el exoma de cada paciente/ grupo familiar.

Secuenciación de las muestras

Una vez firmado el consentimiento por parte de los pacientes (ver sección "Consentimiento informado y hallazgos incidentales"), se procedió a la extracción de sangre periférica y la purificación de ADN de linfocitos circulantes, mediante kits comerciales a una concentración final mínima de 50 ng/ul y una pureza mayor a 1,8 en relación de absorbancia 260nm/280nm. Las muestras fueron analizadas en geles de agarosa al 2% para evaluar la calidad del ADN. La captura exómica se realizó con el kit de "Agilent SureSelect Human All Exon V5" y las muestras fueron secuenciadas mediante la tecnología "Illumina HiSeq 4000" con una longitud de lectura de 100 pares de bases y una profundidad promedio de 100X.

Todas las muestras fueron anonimizadas en todos los pasos del análisis, desde la extracción de sangre hasta la entrega de resultados finales al médico responsable.

Protocolo de procesamiento de datos

Las lecturas obtenidas del proceso de secuenciación fueron procesadas según el *Flujo de procesamiento de los datos* definido en este trabajo de tesis que se describe en detalle en el capítulo **Materiales y Métodos.**

Protocolo de priorización de variantes

A partir de la información anotada en las variantes, se seleccionaron aquellas con mayor probabilidad de ser causantes del cuadro clínico presentado, en un proceso denominado priorización de variantes. ***Para ello se utilizó la herramienta informática B-Platform, desarrollada en el marco de esta tesis*.** Este software funciona dentro de un servidor accesible por internet que facilita las consultas sobre los datos de variantes de cada caso, así como el acceso a los datos por parte del equipo médico. Todos los casos fueron sometidos al protocolo de priorización de variantes diseñado previamente, con la finalidad de descartar variantes que no cumplieran con los requisitos para ser consideradas como "patogénicas" o "posiblemente patogénicas" según los criterios del American College of Medical Genetics and Genomics (ACMG) [6], de enriquecer los resultados con aquellas variantes candidatas a ser las causales de la sintomatología presentada. Dicho protocolo consta de grupos de filtros (descritos en la figura 35), que se aplican secuencialmente y para los cuales se analizan los resultados de manera detallada.

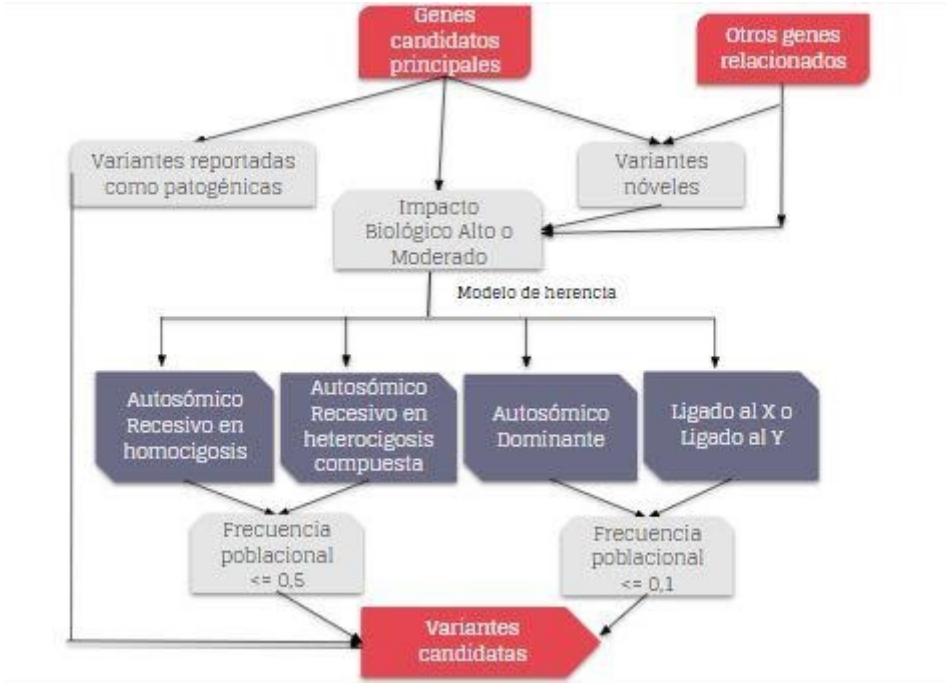

Figura 35: Esquema de priorización de variantes. Según la historia clínica, se seleccionaron genes candidatos a albergar variantes patogénicas que puedan explicar la patología clínica. En primer lugar se buscaron variantes con reportes previos de patogenicidad en bases de datos de asociaciones

clínicas. Luego se procedió a buscar variantes con alto impacto y baja frecuencia poblacional, según el modelo de herencia ó variantes nóveles en los genes candidatos. En los casos donde no se encontraron variantes relevantes en estos primeros pasos, se continuó con la búsqueda de variantes nóveles de impacto alto o moderado por fuera del panel de genes propuesto, pero aún en genes relacionados a la patología.

Cada variante fue categorizada primero de acuerdo a su impacto a nivel molecular. Se le asignó un impacto alto a aquellas variantes que resultan en ganancia o pérdida de codones de inicio, finalización de la traducción y/o cambios en el marco de lectura. El impacto se consideró moderado cuando las modificaciones a nivel proteico involucran cambios no sinónimos de un único aminoácido y/o pequeñas inserciones o deleciones que mantengan el mismo marco de lectura. El resto se consideró de bajo impacto. Luego, para cada variante se analizó su frecuencia poblacional (si la hubiera), la evidencia previa de asociaciones clínicas (calificación de ClinVar considerando valores de ClinSig 4 o 5, correspondientes a las categorías “Likely pathogenic” y “Pathogenic”, respectivamente, ver métodos para màs detalle), el efecto fenotípico previsto, el modelo de herencia para mutaciones ya reportadas en ese gen según OMIM y, si fuera posible, el efecto sobre la estructura y función proteica (utilizando principalmente Uniprot).

Consentimiento informado y hallazgos incidentales. Todos los casos analizados formaron parte de: i) protocolos de investigación aprobados por los comités de ética de las instituciones que atienden a los pacientes, con los consentimientos informados correspondientes; o alternativamente ii) casos enmarcados en un proceso de innovación clínica donde el médico responsable del caso otorga y explica el consentimiento informado al paciente. En todos los casos el análisis se limitó a aquellos genes previamente consensuados con el profesional y directamente relacionados con el diagnóstico presuntivo, para disminuir la posibilidad de hallazgos incidentales. En ninguno de los casos se presentó un hallazgo incidental.

Resultados

Para analizar el grado de éxito alcanzado en la campaña con la secuenciación y análisis de los 100 casos, cada caso fue clasificado en una de las siguientes categorías de acuerdo al tipo y nivel de evidencia disponible para las variantes encontradas, y al grado de asociación clínica entre el diagnóstico presuntivo (o los síntomas) del probando y el fenotipo patológico reportado para defectos en el gen que las contenga: i) categoría 1, casos donde se encontró una o más variantes conocidas con evidencia previa (ClinSig 4 o 5) de asociación con el diagnóstico presuntivo ii) categoría 2, casos donde se encontró, dentro de los genes asociados al diagnóstico presuntivo, una variante nueva potencialmente patogénica, acompañada de una variante conocida en un modelo de heterocigosis compuesta (categoría 2A), o sola en un modelo de herencia dominante (categoría 2B). En caso de que se hayan encontrado dos variantes de estas características en el mismo gen en un modelo de heterocigosis compuesta se clasificó al caso como 2C. iii) Categoría 3, aquellos casos donde se encontró una variante nueva potencialmente patogénica en genes con moderada

asociación con el fenotipo clínico y iv) categoría 4, casos donde no se encontró ninguna variante relevante para ser informada.

Tabla 9. Clasificación de los casos de acuerdo a las categorías 1 a 4.

Categoría		Descripción	# de Casos		# de Variantes encontradas
1		Casos donde se encontró una o más variantes conocidas con evidencia previa de asociación con el diagnóstico presuntivo.	31		45
2	**2A**	Casos donde se encontró, dentro de los genes asociados al diagnóstico presuntivo, una variante nueva potencialmente patogénica, acompañada de una variante conocida en un modelo de heterocigosis compuesta.	27	6	23
	2B	Casos donde se encontró, dentro de los genes asociados al diagnóstico presuntivo, una variante nueva potencialmente patogénica en un modelo de herencia dominante.		14	18
	2C	Casos donde se encontró, dentro de los genes asociados al diagnóstico presuntivo, dos variantes nuevas potencialmente patogénicas en el mismo gen en un modelo de heterocigosis compuesta.		7	25
3		Casos donde se encontró una variante nueva potencialmente patogénica en genes con moderada asociación con el fenotipo clínico.	17		25
4		No se encontró ninguna variante relevante para ser informada.	25		0
Total		Total de casos secuenciados	100		136

De los 100 casos analizados durante la campaña (Tabla 9), 31 corresponden a la categoría 1, que consideramos como casos exitosos y que, dado el conocimiento previo de la variante

y su asociación con el fenotipo patológico, representan un diagnóstico certero. Un ejemplo del mismo se presenta en la BOX1.

BOX 1: Caso Categoría 1

Un ejemplo de diagnóstico exitoso y certero lo comprende el caso de un paciente masculino adulto, aportado por el Dr. César Crespi del Hospital Interzonal Especializado de Agudos y Cronicos San Juan de Dios de La Plata, con un diagnóstico clínico presuntivo para la enfermedad de Wilson, una patología autosómica recesiva que entre los síntomas principales presenta cirrosis hepática, niveles altos de cobre en orina, con anillos de Kayser-Fleischer y cataratas. Se procedió a secuenciar al probando y al padre, y se priorizo un panel de 11 genes. No se encontraron variantes relevantes en homocigosis por lo que se procedió a buscar la coincidencia de dos variantes heterocigotas en el mismo gen (modelo de heterocigosis compuesta).

Se encontraron en el probando dos variantes (una compartida con el padre) en el gen ATP7B, que codifica para la proteína transportadora de cobre ATPasa de tipo P, denominada también proteína de la enfermedad de Wilson, una proteína transmembrana altamente conservada evolutivamente que posee roles esenciales en la fisiología humana, relacionados con el metabolismo del Cobre. Individuos que carecen de proteína ATP7B funcional, evidencian grandes dificultades en las vías de excreción de cobre, dando lugar a la enfermedad de Wilson. Ambas variantes encontradas se encuentran reportadas en ClinVar como patogénicas (NM_000053.3(ATP7B):c.3207C>A; y NM_000053.3(ATP7B):c.3955C>T) y han sido relacionadas a la enfermedad de Wilson por varios autores [68][69][70][71][72], poseen una frecuencia alélica muy baja en la base de datos de ExAC (no existen individuos en homocigosis para ninguna de ellas) y son predichas como "disease causing" por los softwares de predicción de patogenicidad.

La primera variante corresponde a un cambio de histidina por glutamina en la posición 1069 de la proteína, ha sido ampliamente estudiada y es la mutación más frecuentemente asociada a la enfermedad de Wilson. El mecanismo por el cual la mutación afecta la función, se cree está asociado a la desestabilización del sitio de unión a ATP. La otra mutación generaría un codón de terminación prematuro, que se supone da origen a una proteína truncada no funcional. El hecho de que ambas mutaciones hayan sido previamente reportadas como patogénicas y asociadas a la enfermedad de Wilson, y su correcta cigosidad y segregación en el probando y su padre, resultan en un diagnóstico de máxima confianza.

