Segmentos genómicos duplicados y Pseudogenes

¿Cuáles son las Dificultades para validar, interpretar y confirmar los diagnósticos asociados a genes que poseen estas características?

Cuando se utilizan tecnologías de secuenciación masiva para un diagnóstico molecular es importante recordar que las regiones de interés (Genoma, Exoma o Panel de Genes) son secuenciadas mediante la utilización de lecturas "cortas" (100 -300 pb) que luego son mapeadas y alienadas contra el genoma de referencia, para determinar las variantes presentes en la muestra.

Algunos genes, sin embargo, presentan una dificultad particular para su secuenciación y análisis, debido a la presencia en algún "otro" lugar del genoma de una segunda versión del gen -generalmente no funcional (un pseudogen)- o de un segmento del genoma que lo contiene duplicado. La presencia de estas regiones homólogas, dificulta tanto el proceso de secuenciación cómo de análisis posterior, resultando en una limitación para detectar y validar las variantes presentes en los genes asociados.

¿Qué es un pseudogen?

Un pseudogen es una región genómica que tiene una alta similitud de secuencia (es homóloga) con un gen conocido. En general, no es funcional y se ha inactivado en el curso de la evolución al acumular mutaciones deletéreas en su secuencia. Sin embargo, existen otros pseudogenes que cambiaron radicalmente su función original y se mantienen activos.

Por lo general, la identidad de secuencia entre un pseudogen y su gen funcional varía aproximadamente entre el 65 y el 99%.

¿Qué es un segmento genómico duplicado?

Un segmento genómico duplicado (Segmental Duplication, SD) es una región cuyo tamaño varía entre 1 y 200 kb, que aparece duplicada en una (o más) ubicaciones del genoma, pudiendo contener genes y pseudogenes. Se estima que estos segmentos duplicados cubren aproximadamente un 3,6% de la secuencia genómica completa. Actualmente, los SD están definidos como una región en el genoma donde la secuencia está duplicada y la similitud entre la región original y la duplicada es mayor o igual al 90% en una longitud mayor de 1 kilobases (≥1000 pares de bases).

¿Por qué es importante tener en cuenta los pseudogenes (y SD) en el análisis de muestras derivadas para la determinación de un diagnóstico molecular?

La presencia de pseudogenes afecta significativamente el mapeo y el alineamiento de las lecturas, y por lo tanto, el posterior llamado de variantes, debido a que:

Los SD pueden ser indistinguibles entre sí dependiendo del porcentaje de homología y del tamaño de las lecturas que se estén utilizando.

Los altos niveles de similitud de secuencia complican el mapeo y alineamiento preciso de las lecturas. Como se muestra en la Figura 1, durante el proceso de mapeo, las lecturas pueden mapear en más de un lugar y por ello se les asigna un Map Quality bajo. En general, los llamadores de variantes aplican un filtro para descartar las lecturas con Map Quality bajo. Por esta razón, en aquellos genes que presentan pseudogenes y/o SD se subestima la cobertura y consecuentemente el número de variantes (falsos negativos).

Segmentos genómicos duplicados y Pseudogenes

El mapeo y alineamiento erróneo de lecturas derivadas del pseudogen sobre el gen funcional, puede dar lugar al llamado de una variante falsa (falso positivo).
Los genes que tienen pseudogenes son proclives a que haya recombinación homóloga y aparezcan mutaciones en el gen funcional que en realidad corresponden a secuencias de tipo wild type en el pseudogen (de ahí que muchos genes con pseudogenes tengan hot-spots mutacionales).
Debido al alto grado de similitud de secuencia, puede ser extremadamente difícil diseñar primers para la secuenciación del gen (y no del pseudogen) por la técnica de Sanger, imposibilitando muchas veces la correcta verificación de la presencia de una (o más) variante(s).

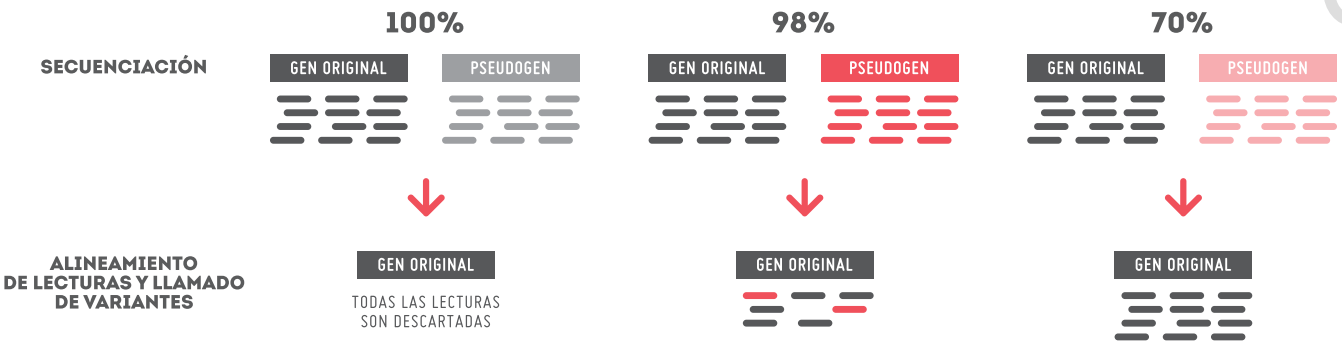


Figura 1: La confianza en el alineamiento de las lecturas disminuye cuando aumenta la homología de secuencia entre regiones. Las lecturas se descartan cuando se alinean igualmente bien a varias posiciones genómicas y tienen una calidad de mapeo igual a 0 (MQ=0).

