Análisis de datos NGS: una revisión de las principales herramientas y marcos de trabajo para el descubrimiento de variantes

Ángel Canal-Alonso1, Noelia Egido1, Pedro Jiménez1, Javier Prieto[1], Juan Manuel Corchado[1]

[1]Departamento de Bioinformática y Biología Computacional, Instituto AIR, Carbajosa de la Sagrada, España

Correo electrónico: acanal@air-institute.com

Abstracto: El análisis de datos genéticos siempre ha sido un problema debido a la gran cantidad de información disponible y la dificultad para aislar la que es relevante. Sin embargo, a lo largo de los años los avances en las técnicas de secuenciación han ido acompañados de un desarrollo de técnicas informáticas hasta la aplicación actual de la inteligencia artificial. Podemos resumir las fases del análisis de secuencia en lo siguiente: evaluación de calidad, alineación, procesamiento previo a la variante, llamada de variante y anotación de variante. En este artículo revisaremos y comentaremos las herramientas utilizadas en cada fase de la secuenciación genética, y analizaremos los inconvenientes y ventajas que ofrece cada una de ellas.

Palabras clave: Aprendizaje automático; Bioinformática; Secuenciación de próxima generación, canalización

---

1. Introducción

En los últimos años hemos experimentado una evolución sin precedentes en cuanto a la secuenciación del ADN, hasta el punto de que se ha convertido en una de las principales herramientas en múltiples áreas de la ciencia biomédica. Todo comenzó con el descubrimiento de la estructura del ácido desoxirribonucleico -ADN- por Watson y Crick [1], el material genético a partir del cual se desarrolla todo organismo vivo, formado por una doble hélice con dos cadenas poliméricas complementarias entre sí. Cada una de estas hebras está formada por la unión de nucleótidos, moléculas orgánicas compuestas por un carbohidrato, un grupo fosfato y una base nitrogenada; estos últimos son compuestos cíclicos cuya identidad -adenina (A), timina (T), citosina (C) o guanina (G)- determinará el tipo de nucleótido, y por tanto, cada una de las cadenas que componen el ADN se puede especificar como una secuencia de estas cuatro letras. De este modo, El proceso de secuenciación del ADN consiste en determinar el orden preciso de estos cuatro nucleótidos a lo largo de una molécula de ADN. La complementariedad que hemos comentado anteriormente se debe a que estas dos cadenas están unidas naturalmente entre sí mediante enlaces químicos específicos entre sus nucleótidos, en la forma AT y CG -debido principalmente a sus propiedades químicas-, por lo que podríamos decir que si Conociendo la secuencia de una de las cadenas podremos averiguar el orden secuencial de la otra debido a la complementariedad de sus bases nitrogenadas.

Paralelamente, Sanger [2] y Maxam y Gilbert [3] lograron desarrollar diferentes métodos para secuenciar moléculas de ADN a finales de los años 1970, aunque la técnica de secuenciación de Sanger o método dideoxi se convirtió en la herramienta predominante durante los años siguientes hasta la aparición de los siguientes. Métodos de secuenciación de generación. Esta técnica se basa en la propia reacción de polimerización del ADN, proceso conocido como replicación que ocurre en todo tipo de organismos biológicos, en el que se sintetiza una cadena de ADN utilizando como plantilla su cadena complementaria; en este proceso, los nucleótidos utilizados llevan consigo una molécula fluorescente diferente según su base nitrogenada, además de una pequeña modificación química que provoca que, cuando se une a la cadena en síntesis, ya no tenga capacidad para unirse a la siguiente. nucleótido, deteniendo el proceso en este punto. En consecuencia, nuestra mezcla de reacción contendrá multitud de moléculas de ADN que difieren en longitud, ya que este tipo de nucleótidos modificados se unen a la cadena por casualidad. Posteriormente se utilizará una técnica de biología molecular conocida como electroforesis capilar, en la que los fragmentos de ADN se separan por tamaño y pasan por un detector que recoge la señal luminosa que emite el último nucleótido de cada fragmento, conociendo así el orden secuencial de las bases. según su paso por el detector. Al final obtendremos como resultado el conjunto de señales recogidas por el detector ordenadas desde el fragmento más pequeño hasta la molécula completa. Se utilizará una técnica de biología molecular conocida como electroforesis capilar, en la que los fragmentos de ADN se separan por tamaño y pasan por un detector que recoge la señal luminosa emitida por el último nucleótido de cada fragmento, conociendo así el orden secuencial de las bases según su paso por el detector. Al final obtendremos como resultado el conjunto de señales recogidas por el detector ordenadas desde el fragmento más pequeño hasta la molécula completa. Se utilizará una técnica de biología molecular conocida como electroforesis capilar, en la que los fragmentos de ADN se separan por tamaño y pasan por un detector que recoge la señal luminosa emitida por el último nucleótido de cada fragmento, conociendo así el orden secuencial de las bases según su paso por el detector. Al final obtendremos como resultado el conjunto de señales recogidas por el detector ordenadas desde el fragmento más pequeño hasta la molécula completa.