En la categoría 2 clasificamos 27 casos. Estos son, desde una perspectiva de descubrimiento, los más interesantes, ya que representan aquellos donde se han encontrado variantes nuevas con un potencial significativo para explicar el fenotipo patogénico. En estos casos es fundamental la evaluación del potencial patogénico de la

variante, para lo cual se consideran importantes diversos factores. Por un lado, se debe verificar que la variante posee una cigosidad correcta de acuerdo al modelo de herencia y verificación de la correcta segregación en el grupo familiar. Por ejemplo, en casos de herencia autosómica dominante (AD) la misma debe estar presente en heterocigosis en el probando y ausente en ambos padres (y si hubiera hermanos) sanos. En casos de enfermedades recesivas (AR), usualmente, si no hay consanguinidad, se presenta un modelo de heterocigosis compuesta, donde cada una de las variantes encontradas, está presente en heterocigosis en el probando y en uno (y sólo uno) de los padres.

Por otro lado, es fundamental determinar el potencial patogénico de la variante desde una perspectiva molecular analizando su impacto a nivel del gen o la proteína. En el caso de variantes que introducen un stop prematuro o producen un cambio en el marco de lectura, es razonable suponer que las mismas den lugar a una proteína no funcional. En el caso de que la variante resulte en un cambio de aminoácido, se debe profundizar el análisis considerando el tipo de cambio de residuo, la conservación del mismo en términos evolutivos, la frecuencia poblacional de la variante (si la hubiere), la predicción de patogenicidad por parte de algoritmos bioinformáticos, y si hubiere un análisis del efecto de la misma sobre la estructura proteica. Idealmente, las variantes candidatas deben ser predichas como patogénicas por todas estas propiedades. Ejemplos de estos casos se presentan en el BOX 2, BOX 3 y BOX 4.

BOX 2: Caso Categoría 2A

Paciente masculino con diagnóstico clínico tentativo de epilepsia mioclónica progresiva, aportado por el Dr. Santiago Chacón, del Hospital Centenario de Gualeguaychú, Entre Ríos. En base a toda su historia clínica, se confeccionó una lista de 65 genes a priorizar basándose en paneles existentes, genes extraídos de publicaciones científicas, y aquellos identificados mediante cruce de datos de distintas procedencias, tales como sintomatología del paciente y genes involucrados en patologías similares.

Al analizar el exoma del paciente se halló una variante en heterocigosis en el gen EPM2A, siendo éste uno de los dos genes responsables de la patología. La variante resultó causar el cambio de aminoácido Arginina 108 por Cisteína (c.322C>T) en la proteína laforina, la cual según Genetic Home Reference [73], parece jugar un rol fundamental en la supervivencia de las neuronas. El impacto de dicha variante sobre la estructura de la laforina puede observarse en la Figura 36, realizada mediante un análisis estructural de la proteína a nivel molecular. Dicha variante se encuentra descrita por ClinVar como variante patogénica para epilepsia mioclónica progresiva con código de acceso rs137852915. Por otra parte, en el mismo gen se halló una deleción en carácter también heterocigota, no descrita previamente, la cual produce la pérdida de 11 aminoácidos, con el consiguiente corrimiento en el marco de lectura. Dicha deleción se encuentra localizada dentro

del dominio cBM20 amino terminal (family 20 carbohydrate-binding module) de la proteína laforina, y muy cerca del sitio de unión a maltohexosa.

En este contexto, si bien la deleción de 11 aminoácidos no ha sido previamente informada (y por ende es de significado incierto), se encuentra en la región de unión a la maltohexosa, por lo que su efecto a nivel molecular y aparición en compañía de la otra variante patogénica ya reportada, ambas en heterocigosis, fortalecen la hipótesis de un caso de heterocigosis compuesta, siendo fuertes candidatas causales de la sintomatología presentada.

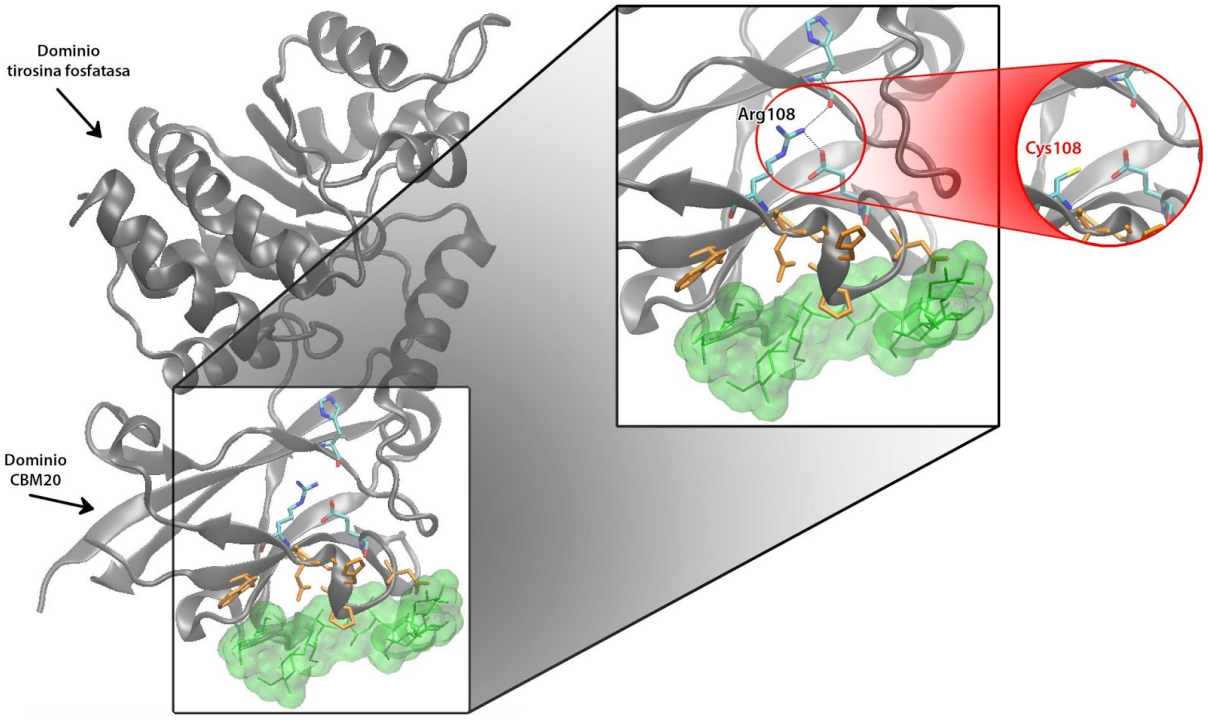

Figura 36. Representación de la estructura cristalográfica de la proteína laforina. Pueden observarse sus dos dominios: tirosina fosfatasa y CBM20. Este último posee un sitio de unión a maltohexosa un azúcar (resaltada en verde), y los aminoácidos involucrados en la unión de dicha molécula se encuentran resaltados en naranja. En la vista ampliada, puede verse cómo la Arg108 contribuye, mediante interacciones de puente de hidrógeno con otros residuos, a la estabilización de la estructura terciaria de cMB20. En la región resaltada en rojo puede observarse que la mutación Arg108 -> Cys imposibilita la formación de los puentes de hidrógeno mencionados, lo cual desestabiliza la estructura del sitio de unión a maltohexosa, resultando en una proteína no funcional.

BOX 3: Caso Categoría 2B

Paciente masculino con diagnóstico clínico tentativo de Disregulación Inmune no caracterizada, aportado por la división de Inmunología del "Hospital Prof. Dr. Juan P. Garrahan", sin un modelo de herencia definido. En base a análisis genéticos previos, los profesionales ya habían descartado la presencia de mutaciones en 4 genes (ALP, CASP8, CASP10, FAS) relacionados con los síntomas del paciente. Confeccionamos una lista de genes candidatos, tomando como punto de partida los genes causantes de más de 150 formas distintas de inmunodeficiencias primarias, a los que agregamos panel de genes para Linfocitosis congénita de células B y genes contenidos en paneles comerciales para desregulaciones inmunes. En total el panel de genes a priorizar contenía 207 genes. Realizando filtros en las variantes obtenidas de este conjunto de genes, (frecuencia poblacional menor al 1%, heterocigosis e impacto alto/moderado) se encontró una variante de tipo missense dentro de la región codificante del gen CARD11 (Caspase Recruitment Domain family member 11).

La variante encontrada produce el cambio de la Treonina 117 por una Prolina (p.Thr117Pro, NM_032415(CARD11):c.349A>C) y es considerada como patogénica por los predictores bioinformáticos SIFT, Polyphen y Mutation Taster; habiéndose reportado previamente en bibliografía un cambio aminoacídico en la misma posición (p.Thr117Ala) también en heterocigosis, en un paciente con un trastorno genético poco frecuente asociado con linfocitosis congénita de células B [74]. La prolina es un aminoácido no cargado que tiende a rigidizar la estructura y afectar, de esta forma, la función de la proteína resultante. Mutaciones similares en CARD11 han sido asociadas a la ganancia de función de un dominio funcional de la proteína, que normalmente desempeña un papel crítico en el mantenimiento de la misma en un estado inactivo, activando en las mutantes espontáneamente a NF-κ B y promoviendo la supervivencia de células B humanas de linfoma in vitro [75]. El gen CARD11 se encuentra asociado con el desarrollo de "Expansión de células B con NFKB y anergia de células T" (OMIM: 616452); trastorno con el que el paciente presenta muchos síntomas solapados, y en un modelo Dominante, consistente con la cigosidad de la variante encontrada. De esta forma, se la considera como potencialmente responsable del fenotipo observado con un alto grado de confianza, si bien es una variante novel y por lo tanto de significado incierto.

BOX 4: Caso Categoría 2C

Un ejemplo de esta categoría lo comprende el caso de un paciente femenino adulto con un diagnóstico clínico presuntivo de síndrome atáxico-oculomotor de tipo 2 (OMIM: 606002), una patología autosómica recesiva que presenta varios síntomas neuromotores. Para analizar el mismo se secuenció al probando y se analizó un panel de 222 genes. No se encontraron variantes relevantes en homocigosis, por lo que se procedió a buscar la coincidencia de dos variantes heterocigotas en un mismo gen (modelo de heterocigosis compuesta). Con este modelo de herencia, se encontraron dos variantes noveles en el gen SETX (OMIM: 608465), un gen que codifica para la proteína senataxina y se expresa en un amplio rango de tejidos, incluyendo el cerebro, la médula espinal y los músculos. La primera variante (p.Ser507fs: c.1518dupA) es de alto impacto funcional, ya que produce el corrimiento del marco de lectura por la adición de un nucleótido; mientras que la segunda (p.Val2385Gly: c.7154T>G) es una variante de cambio de aminoácido, de una valina a una glicina, un aminoácido más pequeño y no quiral. Ambas variantes son predichas como "disease causing" por "Mutation Tester". Mutaciones reportadas en este gen están asociadas al desarrollo de un trastorno neurodegenerativo caracterizado por el inicio en la edad adulta de ataxia cerebelosa progresiva, neuropatía periférica senso-motora axonal y aumento de la alfa-fetoproteína sérica [76][77], síntomas que se solapan con la Historia Clínica del probando.

En resumen, la predicción del efecto patogénico de ambas variantes y su presencia en el principal gen candidato, sugieren fuertemente que ambas variantes combinadas podrían ser las responsables del fenotipo observado.

La categoría 3, con 17 casos, comprende a aquellos que representan hipótesis de trabajo, tanto clínicas como moleculares. En estos casos, si bien se han encontrado variantes de significado incierto, éstas son desde el punto de vista molecular potencialmente patogénicas cumpliendo la mayoría de los criterios de análisis a nivel gen y/o proteína. En este contexto, los pasos a seguir son por un lado verificar la correcta segregación y cigosidad de la variante en el grupo familiar y en concordancia con el modelo de herencia de la enfermedad, si el mismo estuviera establecido. Por otro lado, se podría -al igual que con los casos de la categoría 2- avanzar con ensayos *in-vitro* para analizar el efecto de la mutación sobre la función proteica. En relación con la clínica, el objetivo debe tender a una revisión del solapamiento entre los síntomas presentados por el paciente, y aquellos reportados para otros pacientes con defectos en el gen donde se hallaron las variantes. Un ejemplo de un caso con este tipo de resultados se presenta en el BOX 5.

