Cómo Encontrar la Causa del Cáncer: La Aguja entre 3000 Millones de Datos

Andrea Sabina Llera[1], Juan Martín Sendoya[1], Gabriela Merino[2], Osvaldo Podhajcer[1], Elmer Fernández[2]

[1]Laboratorio GENOCAN-LTMC, Fundación Instituto Leloir, Avenida Patricias Argentinas 435, C1405BWE Buenos Aires, Argentina
[2]Grupo de Minería de Datos en Biociencias, Universidad Católica de Córdoba, Avenida Armada Argentina 3555, X5016DHK Córdoba, Argentina

Abstract. En la actualidad es posible obtener y "leer" la secuencia de bases presentes en cualquier molécula de ADN y así obtener la información hereditaria contenida en cada organismo. Este procedimiento se denomina "secuenciación". Actualmente, las tecnologías disponibles (*Next Generation Sequencing* o NGS) permiten evaluar millones de secuencias a la vez, en un período corto de tiempo. En esta presentación se resume el flujo típico de trabajo de análisis de variantes durante una secuenciación NGS. Se muestran las características que tiene cada etapa, las herramientas informáticas que se requieren y las dificultades a las que el usuario se enfrenta actualmente.

Keywords: Cáncer · Grandes Datos · NGS

1 Introducción

Todos los organismos de la Tierra, sin ninguna excepción conocida, guardan su información hereditaria en forma de moléculas de ADN de doble cadena. Cada una de estas está formada por cuatro tipos de moléculas llamadas nucleótidos o bases [Adenosina (A), Citosina (C), Guanina (G) y Timina (T)] los cuales están unidos entre sí formando una larga secuencia lineal. En el caso del genoma humano esta información ocupa aproximadamente 3 giga-pares de bases, es decir, 3000 millones de esas "letras" para un genoma haploide. Se puede decir que la información genética es digital porque sólo la combinación lineal entre esas cuatro letras alcanza para codificar toda la información funcional de las células de un organismo. Si bien todas las células de un organismo tienen la misma información genética, distintos eventos fisiológicos y ambientales determinan que éstas se diferencien, generando de este modo los distintos órganos. Pero en ciertas condiciones pueden formarse estructuras anormales que denominamos tumores. Los tumores acumulan cambios en la secuencia de ADN (información genética) que dan por resultado un crecimiento descontrolado de las células.

Para poder estudiar la información genética que dio origen al tumor, es decir qué cambios en las bases del ADN lo generaron, es importante entender cuánto cambia o cuánto se conserva el ADN entre individuos, y entre las células que lo conforman.

Dentro de cada especie esta información es muy conservada porque es ella quien le da las características a dicha especie; sin embargo, esta información no es 100% idéntica en todos los individuos de una especie determinada. Cada individuo posee ciertos cambios en las bases de su ADN (variantes genómicas) que lo hacen único. Cada uno de estos cambios puede provocar un efecto positivo (mejor adaptación al ambiente), neutro, o negativo (predisposición a enfermedades de origen genético) según la posición y el tipo de cambio, y por ello es importante identificarlos y poder distinguirlos. El proceso de determinación de las variantes encontradas en una dada secuencia genómica se denomina "análisis de variantes". La información genética se hereda de los progenitores, una copia de cromosomas de la madre y otra del padre, y junto a ellos sus variantes. Pero además, a lo largo de la vida, distintas células pueden sufrir cambios en su ADN, es decir, adquirir más variantes. En el caso de la especie humana, aproximadamente un 0.1-0.3% del genoma varía entre individuos. Esto indica que, *a priori*, cada individuo podría contener 3 millones de bases variantes en su genoma normal. Si este individuo, además, genera un tumor, su genoma acumulará nuevas mutaciones que generarán cambios en la información genética, pero sólo unos pocos de estos cambios son decisivos para la enfermedad. A modo de ejemplo, se puede citar un reciente estudio en 506 tumores de mama, en el que el análisis de los 506 genomas completos del tumor y de células normales de los mismos individuos dio como resultado únicamente 93 cambios de variantes importantes para el desarrollo de los tumores, la mayoría acumulados en sólo 10 genes. La definición de cuáles son los 93 cambios relevantes es el resultado final de un ingente trabajo que requiere el análisis de los 3000 millones de bases en los 506 genomas. [1].