¿Todos los genes con pseudogenes asociados son difíciles de secuenciar y analizar?

El grado en el que la existencia de un pseudogen afecta la capacidad de detectar y mapear con precisión variantes en su gen funcional depende principalmente del grado de similitud entre la región duplicada y el gen funcional.
En general, las variantes en genes que comparten entre un 90% y hasta un 98% de homología con un pseudogen todavía se detectan y mapean con precisión. Cuando la homología es superior al 98%, la detección precisa y el mapeo de variantes se vuelve más difícil.

¿Cuántos genes clínicamente relevantes se ven afectados por pseudogenes y/o SDs?

Segun el proyecto GENCODE, se estima que los humanos tienen más de 10,000 pseudogenes. Es por esta razón que algunos de los genes contenidos en paneles con alta demanda y contemplados en la secuenciación del exoma completo tienen pseudogenes u otras regiones homólogas en el genoma.

El llamado de variantes de estas regiones puede, debido a los problemas enumerados anteriormente, poseer menor calidad.

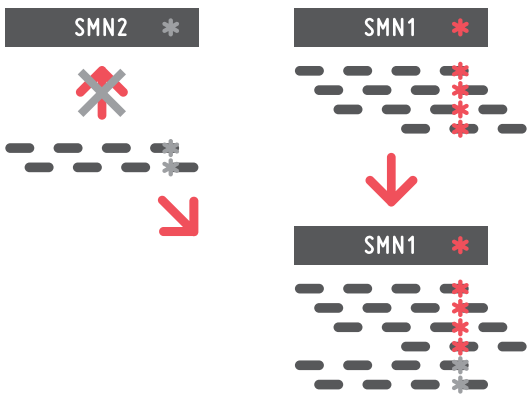
¿Qué hacemos en Bitgenia para mejorar la capacidad de detectar con precisión variantes en genes clínicamente relevantes con pseudogenes y/o SD?

Para hacer frente a las limitaciones mencionadas, cuando un profesional solicita el análisis de un gen con estas características, realizamos un análisis bioinformático de mapeo y alineamiento diferencial, supervisado por un análisis de posiciones y variantes características que maximiza la calidad del llamado de variantes y análisis subsiguiente en el gen de interés.

Segmentos genómicos duplicados y Pseudogenes

Brevemente, tomando como punto de partida el conocimiento de las posiciones genómicas de los segmentos duplicados, enmascaramos la referencia en la región correspondiente al pseudogen y hacemos que el alineador considere sólo nuestras regiones de interés. En consecuencia, podemos identificar las lecturas del pseudogen como aquellas que poseen "variantes" que corresponden a la secuencia salvaje del mismo, siendo las otras las correspondientes al gen funcional de interés.

A modo de ejemplo, tomemos el gen SMN1, solicitado en el estudio de pacientes con AME (Atrofia muscular espinal), que tiene más de un 98% de identidad de secuencia con el pseudogen SMN2, afectando la totalidad de los exones. Los genes SMN1 y el SMN2 sólo difieren en cinco nucleótidos: c.835-44G>A (rs1454173648); c.840C>T (rs1164325688); c.888+100A>G (rs212214); c.888+215A>G (rs1244569826); c.1155G>A (rs1208416968). Si se alinean todas las lecturas a SMN1, "aparecen" estas variantes esperadas y benignas, en las lecturas del pseudogen La ausencia de otras variantes, indica que el gen funcional NO presenta mutaciones, mientras que las variantes del mismo deben observarse "sólo" en aquellas lecturas que no participan en el llamado de las variantes del pseudogen.



LECTURAS DE SMN1 Y SMN2 JUNTAS

Genes afectados por segmentos genómicos duplicados
Las coordenadas genómicas de las regiones afectadas se enumeran en la siguiente tabla, donde se detallan los exones de los genes que se ven afectados por una homología >90% (ó >98%), basada en datos de duplicaciones segmentarias extraídos de la base de datos del navegador del genoma UCSC.

Link a la página con la lista de genes.

En nuestro sitio web (www.bitgenia.com) destacamos en negrita los genes afectados de nuestros paneles aclarando la existencia de esta limitación, con la finalidad de garantizar que los profesionales solicitantes estén adecuadamente informados.

En Bitgenia creemos que es importante que nuestros clientes estén al tanto de las limitaciones asociadas con las tecnologías de secuenciación y las mismas puedan ser contempladas al momento de solicitar estudios que involucren a genes con secuencias repetitivas o con exones que presenten una alta homología con otras regiones del genoma. Por otro lado, constantemente tomamos medidas y nos servimos de recursos informáticos para mantenernos actualizados y hacer frente a las limitaciones mencionadas, buscando alternativas que nos permitan minimizar las dificultades que las mismas nos puedan llegar a presentar durante el procesamiento, llamado de variantes y análisis de los resultados.