BOX 5. Caso Categoría 3

Paciente pediátrico masculino sin antecedentes familiares relevantes, cariotipo normal y diagnóstico tentativo de Síndrome de Sotos, un desorden autosómico dominante caracterizado, entre otros síntomas, por macrocefalia y crecimiento corporal excesivo en la infancia acompañado de retraso mental, problemas de conducta e hipotonía. La mayoría de los casos de síndrome de Sotos (95%) se producen por mutaciones de novo en el gen NSD1, o deleciones en la región 5q35.3, en la persona afectada. También se reportan algunos casos para mutaciones heterocigotas en el gen NFIX, o mutaciones en el gen APC2, en este último caso siguiendo un modelo autosómico recesivo.

A partir de la secuenciación exómica del probando se buscaron, en un primer momento, variantes presentes en los tres genes reportados para el síndrome de Sotos, sin encontrarse ninguna con relevancia clínica. Extendiendo la búsqueda a genes asociados a los síntomas reportados en la historia clínica del probando se encontró una variante novel en heterocigosis (NM_000264(PTCH1):c.1906A>G), en el gen PTCH1, que se traducen en el cambio aminoacídico de asparagina por ácido aspártico en la posición 636 de la proteína Patched-1. Mutation Taster la reporta como patogénica. El gen PTCH1 es considerado un gen supresor de tumores por su rol en la prevención de la proliferación celular incontrolada. Dentro de las patologías asociadas a este gen se encuentra el síndrome de Gorlin o Nevoid basal cell carcinoma syndrome (NBCCS) (OMIM #109400) y es considerada por la bibliografía como una condición que puede confundirse con Síndrome de Sotos [78]. Existe un solapamiento fenotípico moderado con los signos y síntomas del probando, aunque es posible que algunos de ellos que permitan la realización de un diagnóstico diferencial aún no se hayan desarrollado, dada la edad pediátrica del paciente. El síndrome NBCCS presenta herencia autosómica dominante que es compatible con la variante del probando. De esta forma, la variante encontrada se considera como potencialmente responsable del fenotipo observado, si bien es una variante novel y por lo tanto de significado incierto. Una vez informada al médico la analizará junto a la evidencia clínica para validar o rechazar el nuevo diagnóstico hipotético; de ser confirmado, permitirá acompañar la evolución del paciente de manera eficiente.

Finalmente, la categoría 4 comprende aquellos 25 casos donde no se ha encontrado ninguna variante que pueda explicar el fenotipo observado. Estos casos, si bien se los puede considerar abiertos, la falta de una hipótesis de trabajo luego de un análisis exhaustivo, que puede involucrar dos o hasta tres ciclos de evaluación de genes candidatos, sugiere que poseen una causa molecular subyacente que no ha podido ser determinada por la técnica utilizada.

Discusión sobre la campaña 100 exomas.

Los resultados de la iniciativa “100 Exomas”, si bien no representan una cantidad de casos suficientes como para realizar afirmaciones estadísticamente significativas sobre la capacidad de la secuenciación exómica para diagnosticar de manera precisa las causas moleculares de las EPoFs o para avanzar en el descubrimiento de nuevos genes y variantes patogénicas, permiten evaluar en las capacidades locales, con sus virtudes y dificultades, para la implementación de estas tecnologías en contextos de investigación científica e innovación clínica, y compararlos con experiencias realizadas en los países del denominado “primer mundo”. Para llevar adelante la campaña, armamos un protocolo de trabajo con una fuerte base interdisciplinaria, desde los médicos responsables de los casos clínicos a biólogos, bioinformáticos y expertos en computación, que permitió un entendimiento completo y exhaustivo que abarca desde las características fenotípicas del paciente a las causas moleculares intrínsecas.
Si consideramos aquellos casos donde se encontraron variantes de asociación clínica conocida (categoría 1) o con altas chances de ser patogénicas (categoría 2) como exitosos, vemos que la tasa de éxito es superior al 50%, lo que es levemente superior a lo reportado en la literatura (Shashi et.al., 2014)[79]. La aparente performance superior podría deberse a dos motivos: ya sea una selección más sesgada de los casos con alta probabilidad de obtener un diagnóstico y/o la definición de caso exitoso. Si, por ejemplo, consideramos como casos exitosos sólo aquellos incluidos en la categoría 1, entonces el porcentaje de casos que han llegado a un diagnóstico se acerca más al 30%, que es el valor usualmente tomado como de referencia para análisis de exomas.

Desde una perspectiva de ciencia básica, los casos más interesantes son los de categoría 2, donde se encuentran una (o dos) variantes nóveles, con una alta probabilidad de ser responsable del fenotipo observado. Estas variantes muchas veces representan el punto de partida para el estudio de los fenómenos moleculares subyacentes al desarrollo patológico, al señalar cómo su efecto sobre la función proteica repercute en la fisiología molecular y celular. Un ejemplo exitoso de este tipo de caso, derivado de la presente campaña, es el de una inmunodeficiencia combinada que se origina en una mutación en heterocigosis en el gen CARD11, la cual comprende una inserción de 14 aminoácidos. La misma fue estudiada en colaboración con un grupo del Instituto Nacional de Salud de EEUU (NIH), junto con otras tres mutaciones, todas en heterocigosis en diferentes dominios de CARD11, mostrando que el efecto patogénico se debía a una pérdida de función con interferencia sobre la proteína salvaje. Esto dió lugar a un cuadro autosómico dominante, que puso en evidencia detalles del funcionamiento de diferentes vías de señalización de los linfocitos T, algo que culminó en un trabajo en la prestigiosa revista Nature Genetics. (Ma et al., 2017)[81]
Otros de los casos de esta categoría que han dado lugar a resultados del mismo tipo se encuentran actualmente siendo analizados en el grupo y en el marco de colaboraciones nacionales e internacionales.

Es interesante también analizar cuáles son los motivos subyacentes que resultan en casos donde no se ha llegado al diagnóstico molecular, es decir, aquellos casos de categoría 3 o 4 donde falle la hipótesis, no se confirme la asociación propuesta entre genotipo (variante) y fenotipo del paciente y/o la variante propuesta no segregue adecuadamente en el grupo familiar. En este sentido, hay diversos motivos asociados a un resultado de este tipo, dependiendo principalmente de la naturaleza del caso. En algunos casos puede suceder que las mutaciones responsables no fueron reveladas por el tipo de experimento de secuenciación, ya sea porque están en exones pobremente capturados y/o regiones no codificantes o porque corresponden a inserciones o deleciones de tamaño superior al que puede ser determinado por la secuenciación exómica (comúnmente conocidas como variaciones en el número de copias). En estos casos la única posibilidad para avanzar hacia el diagnóstico es realizar una secuenciación genómica completa. En otros casos, un resultado de este tipo puede estar asociado a la falta de mejores - en términos de la precisión sintomatológica - relaciones genotipo-fenotipo. Este área es una de las de mayor expansión, y existen varias iniciativas internacionales para promover su desarrollo. Una de ellas, PhenomeCentral [80] es un portal centralizado que permite cargar casos de EPoF no resueltos. Dicho portal utiliza un sistema automático para evaluar similitud de fenotipos y alertar a los investigadores sobre las coincidencias halladas para, mediante la evaluación de múltiples casos y cooperación internacional, poder reforzar hipótesis de asociación variante-fenotipo y/o avanzar hacia ensayos de validación funcional.

Finalmente, resulta relevante evaluar los resultados de la campaña desde una perspectiva asociada al trabajo conjunto de los distintos tipos de profesionales, con sus experiencias y conocimientos especializados requeridos para la implementación de un servicio de diagnóstico genómico. La experiencia de esta campaña permite darle peso y reforzar la siguiente filosofía de trabajo interdisciplinario: si bien la división de tareas es un proceso lógico, práctico y adecuado para muchos trabajos, tales como el relevamiento de la historia clínica por parte del médico, la toma y procesamiento de la muestra por el bioquímico de laboratorio, el procesamiento de datos por parte del bioinformático, la priorización de variantes por parte de bioquímicos/biólogos moleculares y finalmente la interpretación de las mismas y devolución al paciente por parte del médico; la eficiencia y las chances de éxito se maximizan cuando todos los profesionales trabajan en conjunto. Particularmente, esto sucede con mayor hincapié en la etapa de priorización e interpretación de las variantes, la cual requiere integrar conocimientos asociados al experimento de NGS (cobertura, profundidad, calidad de la variante, entre otras), al impacto a nivel molecular de la misma y, por supuesto, a su potencial relación con la clínica. En este contexto, a través de esta experiencia, se destaca que la mayor probabilidad de éxito se asocia a un diagnóstico presuntivo preciso, y a un profundo conocimiento por parte del equipo médico de los genes y variantes patogénicas conocidas en casos similares. Todo esto contribuye a evidenciar y reforzar la importancia de brindar a nuestros futuros médicos, conocimientos y capacitación en el área de la genómica clínica.

Cierre en el contexto de esta tesis.

El desarrollo exitoso de la campaña “100 Exomas” resultó, en una primera instancia, en un testeo y puesta a punto de la *B-Platform* desarrollada en este trabajo. *Su utilización con éxito en la campaña 100 Exomas, permitió, tomando como punto de partida los datos obtenidos del secuenciador, llegar a la interpretación biológica y clínica de las variantes encontradas en cada caso. Además, de poder validar el protocolo diseñado.*

En este contexto la plataforma ya fue utilizada por diferentes profesionales de la Salud (Médicos, Bioquímicos, Genetistas, Biólogos) de más de 30 instituciones de distintas provincias, para el procesamiento y análisis de los exomas secuenciados para la campaña.

Como prueba de concepto, los resultados de esta tesis muestran que fue posible implementar a nivel local servicios de diagnóstico molecular basados en secuenciación de próxima generación. Dicha implementación se sustenta en el conocimiento y capacidad de los profesionales locales, y es un ejemplo que puede servir como semilla para extenderlo a escala regional. Esta prueba de concepto es fundamental para la nacionalización de este tipo de servicios que hoy en día se encuentran disponibles internacionalmente.

Caso II: Evaluación de un perfil de riesgo

Introducción

Como parte de este trabajo de tesis, y utilizando el protocolo definido anteriormente en los capítulos de cálculo de perfil de riesgo y perfil farmacogenético, se decidió realizar una prueba de concepto, mediante la evaluación de un individuo adulto, sano, que entre los puntos a destacar de su historia clínica presentaba antecedentes familiares de enfermedad vascular e hipertensión.

Como ya mencionamos, entre las características de una enfermedad compleja como la enfermedad vascular o la hipertensión está su heterogeneidad genética. Es decir, que las mutaciones en diferentes genes pueden conducir a la misma enfermedad. También lo es su expresividad variable, o sea que individuos con el mismo genotipo también pueden mostrar diferentes grados del mismo fenotipo. Y finalmente, el gran impacto del ambiente (en sentido amplio), que mediante su control, permite tomar medidas preventivas.

Dentro de los signos (o características) que podrían sugerir una predisposición genética subyacente para el desarrollo de una enfermedad compleja, podemos mencionar :

- La edad de aparición es más temprana de lo esperado. Un ejemplo aquí sería un infarto de miocardio o un ataque al corazón que ocurre en una persona de 25 años contra uno que ocurre en una persona de 80 años.
- Cuándo una condición que ocurre en el sexo menos afectado. Por ejemplo, para el cáncer de mama, por supuesto, las mujeres son las principales afectadas, pero si se tiene un hombre afectado, podría ser un indicio de que hay una etiología genética subyacente.
- La historia familiar, si existe una historia familiar de una enfermedad, en particular una enfermedad que ocurre en múltiples generaciones, tal vez a una edad más temprana de aparición que lo habitual. Eso podría ser un indicio de que hay una base genética.
- Y por último, cuando la enfermedad ocurre en ausencia de factores de riesgo ambientales conocidos. Por ejemplo, una enfermedad como la diabetes, donde la mayoría de las personas con diabetes también tienen obesidad. Es una comorbilidad y quizás una causa de la diabetes. Pero entonces si una persona delgada tiene diabetes se podría pensar que tal vez hay alguna etiología genética.