En la actualidad es posible obtener y "leer" la secuencia de bases presentes en cualquier molécula de ADN y así obtener la información hereditaria contenida en cada organismo. Este procedimiento se denomina "secuenciación". Actualmente, las tecnologías disponibles (*Next Generation Sequencing* o NGS) permiten evaluar millones de secuencias a la vez. Estas tecnologías producen grandes volúmenes de datos en un período de tiempo relativamente corto [2].

2 Análisis de variantes de NGS: flujo estándar

Durante el análisis de variantes existen 3 etapas bien definidas: 1) el control de la calidad de la secuenciación (para distinguir el dato bueno del espurio), 2) la asignación y 3) la interpretación de variantes. Si bien existen herramientas que analizan cada paso del proceso por separado, el rendimiento de las mismas no es equivalente para distintas tecnologías de secuenciación ni para distintos problemas biológicos. Es por eso que se requieren flujos de trabajo modulares, adaptados a cada centro de análisis y a cada proyecto de investigación o servicio realizado [3].

2.1 Control de calidad de la secuenciación

Para la etapa de control de calidad, si bien existen herramientas como TarSeqQC, no son lo suficientemente versátiles para ambientes con la dinámica de un laboratorio

de servicios. Es necesario contar con herramientas amigables que requieran capacidades mínimas y a su vez permitan una rápida visualización de los parámetros de calidad. En este sentido, es importante desarrollar herramientas de visualización rápida para evaluar performance de la secuenciación, a nivel muestra, a nivel experimento y también como carta de control entre experimentos. Este paso debe permitirnos tomar decisiones rápidas como por ejemplo si una muestra o un experimento debe repetirse, si un set de reactivos falla, etc.

2.2 Asignación de variantes (*variant calling*)

Para la etapa de asignación de variantes o *variant calling* existen múltiples opciones de software, donde cada uno posee sus propios supuestos (por ej. calidad de la secuenciación por base, etc.), que impactan en el resultado (variantes asignadas a cada muestra) y su consecuente interpretación. Se debe considerar que la asignación debe distinguir una variante verdadera de un error técnico, el cual tendrá mayor o menor probabilidad de ocurrir según sea el tipo de variante y la tecnología de secuenciación utilizada. La utilización de varias herramientas de asignación de variantes en paralelo podría mejorar la calidad de los resultados, sin embargo cada uno de ellos trabaja con formatos distintos de datos que dificultan su integración y su comunicación con repositorios de información. Es por ello que el desarrollo de herramientas que permitan una integración transparente para el usuario final sería de gran utilidad en este escenario.

2.3 Interpretación de variantes

Una vez identificadas las variantes que presentan grados de calidad aceptables, es necesario interpretar las mismas dentro del contexto de la patología a evaluar. Dicha interpretación requiere de una anotación o nomenclatura estandarizada de las variantes y de la búsqueda y consulta en diferentes repositorios de información biológica contextual [frecuencia de aparición de la variante en la población general (HapMap, 1000 Genomes, etc), frecuencia de aparición en determinadas patologías (HGMD, ClinVar, COSMIC, etc), disponibilidad de drogas dirigidas, etc.]. En la actualidad, una gran parte de este trabajo se realiza manualmente. Dependiendo del tipo de análisis (e.g si se analiza un genoma completo o un *subset* de genes ya asociado con una determinada enfermedad), este proceso de interpretación, que combina en mayor o menor medida herramientas automáticas con trabajo manual, puede llevar a un analista de 6 horas a 20 días para cada muestra. Si en un laboratorio de diagnóstico molecular genómico se procesan habitualmente entre 5 y 30 muestras por mes, es apreciable la envergadura del trabajo y la necesidad imperiosa de generar herramientas automatizadas, amén de contar como poder computacional suficientemente alto como para procesar ese volumen de información.

3 Observaciones finales

En esta presentación se ha sintetizado el flujo típico de trabajo de análisis de variantes durante la secuenciación NGS. Habiendo descripto las características de cada etapa, resulta evidente la importancia crucial de las herramientas de grandes datos a la hora del análisis y entendimiento de los datos recolectados en un laboratorio de genómica aplicada a la salud humana.

Referencias

1. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47-54 (2016)
2. Eisenstein, M. Big data: The power of petabytes. Nature 527, S2–S4 (2015)
3. Ding, L., Wendl, M., McMichael, J. & Raphael, B. Expanding the computational toolbox for mining cancer genomes. Nature reviews. Genetics 15, 556–70 (2014)