Entonces para este caso de estudio, si bien la determinación del perfil de riesgo y farmacogenómico incluyó el análisis de los 290 marcadores (polimorfismos de un solo nucleótido) asociados a las patologías y fármacos listados en las “Tabla 6: Lista de patologías seleccionadas.” y “Tabla 7: Fármacos seleccionados de PharmGKB”, respectivamente, a los fines de ajustar a la extensión de este trabajo solo se mencionan los rasgos más relevantes asociados a la historia clínica del paciente.

Métodos

Genotipificación de la muestra

La genotipificación se realizó mediante un microarray (SNP Genotyping Microarray). Para este caso se utilizó el chip *Infinium Global Screening Array v2.0*. Mediante un servicio provisto por la compañía Macrogen.

Protocolo de procesamiento de datos

Como resultado del genotipado se recibe un archivo llamado gtReport.txt que contiene los genotipos de cada muestra y al cual nos referimos como “array”. Junto con este archivo viene uno llamado GSAMG2_SNP_Info.txt (al cual nos referimos como “Info”). Este archivo contiene información sobre cada posición (SNP) interrogada en el array, como cuáles son los dos posibles alelos que lee, etc. Además es muy importante porque relaciona el nombre de la variante que figura en el array con un código rs. Los alelos reportados en el array no corresponden todos a una misma hebra. Para nuestro análisis usamos la hebra reportada por dbSNP, y para corregir esto usamos el archivo manifest, que puede descargarse de la página de Illumina.

Riesgo

Si bien a la persona se le evaluaron los 290 marcadores que permitieron calcular el riesgo para más de 75 rasgos (patologías) asociados a su perfil genético, en esta tesis para no excedernos en la extensión de la misma, sólo se mencionan los dos rasgos de interés asociados a la historia clínica: Infarto agudo de miocardio e Hipertensión arterial.

Infarto agudo de miocardio

El infarto agudo de miocardio (IAM) es una patología con alta incidencia en la población y es causada por la obstrucción u oclusión de alguna de las principales arterias que llevan sangre al corazón para nutrirse, generando así un déficit del aporte de los nutrientes que el músculo cardíaco necesita para subsistir y consecuentemente muerte célular. Como consecuencia, pueden surgir diversas complicaciones: insuficiencia cardíaca, arritmias graves, ruptura cardíaca, daño de alguna de las válvulas del corazón, etc.

El desarrollo de la enfermedad coronaria depende de múltiples factores, entre los que se asocian a mayor riesgo de desarrollar IAM son: la edad (a mayor edad, más probabilidades de desarrollar enfermedad coronaria), sexo masculino, antecedentes familiares de IAM o muerte súbita, consumo de tabaco, dieta no saludable, sedentarismo, hipertensión arterial, sobrepeso/obesidad, consumo de alcohol, consumo de drogas como cocaína ydiabetes.

Se han identificado múltiples variantes genéticas que se asocian con mayor riesgo de desarrollarla. En este caso analizamos 11 marcadores, que en conjunto pueden aumentar el riesgo de atravesar un infarto de miocardio hasta 4 veces.

De acuerdo a los marcadores genéticos que se analizaron y para los cuales se calculó el Odd Ratio (OR), según lo explicado en la sección correspondiente, el individuo analizado posee una probabilidad levemente aumentada de desarrollar infarto agudo de miocardio. Esto significa que el OR calculado (score=2,6846) dio mayor al percentil 5th qué era 2.52. Es decir, se encuentra dentro del 5% de las personas de su edad y sexo para las cuales su genoma contribuye con una probabilidad aumentada de presentar IAM.

En base a esta información la persona puede tomar acciones para reducir este riesgo de IAM, como, por ejemplo, ajustar su estilo de vida, realizarse controles médicos de rutina que incluyan chequeos cardiológicos, de presión arterial y análisis de laboratorio. Estar atento ante la presencia de alguno de los síntomas, ya que la detección temprana de un infarto es vital para realizar el tratamiento precoz y reducir complicaciones. Puede seguir una dieta saludable, hacer ejercicio en forma regular, mantener un peso saludable acorde a su edad, altura y momento de la vida.

Hipertensión arterial

La tensión arterial (TA) es la medición de la fuerza que la sangre ejerce sobre las paredes de la arteria al momento que el corazón realiza un latido. Se compone por dos variables: Presión arterial sistólica o “máxima” y presión arterial diastólica o “mínima”. Según las guías argentinas de cardiología, la HiperTensión Arterial (HTA) se define por una TA > 140 mmHg de sistólica y una diastólica >90 mmHg. La HTA, también conocida como “el mal silencioso” lleva ese sobrenombre porque, en la gran mayoría de los casos, se presenta sin síntomas o con algunos poco específicos que podrían atribuirse a otras causas.

Cuando la hipertensión arterial se encuentra presente por largo tiempo puede causar daños en algunos órganos por lo que la detección temprana es importante para evitar estos riesgos. A pesar de que el ambiente juega un papel importante en esta enfermedad, se han identificado múltiples variantes genéticas que se asocian con mayor riesgo de desarrollarla. Un diagnóstico temprano de presión arterial alta, con un tratamiento adecuado, puede ayudar a prevenir complicaciones.

La HTA no controlada puede ocasionar lesión en diferentes órganos, dentro de los cuales se incluyen el corazón, el riñón, ojo y cerebro, aumentando el riesgo de desarrollar ataque cardíaco (infarto agudo de miocardio) o accidente cerebrovascular (ACV), aneurismas arteriales, insuficiencia cardíaca, enfermedad renal vascular (que puede conducir a insuficiencia renal), enfermedad vascular ocular (que puede conducir a ceguera), enfermedad vascular cerebral (que puede manifestarse con problemas con la memoria o el entendimiento o demencia).

Para el individuo bajo estudio, al igual que con IAM, se calculó el Odd Ratio en función de los marcadores analizados y resultó en un score=1,6297 que se encuentra entre los percentiles 5th=10.01 y 20th=1.01, lo que representa una probabilidad de presentar HTA moderada.

Al contar con esta información, al igual que con la de infarto de miocardio, sobre HTA la persona puede accionar sobre su salud.

Discusión

La medicina del futuro (y en este caso el futuro es hoy) requiere de la capacidad de procesar rápidamente, integrar y analizar cuidadosamente la enorme cantidad de datos que se generan en relación con nuestra salud. El cuerpo humano se convierte entonces en una fuente de datos, comenzando por su genoma (o parte de él), e integrando progresivamente todos aquellos datos derivados de análisis clínicos, en una gran historia clínica digital. En este contexto, de uso y aplicación de la genómica en la clínica, la bioinformática se tornó esencial para el correcto uso y aprovechamiento de esta información . En particular, la genómica, mediante la aplicación de tecnologías de NGS, está cambiando el paradigma de la medicina en sus tres aspectos fundamentales: la prevención, el diagnóstico y el tratamiento, potenciando, de esta manera, la revolución de la medicina personal y de precisión.

En el diagnóstico, principalmente de enfermedades poco frecuentes, es donde las tecnologías NGS tuvieron mayor impacto, al permitirnos buscar de manera precisa cuál es la variante (o mutación) responsable del desarrollo de la enfermedad, otorgando al paciente y su familia un diagnóstico molecular preciso y certero. Respecto al tratamiento, el impacto producido por la genómica una vez obtenido un diagnóstico, si es que existe un tratamiento disponible, se podría ajustar el tratamiento farmacológico en función de las características genéticas del paciente para mejorar la eficacia, o reducir los efectos secundarios, de los fármacos en su organismo. Este es el paradigma de la farmacogenómica.
Por último, en cuanto a la prevención, como mencionamos varias veces a lo largo de este trabajo de tesis, el genoma del individuo puede contribuir, al menos parcialmente en el potencial -o riesgo- de desarrollar enfermedades complejas de alta prevalencia en el adulto, como por ejemplo las enfermedades cardiovasculares, la diabetes tipo 2, la obesidad, el alzheimer, parkinson o el cáncer entre otras. Entonces el análisis de riesgo diferencial, permite tomar acciones preventivas, cambios de hábitos, de dieta. etc.
Lo que buscamos cubrir con este trabajo son los desarrollos bioinformáticos necesarios para avanzar en estos tres aspectos y de este modo abordar los desafíos de la genómica en la clínica.

Nuestra hipótesis de trabajo, es que una de las cuestiones clave para que la medicina de precisión incremente su adopción en la práctica clínica, es contar con una herramienta sólida de integración de datos genómicos que sirva de apoyo a la toma de decisiones mediante el análisis de los datos y presente los resultados en un formato fácilmente interpretable para los profesionales de la salud.

Relevancia e importancia del protocolo de procesamiento de datos genómicos

Tomando como punto de partida el desafío bioinformático que implica analizar los datos genómicos humanos, nos propusimos durante el desarrollo de esta tesis implementar los protocolos y herramientas de software *B-Platform* para que aquellos profesionales, tanto en el ámbito académico, como en la práctica médica, que contaran con datos genómicos, pudieran analizarlos de manera eficiente y, de este modo, llevar realmente la genómica a la clínica.

Los resultados, de los casos de aplicación de esta tesis (*Caso I: La campaña 100 exomas y Caso II: Evaluación de un perfil de riesgo)* muestran que combinando e integrando metodologías, componentes de software (BWA, SNPeff, SAMtools, etc.), y bases de datos (OMIM, ClinVar, GWAS, etc) disponibles y ampliamente utilizados y validados por la comunidad, sumados a un conjunto de desarrollos propios para vincular y completar y complementar los mismos, fue posible desarrollar un protocolo, que permite hoy en día de manera exitosa, a partir de los datos de secuenciación de NGS (y/o SNP microarray si fuera necesario) precisar y anotar con una elevada cantidad de contenido biológico todas las variantes (mutaciones) de la muestra correspondiente.

El archivo resultante, referido usualmente como VCF anotado, es conceptualmente una descripción detallada de la información genómica del individuo del que se tomó la muestra, y técnicamente una versión resumida (y anotada) de su genoma, que comprende el punto de partida para el análisis del mismo por parte de los profesionales de la salud.

La relevancia del VCF anotado (y el proceso que lo sustenta) es enorme, ya que un error en alguno de los pasos resultará en un archivo mal anotado, que contendrá inexactitudes y puede resultar en un análisis y por lo tanto diagnóstico erróneo o la incapacidad de llegar a uno. Por otro lado, el VCF anotado es un registro portable que permite re-analizar los datos en cualquier momento del futuro. Finalmente, la cantidad y calidad de la anotación son determinantes para el potencial éxito del análisis clínico, y por ende para llegar a un diagnóstico y/o hacer una correcta y relevante evaluación de perfil de riesgo y perfil farmacogenómico. El protocolo presentado en esta tesis ha sido (y es actualmente) utilizado por varios grupos de investigación y profesionales de la salud en nuestro país y también de diferentes grupos de otros países de latinoamérica.

La importancia de una herramienta de análisis eficiente e intuitiva.

Nuevamente, el VCF anotado es el punto de partida para el análisis subsiguiente, y es aquí donde se evidencia el enorme potencial de una herramienta como la desarrollada en esta tesis, que permite realizar un análisis profundo y detallado de la información genómica del individuo de una manera amigable. Debemos aquí distinguir dos casos, que oportunamente fueron presentados en capítulos separados, por un lado el análisis para diagnóstico molecular de casos de enfermedades mendelianas, y por otro, el análisis de un perfil de riesgo de desarrollar enfermedades complejas y su perfil farmacogenómico.

Diagnóstico molecular de enfermedades mendelianas por NGS: de la promesa a la realidad

Como mencionamos en la introducción de este trabajo quizás los primeros beneficiados de los desarrollos asociados al genoma humano y la secuenciación de próxima generación son aquellos pacientes que poseen (o al menos esa es la sospecha clínica) una enfermedad genética (típicamente mendeliana), pero que dada su condición “poco” frecuente, usualmente resulta en un diagnóstico tardío, equivocado, luego de un proceso extremadamente tedioso para el paciente y su grupo familiar, que ha llevado a acuñar el término “odisea al diagnóstico”.

En algunas situaciones, se tiene un paciente, que tiene lo que parece ser un trastorno mendeliano, pero en realidad no se puede hacer un diagnóstico clínico certero utilizando todas las pruebas estándar que se realizan para intentar llegar a un diagnóstico, como análisis bioquímicos, estudios de sangre y orina, radiológicos, etc. En estas situaciones, a veces se puede encontrar una alteración genética (una variante) que podría explicar por qué el paciente tiene el problema que tiene, siendo la misma entonces diagnóstica, y así podemos terminar con la odisea.

En cuanto al diagnóstico molecular de casos de enfermedades mendelianas, los resultados de la campaña 100 exomas muestran que contar con un protocolo de procesamiento de datos y herramientas bioinformáticas como las definidas en esta tesis permitieron que la tasa de éxito - casos donde se llegó al diagnóstico molecular – se ubique entre el 30-50% mostrando un métrica consistente con estudios realizados a nivel internacional. Varios de los casos, además han permitido encontrar variantes nóveles que representan el punto de partida para el estudio de los fenómenos moleculares subyacentes al desarrollo patológico, algunos de los cuales han dado lugar a publicaciones de primer nivel en el marco de colaboraciones internacionales (Ma et al., 2017)[81].
Desde una perspectiva asociada al trabajo en la clínica, la experiencia también muestra que la eficiencia y las chances de éxito se maximizan cuando los profesionales de la salud cuentan con herramientas adecuadas (protocolo de procesamiento y plataforma

bioinformática) y trabajan en conjunto, particularmente en la etapa de priorización e interpretación de las variantes.
Esto quedó en evidencia cuando pudimos disponer la plataforma para que sea utilizada por diferentes laboratorios de investigación y servicios internos en dos hospitales públicos ***volviendo realidad las promesas de llevar la genómica a la práctica clínica.*** La interacción fluida con los diferentes profesionales de la salud de estos grupos de investigación permitió una retroalimentación positiva que derivó en un proceso de mejora continua de la plataforma. Fue así que se logró construir una herramienta flexible, ajustada a los requerimientos de los profesionales locales, sin restricciones, convirtiéndo esto en una ventaja sobre otras plataformas disponibles. Además contar con el feedback de estos diferentes usuarios nos permite desarrollar un plan de desarrollos futuros, de nuevas funcionalidades a agregar a la plataforma como, por ejemplo, la incorporación del análisis y visualización de los CNVs (Variación en el número de copias), la reclasificación manual de las variantes según los criterios de ACMG, creación de nuevos filtros basados en HPO (Human Phenotype Ontology), y/o la incorporación de algoritmos de inteligencia artificial que nos permitan priorizar los genes en función del fenotipo del paciente, entre otras.

Finalmente, los resultados y la experiencia obtenidos durante el desarrollo de esta tesis evidencian la enorme relevancia e impacto que tiene la utilización de este tipo de herramientas y el trabajo multidisciplinario a la hora de contar con asesoramiento y una posible orientación diagnóstica en un alto (o moderado) porcentaje de los casos. Se espera continuar trabajando de esta forma para acercar estas tecnologías a pacientes de todo el país, de nuestros hermanos latinoamericanos y por qué no, del mundo.

Evaluación de un perfil de riesgo y perfil farmacogenómico: hacia una medicina preventiva

Con respecto al cálculo del perfil de riesgo y perfil farmacogenómico, como ya mencionamos anteriormente en la sección correspondiente, el objetivo en estos casos es actuar no en el contexto de diagnóstico, sino en la prevención y tratamiento, respectivamente. En particular, cuando evaluamos perfil de riesgo, debido al bagaje genético del individuo, nos focalizamos en enfermedades que desde el punto de vista de su etiología son complejas, o sea son poligénicas y poseen un alto impacto del ambiente, y que además son de alta prevalencia en la población, ya que uno de los elementos clave es comprenderlas en el contexto poblacional.

Como muestran los resultados en la sección *Caso II: Evaluación de un perfil de riesgo,* una de las claves para arribar a un diagnóstico preciso y la consecuente evaluación de riesgo, reside en la capacidad de interpretar y valorar las variantes encontradas, lo que sumado a la información de nuestro estado metabólico (clásico análisis de sangre), historia familiar, y estilo de vida (entre otros parámetros), nos brinda una mejor estimación de los riesgos, de padecer enfermedades comunes y complejas como los desórdenes cardiovasculares (infarto de miocardio, hipertensión), diabetes, alzheimer, parkinson, cáncer etc, lo que nos permitirá tomar medidas preventivas o guiar un tratamiento.

Una vez realizada la evaluación de riesgo, el resultado le permitirá al individuo consultar al profesional médico para obtener recomendaciones y determinar la estrategia terapéutica preventiva a seguir. Por ejemplo, las recomendaciones podrían ir desde el incremento en la frecuencia de los chequeos, la disminución en la edad a la que se aconseja comenzar un tratamiento específico, cambios en el estilo de vida y/o la alimentación, hasta la realización de cirugías preventivas y/o la toma de medicación preventiva.

Reconocer que se posee una o varias variantes que confieren un alto riesgo de desarrollar una patología, permite al profesional médico abordar estrategias preventivas y en particular saber cuáles son las mutaciones específicas podrían permitir determinar cuál es el plan terapéutico conveniente a utilizar en cada caso, contribuyendo a una medicina personalizada y de precisión.

Perspectivas Futuras:

La secuenciación genómica se perfila como la herramienta ideal para hacer realidad las promesas de la Medicina Personalizada o mejor dicho Medicina de Precisión. Es fácil ver el potencial que tiene la secuenciación genómica para incorporarse en la práctica clínica rutinaria en un plazo relativamente breve. Ante este escenario es preciso plantearse el impacto que esta práctica va a tener en los sistemas sanitarios y en la toma de las medidas adecuadas para su futura implantación. Sin embargo, a pesar de su indiscutible potencial, hay diversos aspectos que deben ser trabajados en detalle para garantizar el éxito.

Primero, hay un enorme reto tecnológico en cuanto a la propia obtención de estos datos, su fiabilidad y su certificación, como prueba clínica aceptada. Esta etapa está avanzando rápidamente pero se debe estar actualizado y conocer las limitaciones de cada una de las metodologías disponibles. En este sentido, otro de los problemas importantes es la falta de formación de los profesionales de la salud, que deben enfrentarse a un conjunto de datos completamente distinto de lo que estaban habituados. La gran mayoría de los médicos en ejercicio, no aprendió de genómica durante su formación (ya que esta no existía), por eso debemos también prestar gran atención a los aspectos asociados con la formación profesional en el área.

Es importante remarcar, que existen obstáculos potenciales al aprovechamiento óptimo de los datos genómicos que son de carácter legal, en relación con la legislación sobre la protección de datos, o de carácter ético, en el sentido de la información no solicitada o los hallazgos incidentales (en inglés incidental findings) que se obtiene colateralmente en una secuenciación genómica hecha con un objetivo determinado. El consentimiento informado y una clara discusión con el paciente son necesarias tal como se hace con otras técnicas de manera rutinaria en la clínica.

Desde un punto de vista bioinformático, el manejo, el almacenamiento y la consulta del enorme volumen de datos que estas tecnologías producen es un problema de complicada solución con el hardware y software existentes en la actualidad. Por otra parte, la utilidad clínica de la secuenciación genómica depende críticamente de la presentación de los resultados a los profesionales de la salud de una forma que le sea inteligible y clínicamente útil. Es en este sentido que esta problemática debe atacarse para permitir que las técnicas de NGS se usen de forma sistemática. Este desafío requiere del armado de equipos interdisciplinarios para el desarrollo de soluciones informáticas que cumplan con todos los requerimientos.

Siendo un poco más específicos en relación con los desafíos relacionados a los temas abordados en la presente tesis, podemos destacar en la siguiente sección las áreas donde se esperan desarrollos futuros, que mejoren la eficiencia y calidad de los procesos.

Desafíos futuros para el diagnóstico molecular de enfermedades mendelianas por NGS

Un primer aspecto a considerar para desarrollos futuros, es por ejemplo, dentro de los desafíos del diagnóstico molecular de enfermedades mendelianas por NGS, es la clasificación de las variantes noveles ya sea como benignas o patogénicas. Cualquiera que haya realizado un análisis genómico (ya sea de panel, exoma, o genoma) sabe que la mayoría de las variantes encontradas son de tipo VUS (Variant of Uncertain Significance) y esto es un enorme obstáculo al momento de decidir su relevancia en la clínica. Si bien, la continua aplicación de NGS en este área, genera un aumento continuo en la cantidad de variantes sobre las que conocemos su efecto, el número de variantes en la práctica es casi infinito, y en el corto o mediano plazo es imposible que todas hayan sido analizadas. Es por ello que se deben desarrollar nuevas y mejores metodologías para su clasificación. Al respecto, podemos dividir el trabajo en 3 áreas.

Por un lado, existe un continuo desarrollo de métodos bioinformáticos para la valoración de variantes. Estas metodologías son cada vez más precisas, y más complejas, y han incorporado recientemente elementos de Inteligencia Artificial o Aprendizaje Automático (Machine Learning), que mejoran significativamente su performance. Sin embargo, sigue existiendo un sesgo hacia las regiones codificantes, y también problemas de recursividad entre los conjuntos de entrenamiento y testeo, que resultan en cierto sobreajuste de los parámetros. *El problema, de estos sesgos en la práctica clínica, es que la mayoría de los métodos siguen sin poder proveer información relevante en aquellas variantes donde más se lo necesita, y dan respuestas certeras, para aquellas variantes donde el nivel de conocimiento disponible suele ser suficiente para su clasificación.*

Por otro lado, la falta de un proyecto sistemático para la evaluación funcional masiva de variantes en ciertos genes claves responsables de este tipo de patologías. En los últimos años la combinación de técnicas de NGS con síntesis de oligos en arreglos de alta densidad ha permitido el desarrollo de ensayos funcionales que permiten evaluar en único experimento todas las variantes de una región de un gen de interés (usualmente algún dominio proteico particular). Cuando nos referimos a todas las variantes, hablamos por ejemplo de que si el dominio posee 100 residuos de longitud, se evalúan las 100 x 20 variantes posibles. Estas técnicas han sido aplicadas a algunos genes de enorme relevancia como BRCA1, entre otros, pero no existen proyectos de gran escala internacional que sistematicen y organicen los esfuerzos en esta dirección.

Por último, y en relación con el punto anterior, existen algunas características relevantes de las variantes para las cuales no existen bases de datos públicas. Dos de ellas, muy necesarias, son una base de datos de ensayos funcionales de variantes en genes/proteínas, o base de datos de variantes que alteran el proceso de splicing.
En resumen, en el área de la clasificación de las variantes en relación con su relevancia clínica existen varios puntos donde esperamos contribuir con desarrollos futuros.

Un segundo aspecto, como ya se mencionó, es la priorización de genes. Tener una lista de genes priorizados en función del fenotipo patológico observado en el paciente, es fundamental para acotar la búsqueda de variantes diagnósticas. Actualmente, si bien existen diversas herramientas (findZebra, phenomizer, phenolyzer, etc.) que permiten dada una historia clínica, acotar una lista de genes candidatos, las mismas son poco precisas, y suelen otorgar lista de genes candidatos muy extensas, donde los mismos no están ordenados según cuál es su probabilidad (o chance) de contener la variante diagnóstica, dada una historia clínica particular. Por lo tanto, avanzar en el desarrollo de modelos cuantitativos que puedan priorizar y valorar los genes en función de la histórica clínica ingresada, es uno de los próximos objetivos en el corto plazo.

Un tercer aspecto a considerar es la determinación de las variaciones en el número de copias (CNV, copy number variants) a partir de NGS. Los CNV, son deleciones o duplicaciones de pequeños segmentos de ADN en el genoma. La patogenicidad de un CNV depende del tamaño y la ubicación del mismo. El tamaño es importante, porque cuanto más grandes sea el CNV más número de genes se verán afectados y el efecto causado será más deletéreo o patogénico. Pero la ubicación también es muy importante. Los CNV intergénicos, es decir, los que se producen entre genes, no suelen ser patogénicos, mientras que las que abarcan regiones codificantes sí lo son.

¿Cómo es que estas variantes del número de copias causan enfermedades? Si se borra una de las copias de un gen, sólo se producirá la mitad de las proteínas expresadas por dicho gen. O en su defecto, si se adiciona una copia del gen, se pueden tener demasiadas proteínas. La expresión de proteínas en un tipo de célula en la cantidad equivocada, puede ser muy perjudicial para la salud. Por ejemplo, cuando se tiene una enfermedad recesiva es consecuencia de una mutación en ambas copias del gen. Pero si tenemos un CNV que es una deleción sólo se requiere tener una copia del gen mutado para desarrollar la enfermedad, porque su otra copia se elimina completamente causando una pérdida de función de ese gen.

La técnica estándar para detectar este tipo de variantes es MLPA (multiplex ligation-dependent probe amplification) o la utilización de hibridación genómica completa por microarreglos (CGH microarray). Sin embargo, cuando tenemos datos de secuenciación NGS, algunos CNV también pueden ser detectados de forma fiable dependiendo de dónde están y de su tamaño.

Existen diferentes algoritmos para detectar CNV a partir de datos NGS, pero tienen ciertas limitaciones. Estos se basan en la inferencia estadística de los CNVs a partir de la comparación de muestras entre sí. Entonces para un nivel de cobertura de referencia confiable para los cálculos, se necesita un número suficiente, usualmente alto, de muestras de referencia. Normalmente se utilizan como mínimo 20 muestras de una misma corrida, es decir, que las muestras utilizadas en el cálculo idealmente deben tener las mismas condiciones experimentales.

Además, se supone que, para cada región, un CNV sólo está presente en una baja cantidad de muestras. Por lo que es posible que no se detecte un CNV entre el conjunto de muestras

que se están analizando si una o varias de las muestras tienen un CNV en la misma posición.

Por último, los primeros exones de los genes frecuentemente tienen fluctuaciones de cobertura más altas debido a su mayor contenido de GC. Como consecuencia, la detección de CNVs en estas regiones es menor.

El trabajo futuro estará enfocado en desarrollar un protocolo para el análisis de CNV y agregar estas capacidades en la plataforma, para que esta información esté disponible a los profesionales de la salud.

Hacia una mayor eficiencia diagnóstica: incorporación de elementos big data e IA

El desafío general en el contexto de la utilización de NGS para diagnóstico, y del cual los puntos mencionados anteriormente son componentes esenciales, es lograr una mayor eficiencia diagnóstica. Si bien en una primera perspectiva, pareciera que lo que buscamos es mejorar el porcentaje de casos que cuando son analizados resultan en un diagnóstico molecular certero, un segundo punto, muchas veces pasado por alto, consiste en el tiempo requerido para llegar al mismo.

El análisis de un exoma (o genoma) puede llevar horas, y requiere muchas veces revisar una a una, decenas de variantes para determinar cuál de ellas es relevante en el contexto de la historia clínica del paciente. Este proceso se puede hacer más eficiente (reduciendo los tiempos) mediante la incorporación de mejoras al proceso de anotación y a los programas de análisis como la plataforma desarrollada en esta tesis.

Dos mejoras en las que proponemos trabajar en el futuro son la incorporación al proceso de elementos de big data e Inteligencia Artificial (IA). En este contexto el análisis de big data se refiere a la integración y luego síntesis de la información disponible en las bases de datos externas (Uniprot, ClinVar, OMIM, etc.) que hoy en día se encuentran disponibles en la plataforma y que usualmente son consultadas una a una y para cada variante por parte del profesional, para valorar las mismas. La integración y síntesis, buscaría reducir la necesidad de análisis de cada una de las variantes de interés en cada una de las bases de datos, reduciendo de este modo el tiempo de análisis de manera significativa. Esto se podría hacer, mediante una mejor presentación y visualización de los datos clave que el profesional analiza en las bases de datos y mediante el uso de algoritmos de IA que integren y sinteticen los mismos.

Esto nos lleva al segundo punto que proponemos desarrollar a futuro, que consiste en la implementación de algoritmos de IA en el proceso de análisis de variantes. Como describimos en la sección de *Análisis y priorización de variantes* con la plataforma, el análisis de variantes se realiza mediante una serie de filtros y criterios que permiten ir descartando las mismas hasta llegar a una (o unas pocas) variantes candidatas como posible diagnóstico. Estos filtros y criterios pueden embeberse en el contexto de un algoritmo de IA que vaya “aprendiendo” junto con el analista cómo los mismos son

aplicados y luego, una vez entrenado, ayude al analista sugiriendo de manera automática aquellas variantes que para cada caso poseen la mayor chance de ser diagnósticas, reduciendo idealmente el análisis manual a unas pocas variantes claves.

Mejorando los modelos de riesgo poligénico

En el contexto de la utilización de la genómica personal para la prevención, hay que destacar que este área se encuentra mucho menos desarrollada que la de diagnóstico, con una diferencia de aproximadamente 5 años. Tal es así, que hoy todas las aplicaciones se dan, aún, en contexto de investigación, salvo algunas aplicaciones a enfermedades específicas que incluyen, usualmente, sólo unas pocas variantes. Es por ello, que la primera perspectiva en el futuro cercano de la utilización de la genómica para la prevención, consiste en avanzar en su incorporación a la práctica médica. Más allá de las dificultades asociadas a la logística, el costo, y el know-how por parte de los profesionales, una de las principales dificultades de este tipo de aplicaciones, es su aparente baja precisión y su bajo impacto a nivel individual. Si bien, éstas críticas provienen muchas veces de una dificultad para la comprensión de este tipo de evaluaciones probabilísticas, en la actualidad su potencial, debido al nivel de certeza e impacto, es mucho mayor a nivel poblacional que a nivel individual. Para mejorar ambos aspectos, recientemente se ha trabajado mucho en el desarrollo de mejores modelos de riesgo poligénicos, que sean capaces de utilizar un mayor número de marcadores y posean mayor capacidad predictiva y un rango más amplio del potencial impacto evaluado.

Si recordamos la descripción del modelo utilizando en el presente trabajo, se destaca que el mismo se basa en primero estimar el OR*, que es una transformación del OR derivado directamente de GWAS correspondiente (que asocia y valora el alelo de riesgo) en función del genotipo del individuo para el marcador y normalizado respecto de la población de referencia. Luego, los OR* de cada marcador, son combinados en un modelo multiplicativo, asumiendo que son independientes. Estos modelos, entonces, asumen 3 cosas: i) que el efecto de los alelos es multiplicativo en el OR, ii) que la población de referencia utilizada para normalizar es equivalente a aquella donde se realizó el GWAS y que los alelos de referencia y riesgo se encuentran en equilibrio hardy-weinberg y iii) que la contribución de cada marcador es independiente. Estas asunciones, si bien razonables, no son necesariamente correctas y en líneas generales tienden a acotar el rango de los efectos. Los nuevos modelos de riesgo poligénicos se saltean las tres asunciones buscando un modelo integral para todos los marcadores de interés (no asume independencia), que evalúe directamente el genotipo del individuo y que idealmente se encuentre desarrollado y calibrado con una población de referencia que sea equivalente a la cual el mismo luego será aplicado. Es posible que en el futuro cercano veamos un creciente desarrollo de este tipo de modelos para cada una de las enfermedades complejas de alta prevalencia que hemos mencionado. Un punto a destacar, es que para el adecuado desarrollo de este tipo de modelos es necesario contar con una historia clínica detallada y precisa asociada a la información genómica, lo que nos lleva al punto siguiente.

Integración de la genómica con la EHR

Una historia clínica electrónica (EHR por su siglas en inglés, Electronic Health Record) es una versión digital de la historia clínica en papel de un paciente. La EHR son registros que contienen información de salud y de tratamiento del paciente, que están disponibles inmediatamente, en tiempo real y de forma segura para los usuarios autorizados. Una de las características clave de un sistema de EHR es que está construido para ir más allá de los datos clínicos estándar recogidos en el consultorio del médico y puede incluir una visión más amplia de la atención del paciente. La información puede ser creada, gestionada y compartida en un formato digital con otros profesionales o instituciones de salud - tales como laboratorios, especialistas, centros de imágenes médicas, servicios de emergencia y clínicas - ya que contienen información de todos los médicos involucrados en el cuidado del paciente. La EHR puede contener diagnósticos, medicamentos, tratamientos, fechas de vacunación, alergias, imágenes de radiología y laboratorio, resultados de diferentes exámenes *y por supuesto información genómica del individuo.*

Es uno de los desafíos futuros poder colaborar en la incorporación en la historia clínica digital de la información genómica del paciente. Contar con esta información permitiría que pueda ser utilizada por los profesionales de la salud para ayudar a tomar mejores decisiones fundamentadas en la evidencia y coordinadas entre los diferentes profesionales. Antes de (o quizás en paralelo a) incorporar la información genómica en la historia clínica digital, hay que profundizar el grado de adopción de la EHR en el sistema de salud Argentino, ya que en la actualidad es dispar, y cada institución maneja su infraestructura de sistema para la administración y gestión de la información de los pacientes de manera descoordinada y aislada. Además, uno de los puntos claves, es que el paciente no tiene acceso fácil y portable a su propia historia clínica. Cuando se logre que los sistemas de historia clínica estén completamente funcionales e interconectados, los beneficios sobre la historia clínica en papel se incrementarán notablemente: se mejorará la calidad y la comodidad de la atención, se aumentará la participación del paciente en su cuidado, se mejorará la precisión de los diagnósticos y los resultados, entre otros beneficios.

Conclusión

En la presente tesis se planteó como objetivo general el desarrollo y diseño de un marco de referencia tanto tecnológico como conceptual para la gestión, procesamiento y análisis de información genómica humana derivada de experimentos de secuenciación (NGS y arrays), para su aplicación a los tres pilares de la práctica clínica: diagnóstico, tratamiento y prevención. Además se planteó realizar una prueba de concepto de su aplicación en casos reales de cada uno de ellos. Los resultados muestran que fuimos capaces de desarrollar por un lado un protocolo bioinformático que permite de manera eficiente y precisa llevarnos desde los datos obtenidos de un experimento de NGS (lecturas) hasta un archivo de tipo VCF altamente anotado (una suerte de resumen del genoma -o la fracción analizada de él- del individuo del cual se tomó la muestra), que posee un gran contenido de información biológica y clínica. Esta información es la que luego es utilizada por los profesionales para incorporar en la clínica. Para ello, como también se muestra en los resultados, hemos desarrollado una plataforma de software que permite por parte de los profesionales el análisis profundo y detallado de la información genómica en cualquiera de los tres contexto de interés mencionados (diagnóstico, tratamiento, prevención). Como muestran los resultados de las pruebas de concepto los desarrollos mencionados permiten a partir de datos genómicos, por ejemplo un exoma, llegar al diagnóstico molecular preciso en el caso de un paciente que posee un diagnóstico clínico de una enfermedad mendeliana, u obtener un detallado perfil del riesgo relativo de un paciente de desarrollar diversas enfermedades complejas. Más aún, el trabajo realizado con los presentes desarrollos en colaboración con diversos profesionales de la salud, muestra la importancia de contar con este tipo de herramientas que facilitan el acceso, son intuitivas y se encuentran disponibles en el ámbito local.

Anexo I

Publicaciones en revistas especializadas y presentaciones a congresos relacionadas con el trabajo de tesis.

Premios

- Premios Científicos Bernardo A. Houssay 2017 en la categoría ASDIN, otorgado por CEDIQUIFA. TECNOLOGÍAS GENÓMICAS DE NUEVA GENERACIÓN Y MEDICINA DE PRECISIÓN. “Expandiendo las fronteras de la tecnología genómica en Latinoamérica: Medicina de precisión Made in Argentina.” German Biagioli, Jonathan Zaiat, Sebastián A. Vishnopolska, Geronimo Dubra, Guadalupe Buda, Nelba Pérez, María T. Bernardi, Sergio I. Nemirovsky, Juan P. Bustamante, Adrián G. Turjanski y Marcelo A. Marti

Congresos

- 8vo Congreso Argentino de Bioinformática y Biología Computacional (8CA2BC). 26 – 29 de Noviembre de 2017 Posadas, Misiones, Argentina. (Presentación oral).

- “International Congress of Parkinson's Disease and Movement Disorders: Spastic ataxia with Eye-of-the-Tiger Sign due to novel compound heterozygous AFG3L2 mutation.” www.mdscongress.org/Congress-2019.htm. Accessed 15 Dec. 2021.

Publicaciones

- Calandra CR, Mocarbel Y, Vishnopolska SA, Toneguzzo V, Oliveri J, Cazado EC, Biagioli G, Turjanksi AG, Marti M. Gordon Holmes Syndrome Caused by RNF216 Novel Mutation in 2 Argentinean Siblings. Mov Disord Clin Pract. 2019 Jan 16;6(3):259-262. doi: 10.1002/mdc3.12721. PMID: 30949559; PMCID: PMC6417841.

- Calandra CR, Buda G, Vishnopolska SA, Oliveri J, Olivieri FA, Pérez Millán MI, Biagioli G, Miquelini LA, Pellene AL, Marti MA. Spastic ataxia with eye-of-the-tiger-like sign in 4 siblings due to novel compound heterozygous AFG3L2 mutation. Parkinsonism Relat Disord. 2020 Apr;73:52-54. doi: 10.1016/j.parkreldis.2020.03.020. Epub 2020 Mar 24. PMID: 32248051.

- German Biagioli, Jonathan Zaiat, Sebastián A. Vishnopolska, Geronimo Dubra, Guadalupe Buda, Nelba Pérez, María T. Bernardi, Sergio I. Nemirovsky, Juan P. Bustamante, Adrián G. Turjanski y Marcelo A. Marti. “Expandiendo las fronteras de la tecnología genómica en Latinoamérica: Medicina de precisión Made in Argentina.”. Revista Química Viva. Revista Electrónica del Dpto. de Química Biológica, Fac. de Ciencias Exactas y Naturales, Univ. de Buenos Aires, Argentina. QuimicaViva, vol18

num1/2019, www.quimicaviva.qb.fcen.uba.ar/v18n1/E0153.html. Accessed 15 Dec. 2021.

- Revista SAEGRE, Sociedad Argentina de Endocrinología Ginecológica y Reproductiva: Volumen XXVI - Número 1 - Suplemento, Enero - Junio de 2019 – Págs. N° 59- 65. “Análisis genómicos aplicados a la reproducción”- www.saegre.org.ar/revista/numeros/2019/Revista_SAEGRE_1-19_Suplemento.pdf. Accessed 15 Dec. 2021.

- Buda, G., Valdez, R.M., Biagioli, G. et al. Inflammatory cutaneous lesions and pulmonary manifestations in a new patient with autosomal recessive ISG15 deficiency case report. Allergy Asthma Clin Immunol 16, 77 (2020). https://doi.org/10.1186/s13223-020-00473-7

- Mendez R, Iqbal S, Vishnopolska S, Martinez C, Dibner G, Aliano R, Zaiat J, Biagioli G, Fernandez C, Turjanski A, Campbell AJ, Mercado G, Marti MA. Oculocutaneous albinism type 1B associated with a functionally significant tyrosinase gene polymorphism detected with Whole Exome Sequencing. Ophthalmic Genet. 2021 Jun;42(3):291-295. doi: 10.1080/13816810.2021.1888129. Epub 2021 Feb 18. PMID: 33599182.

Bibliografia

[1] WATSON JD, CRICK FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953 Apr 25;171(4356):737-8. doi: 10.1038/171737a0. PMID: 13054692.

[2] Alberts, Bruce. Molecular Biology of the Cell, 5th Edition. Garland Science, 2008.

[3] “The Human Genome Project” 22 Dec. 2020, www.genome.gov/human-genome-project. Accessed 15 Dec. 2021..

[4] “Human Genome Overview” Genome Reference Consortium, 21 July 2021, www.ncbi.nlm.nih.gov/grc/human. Accessed 15 Dec. 2021.

[5]“Acerca del Instituto” www.genome.gov/acerca-del-instituto. Accessed 15 Dec. 2021.

[6] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015 May;17(5):405-24. doi: 10.1038/gim.2015.30. Epub 2015 Mar 5. PMID: 25741868; PMCID: PMC4544753.

[7] Sanger F; Coulson AR (May 1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". J. Mol. Biol. 94 (3): 441–8. doi:10.1016/0022-2836(75)90213-2. PMID 1100841.

[8] Sanger F; Nicklen S; Coulson AR (December 1977). "DNA sequencing with chain-terminating inhibitors". Proc. Natl. Acad. Sci. U.S.A. 74 (12): 5463–7. Bibcode:1977PNAS...74.5463S. doi:10.1073/pnas.74.12.5463. PMC 431765. PMID 271968.

[9] “Microarrays |” Microarray analysis techniques and products, www.illumina.com/techniques/microarrays.html. Accessed 15 Dec. 2021.

[10] “Sequencing Technology |” Sequencing by synthesis, emea.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html. Accessed 15 Dec. 2021.

[11] “National Institutes of Health (NIH)” Turning Discovery Into Health, www.nih.gov/. Accessed 15 Dec. 2021.

[12] “The European Bioinformatics Institute. EMBL-EBI” www.ebi.ac.uk/. Accessed 15 Dec. 2021.

[13] Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008;9:387-402. doi: 10.1146/annurev.genom.9.081307.164359. PMID: 18576944.

[14] Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of next-generation sequencing technologies. Anal Chem. 2011 Jun 15;83(12):4327-41. doi: 10.1021/ac2010857. Epub 2011 May 25. PMID: 21612267; PMCID: PMC3437308.

[15] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. Epub 2010 Jul 19. PMID: 20644199; PMCID: PMC2928508.

[16] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491-8. doi: 10.1038/ng.806. Epub 2011 Apr 10. PMID: 21478889; PMCID: PMC3083463.

[17] Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1110):11.10.1-11.10.33. doi: 10.1002/0471250953.bi1110s43. PMID: 25431634; PMCID: PMC4243306.

[18] Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011 Mar 15;27(6):863-4. doi: 10.1093/bioinformatics/btr026. Epub 2011 Jan 28. PMID: 21278185; PMCID: PMC3051327.

[19] Andrews S. (2010). “Babraham Bioinformatics” FastQC A Quality Control tool for High Throughput sequence data., www.bioinformatics.babraham.ac.uk/projects/fastqc. Accessed 15 Dec. 2021.

[20] “Picard Tools” By Broad Institute, broadinstitute.github.io/picard/. Accessed 15 Dec. 2021.

[21] Heng Li, Richard Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics, Volume 25, Issue 14, 15 July 2009, Pages 1754–1760, https://doi.org/10.1093/bioinformatics/btp324

[22] “Alignment/Map Format Specification” samtools.github.io/hts-specs/SAMv1.pdf. Accessed 15 Dec. 2021.

[23] “ Sequence Alignment/Map Optional Fields Specification” samtools.github.io/hts-specs/SAMtags.pdf. Accessed 15 Dec. 2021..

[24] “Picard Metrics Definitions.CollectQualityYieldMetrics” broadinstitute.github.io/picard/picard-metric-definitions.html#CollectQualityYieldMetrics. Accessed 15 Dec. 2021.

[25] “Picard Metrics Definitions.CollectHsMetrics” broadinstitute.github.io/picard/command-line-overview.html#CollectHsMetrics. Accessed 15 Dec. 2021.

[26] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8. PMID: 19505943; PMCID: PMC2723002.

[27] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841-2. doi: 10.1093/bioinformatics/btq033. Epub 2010 Jan 28. PMID: 20110278; PMCID: PMC2832824.

[28] Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF; WGS500 Consortium, Wilkie AOM, McVean G, Lunter G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014 Aug;46(8):912-918. doi: 10.1038/ng.3036. Epub 2014 Jul 13. PMID: 25017105; PMCID: PMC4753679.

[29] Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 [q-bio.GN] 2012. github.com/ekg/freebayes. Accessed 15 Dec. 2021.

[30] Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012 Apr-Jun;6(2):80-92. doi: 10.4161/fly.19695. PMID: 22728672; PMCID: PMC3679285.

[31] Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012 Mar 15;3:35. doi: 10.3389/fgene.2012.00035. PMID: 22435069; PMCID: PMC3304048.

[32] Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). Hum Mutat. 2000;15(1):57-61. doi: 10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G. PMID: 10612823.

[33] Daniel R. Zerbino, Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R. Laird, Ilias Lavidas, Zhicheng Liu, Jane E. Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N. Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E. Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J. Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M. Staines, Stephen J. Trevanion, Bronwen L. Aken, Fiona Cunningham, Andrew Yates, Paul Flicek, ENSEMBL 2018.,PubMed PMID: 29155950. doi:10.1093/nar/gkx1098 . www.ensembl.org/index.html. Accessed 15 Dec. 2021.

[34] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506-D515. doi: 10.1093/nar/gky1049. PMID: 30395287; PMCID: PMC6323992.

[35] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11. doi: 10.1093/nar/29.1.308. PMID: 11125122; PMCID: PMC29783.

[36] Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA,

Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG; Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285-91. doi: 10.1038/nature19057. PMID: 27535533; PMCID: PMC5018207.

[37] 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68-74. doi: 10.1038/nature15393. PMID: 26432245; PMCID: PMC4750478.

[38]Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott DR. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016 Jan 4;44(D1):D862-8. doi: 10.1093/nar/gkv1222. Epub 2015 Nov 17. PMID: 26582918; PMCID: PMC4702865.

[39] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan;Chapter 7:Unit7.20. doi: 10.1002/0471142905.hg0720s76. PMID: 23315928; PMCID: PMC4480630.

[40] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81. doi: 10.1038/nprot.2009.86. Epub 2009 Jun 25. PMID: 19561590.

[41] Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010 Aug;7(8):575-6. doi: 10.1038/nmeth0810-575. PMID: 20676075.

[42] Siepel et al., Genome Res, 2005, for phyloFit, cite Siepel and Haussler, Mol Biol Evol, 2004, for exoniphy cite Siepel and Haussler, RECOMB, 2004, for phyloP cite Pollard et al, Genome Res, 2010, and for dless, cite Siepel, Pollard, and Haussler, RECOMB 2006.

[43] Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010 Jan;20(1):110-21. doi: 10.1101/gr.097857.109. Epub 2009 Oct 26. PMID: 19858363; PMCID: PMC2798823.

[44] Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010 Dec 2;6(12):e1001025. doi: 10.1371/journal.pcbi.1001025. PMID: 21152010; PMCID: PMC2996323.

[45] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep;38(16):e164. doi: 10.1093/nar/gkq603. Epub 2010 Jul 3. PMID: 20601685; PMCID: PMC2938201.

[46] Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. Am J Hum Genet. 2017 Feb 2;100(2):267-280. doi: 10.1016/j.ajhg.2017.01.004. Epub 2017 Jan 26. PMID: 28132688; PMCID: PMC5294755.

[47] Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, Yates B, Bruford E. Genenames.org: the HGNC and VGNC resources in 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D786-D792. doi: 10.1093/nar/gky930. PMID: 30304474; PMCID: PMC6324057.

[48] “National Center for Biotechnology Information” www.ncbi.nlm.nih.gov/. Accessed 15 Dec. 2021.

[49] McEntyre, Jo. “The NCBI Handbook” NCBI Bookshelf, www.ncbi.nlm.nih.gov/books/NBK21101. Accessed 15 Dec. 2021.

[50] “University of California, Santa Cruz” 6 Dec. 2021, www.ucsc.edu. Accessed 15 Dec. 2021.

[51] “Home Wellcome Sanger Institute” 14 Dec. 2015, www.sanger.ac.uk/. Accessed 15 Dec. 2021.

[52] Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, Hamamsy T, Lek M, Samocha KE, Cummings BB, Birnbaum D; The Exome Aggregation Consortium, Daly MJ, MacArthur DG. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2017 Jan 4;45(D1):D840-D845. doi: 10.1093/nar/gkw971. Epub 2016 Nov 28. PMID: 27899611; PMCID: PMC5210650.

[53] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME; Genome Aggregation Database Consortium, Neale BM, Daly MJ, MacArthur DG. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434-443. doi: 10.1038/s41586-020-2308-7. Epub 2020 May 27. Erratum in: Nature. 2021 Feb;590(7846):E53. PMID: 32461654; PMCID: PMC7334197.

[54] Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, Danis D, Balagura G, Baynam G, Brower AM, Callahan TJ, Chute CG, Est JL, Galer PD,

Ganesan S, Griese M, Haimel M, Pazmandi J, Hanauer M, Harris NL, Hartnett MJ, Hastreiter M, Hauck F, He Y, Jeske T, Kearney H, Kindle G, Klein C, Knoflach K, Krause R, Lagorce D, McMurry JA, Miller JA, Munoz-Torres MC, Peters RL, Rapp CK, Rath AM, Rind SA, Rosenberg AZ, Segal MM, Seidel MG, Smedley D, Talmy T, Thomas Y, Wiafe SA, Xian J, Yüksel Z, Helbig I, Mungall CJ, Haendel MA, Robinson PN. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D1207-D1217. doi: 10.1093/nar/gkaa1043. PMID: 33264411; PMCID: PMC7778952. (HPO: https://hpo.jax.org/)

[55] “Orphanet” 18 Oct. 2022, www.orpha.net/consor/cgi-bin/index.php. Accessed 15 Dec. 2021.

[56] “DECIPHER v11.9: Mapping the clinical genome” decipher.sanger.ac.uk. Accessed 15 Dec. 2021.

[57] Peng, J., Hui, W. & Shang, X. Measuring phenotype-phenotype similarity through the interactome. BMC Bioinformatics 19, 114 (2018). https://doi.org/10.1186/s12859-018-2102-9

[58] Peng, J., Li, Q. & Shang, X. Investigations on factors influencing HPO-based semantic similarity calculation. J Biomed Semant 8, 34 (2017). https://doi.org/10.1186/s13326-017-0144-y

[59] Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, Mundlos C, Horn D, Mundlos S, Robinson PN. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009 Oct;85(4):457-64. doi: 10.1016/j.ajhg.2009.09.003. PMID: 19800049; PMCID: PMC2756558.

[60] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousgou O, Whetzel PL, Amode R, Guillen JA, Riat HS, Trevanion SJ, Hall P, Junkins H, Flicek P, Burdett T, Hindorff LA, Cunningham F and Parkinson H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research, 2019, Vol. 47 (Database issue): D1005-D1012

[61] “The Clinical and Functional TRanslation of CFTR (CFTR2)” CFTR2, 24 Sept. 2021, cftr2.org. Accessed 15 Dec. 2021.

[62] Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012 Oct;92(4):414-7. doi: 10.1038/clpt.2012.96. PMID: 22992668; PMCID: PMC3660037.

[63] Relling MV, Klein TE. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. Clin Pharmacol Ther. 2011 Mar;89(3):464-7. doi: 10.1038/clpt.2010.279. Epub 2011 Jan 26. PMID: 21270786; PMCID: PMC3098762.

[64] Ku CS, Cooper DN, Polychronakos C, Naidoo N, Wu M, Soong R. Exome sequencing: dual role as a discovery and diagnostic tool. Ann Neurol. 2012 Jan;71(1):5-14. doi: 10.1002/ana.22647. PMID: 22275248.

[65] Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. J Pathol Inform. 2012;3:40. doi: 10.4103/2153-3539.103013. Epub 2012 Oct 31. PMID: 23248761; PMCID: PMC3519097.

[66] Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 2010 Apr 15;26(8):1112-8. doi: 10.1093/bioinformatics/btq099. Epub 2010 Mar 3. PMID: 20200009; PMCID: PMC2853691.

[67] Revista QuímicaViva. Número 1, año 18, Abril 2019
http://www.quimicaviva.qb.fcen.uba.ar/v18n1/E0153.html

[68] Rodriguez-Granillo A, Sedlak E, Wittung-Stafshede P. Stability and ATP binding of the nucleotide-binding domain of the Wilson disease protein: effect of the common H1069Q mutation. J Mol Biol. 2008 Nov 28;383(5):1097-111. doi: 10.1016/j.jmb.2008.07.065. Epub 2008 Jul 29. PMID: 18692069.

[69] Firneisz G, Lakatos PL, Szalay F, Polli C, Glant TT, Ferenci P. Common mutations of ATP7B in Wilson disease patients from Hungary. Am J Med Genet. 2002 Feb 15;108(1):23-8. doi: 10.1002/ajmg.10220. PMID: 11857545.

[70] Thomas GR, Forbes JR, Roberts EA, Walshe JM, Cox DW. The Wilson disease gene: spectrum of mutations and their consequences. Nat Genet. 1995 Feb;9(2):210-7. doi: 10.1038/ng0295-210. Erratum in: Nat Genet 1995 Apr;9(4):451. PMID: 7626145.

[71] Czlonkowska A, Rodo M, Gajda J, Ploos van Amstel HK, Juyn J, Houwen RH. Very high frequency of the His1069Gln mutation in Polish Wilson disease patients. J Neurol. 1997 Sep;244(9):591-2. doi: 10.1007/s004150050149. PMID: 9352458.

[72] Abdelghaffar TY, Elsayed SM, Elsobky E, Bochow B, Büttner J, Schmidt H. Mutational analysis of ATP7B gene in Egyptian children with Wilson disease: 12 novel mutations. J Hum Genet. 2008;53(8):681. doi: 10.1007/s10038-008-0298-7. Epub 2008 May 16. PMID: 18483695.

[73] “MedlinePlus: Genetics” 11 Aug. 2021, ghr.nlm.nih.gov/. Accessed 15 Dec. 2021.

[74] Chan W, Schaffer TB, Pomerantz JL. A quantitative signaling screen identifies CARD11 mutations in the CARD and LATCH domains that induce Bcl10 ubiquitination and human lymphoma cell survival. Mol Cell Biol. 2013 Jan;33(2):429-43. doi: 10.1128/MCB.00850-12. Epub 2012 Nov 12. PMID: 23149938; PMCID: PMC3554118.

[75] Brohl AS, Stinson JR, Su HC, Badgett T, Jennings CD, Sukumar G, Sindiri S, Wang W, Kardava L, Moir S, Dalgard CL, Moscow JA, Khan J, Snow AL. Germline CARD11 Mutation in a Patient with Severe Congenital B Cell Lymphocytosis. J Clin Immunol. 2015 Jan;35(1):32-46. doi: 10.1007/s10875-014-0106-4. Epub 2014 Oct 29. PMID: 25352053; PMCID: PMC4466218.

[76] Yüce Ö, West SC. Senataxin, defective in the neurodegenerative disorder ataxia with oculomotor apraxia 2, lies at the interface of transcription and the DNA damage response. Mol Cell Biol. 2013 Jan;33(2):406-17. doi: 10.1128/MCB.01195-12. Epub 2012 Nov 12. PMID: 23149945; PMCID: PMC3554130.

[77] Ichikawa Y, Ishiura H, Mitsui J, Takahashi Y, Kobayashi S, Takuma H, Kanazawa I, Doi K, Yoshimura J, Morishita S, Goto J, Tsuji S. Exome analysis reveals a Japanese family with spinocerebellar ataxia, autosomal recessive 1. J Neurol Sci. 2013 Aug 15;331(1-2):158-60. doi: 10.1016/j.jns.2013.05.018. Epub 2013 Jun 18. PMID: 23786967.

[78] Tatton-Brown K, Cole TRP, Rahman N. Sotos Syndrome. 2004 Dec 17 [updated 2019 Aug 1]. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mirzaa G, Amemiya A, editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993–2021. PMID: 20301652.

[79] Vandana Shashi MD, Allyn McConkie-Rosell PhD, Bruce Rosell BS, Kelly Schoch MS, Kasturi Vellore MD, Marie McDonald MD, Yong-Hui Jiang MD, PhD, Pingxing Xie PhD, Anna Need PhD & David B Goldstein PhD. The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. Genet Med 15, 849 (2013). https://doi.org/10.1038/gim.2013.147

[80] Buske OJ, Girdea M, Dumitriu S, Gallinger B, Hartley T, Trang H, Misyura A, Friedman T, Beaulieu C, Bone WP, Links AE, Washington NL, Haendel MA, Robinson PN, Boerkoel CF, Adams D, Gahl WA, Boycott KM, Brudno M. PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. Hum Mutat. 2015 Oct;36(10):931-40. doi: 10.1002/humu.22851. Epub 2015 Aug 31. PMID: 26251998; PMCID: PMC5467641.

[81] Ma CA, Stinson JR, Zhang Y, Abbott JK, Weinreich MA, Hauk PJ, Reynolds PR, Lyons JJ, Nelson CG, Ruffo E, Dorjbal B, Glauzy S, Yamakawa N, Arjunaraja S, Voss K, Stoddard J, Niemela J, Zhang Y, Rosenzweig SD, McElwee JJ, DiMaggio T, Matthews HF, Jones N, Stone KD, Palma A, Oleastro M, Prieto E, Bernasconi AR, Dubra G, Danielian S, Zaiat J, Marti MA, Kim B, Cooper MA, Romberg N, Meffre E, Gelfand EW, Snow AL, Milner JD. Germline hypomorphic CARD11 mutations in severe atopic disease. Nat Genet. 2017 Aug;49(8):1192-1201. doi: 10.1038/ng.3898. Epub 2017 Jun 19. Erratum in: Nat Genet. 2017 Oct 27;49(11):1661. PMID: 28628108; PMCID: PMC5664152.

[82] Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM, Sherry ST, Watson MS; ClinGen. ClinGen--the Clinical Genome Resource. N Engl J Med. 2015 Jun

4;372(23):2235-42. doi: 10.1056/NEJMsr1406261. Epub 2015 May 27. PMID: 26014595; PMCID: PMC4474187.

[83] “Broad Institute” www.broadinstitute.org/. Accessed 15 Dec. 2021.

[84] McCarthy, J.,Nussbaum, R. “Genomic and Precision Medicine”.[MOOC]. Coursera. www.my-mooc.com/en/mooc/genomicmedicine/. Accessed 15 Dec. 2021.

[85] Young, A., Gillung, J.P.. Phylogenomics – principles, opportunities and pitfalls of big-data phylogenetics. onlinelibrary.wiley.com/doi/full/10.1111/syen.12406. Accessed 15 Dec. 2021.

[86] “Starting off in Bioinformatics RNA Transcription and Translation” Towards Data Science, 18 Aug. 2017, towardsdatascience.com/starting-off-in-bioinformatics-rna-transcription-and-translation-aaa7a91db031. Accessed 15 Dec. 2021.