Durante los años siguientes, la secuenciación del ADN trajo consigo una enorme cantidad de logros y aplicaciones que culminaron con la finalización del Proyecto Genoma Humano en 2004 [4], donde se obtuvo por primera vez la secuencia del genoma humano; un genoma se define como la secuencia completa de ADN. de un organismo -. El desarrollo de este enorme y largo proyecto -que duró casi 15 años-, además de la enorme variedad de aplicaciones que podría tener en el futuro, puso de manifiesto la urgente necesidad de tecnologías de secuenciación más avanzadas que permitieran obtener el genoma de un organismo de forma rápida y relativamente accesible en términos económicos para la mayoría de los laboratorios. Así, tras la finalización del proyecto, el Instituto Nacional de Investigación del Genoma Humano (NHGRI) inició un programa de financiación conocido como proyecto Genoma de 1.000 dólares [5]. con el objetivo de disponer de métodos de secuenciación de alta precisión -menos de 1 error por 10.000 bases-, longitudes de lectura largas, alto rendimiento y una reducción del coste de secuenciar un genoma a 1.000 dólares en 10 años. Esto aceleró la llegada de nuevas tecnologías de secuenciación, acuñadas con el término de secuenciación de próxima generación para diferenciarlas de los métodos clásicos de secuenciación de primera generación de Sanger, Maxam y Gilbert.

2. Tecnologías de secuenciación de próxima generación

Los nuevos métodos de secuenciación que comenzaron a surgir a partir de este momento comparten una serie de características que mejoran ostensiblemente el rendimiento de las técnicas clásicas [6]. En primer lugar, se basan en la preparación de bibliotecas de ADN para adaptar el genoma según la técnica de secuenciación a utilizar; en segundo lugar, el proceso de secuenciación se produce en paralelo en múltiples fragmentos del genoma inicial, lo que permite que se produzcan miles o millones de reacciones de secuenciación al mismo tiempo; y finalmente, la detección de las bases secuenciadas se realiza directamente sin necesidad de electroforesis, acelerando enormemente

todo el proceso. Dentro de este amplio abanico de nuevas técnicas de secuenciación diferenciaremos dos grandes grupos: por un lado los métodos de secuenciación de segunda generación, también llamados secuenciación de lecturas cortas o basados en amplificación por PCR,

2.1. Secuenciación de segunda generación

Este tipo de tecnología fue la primera que surgió tras la finalización del Proyecto Genoma Humano desde 2005, y sus rasgos más característicos son la amplificación de los fragmentos a secuenciar mediante PCR y la alta paralelización del proceso. La Reacción en Cadena de la Polimerasa (PCR) es una técnica fundamental en biología molecular en la que se amplifican moléculas de ADN realizando un proceso de replicación in vitro; El objetivo principal es imitar este proceso celular mezclando todos los componentes moleculares necesarios en un tubo de ensayo, lo que da como resultado una gran cantidad de copias exactas del fragmento original. Esta etapa permitirá que la señal óptica detectada por la máquina sea mucho mayor en cada reacción de secuenciación, mejorando mucho la precisión del proceso. La preparación de la biblioteca de ADN es específica del método de secuenciación, aunque todos ellos tienen varias características en común. En primer lugar, el ADN debe fragmentarse en fragmentos que oscilan entre 400 y 1.200 pares de bases en un proceso conocido como "secuenciación rápida", porque la fragmentación se produce al azar en múltiples posiciones del genoma, lo que da como resultado múltiples fragmentos de ADN que se superponen entre sí. Cabe señalar que en un experimento de secuenciación tradicionalmente es necesario extraer el ADN de la muestra biológica que se estudia, normalmente formada por millones de células, cada una con su propia molécula de ADN que se pretende secuenciar; es por eso que cada región del genoma estará representada por múltiples fragmentos, que pueden superponerse o no entre sí. En segundo lugar, todos estos fragmentos están unidos a secuencias adaptativas, pequeños fragmentos de ADN de secuencia conocida que tienen una doble función; por un lado, se encargan de iniciar el proceso de amplificación por PCR -la maquinaria de replicación celular siempre necesita una pequeña secuencia para iniciar la síntesis de ADN- y por otro lado, permitirán que todos los fragmentos a secuenciar sean anclado a un soporte sólido donde tiene lugar el propio proceso de secuenciación [7].

Las tecnologías de secuenciación de segunda generación se dividen en dos grandes categorías [8]: secuenciación por métodos de ligación (SBL), donde la detección de bases se realiza uniendo oligonucleótidos marcados con una molécula fluorescente, y secuenciación por síntesis (SBS), donde la detección de la señal se realiza mediante métodos de ligación. Se produce incorporando un nucleótido en una cadena alargada. En concreto, a continuación se describirán las principales tecnologías SBL - SOLiD - y SBS, esta última clasificada en métodos de terminación cíclica reversible (CRT), con la plataforma Illumina a la cabeza, y de adición de un solo nucleótido (SNA), con pirosecuenciación 454 e Ion Torrent como método. principales buques insignia.

SOLiD no es una de las técnicas más utilizadas en la actualidad, siendo ampliamente desplazada por otros métodos de secuenciación por diversos motivos que se entenderán tras describir su funcionamiento. En primer lugar, es necesario destacar un aspecto de la preparación de la biblioteca de ADN explicado anteriormente, ya que en estos métodos la amplificación de fragmentos por PCR se realiza en diminutas esferas de resina que actúan como microrreactores, en un proceso conocido como emulsión. PCR; su principal diferencia es que se agrega una mezcla formada por una fase acuosa, la cual contiene todos los reactivos químicos necesarios para la amplificación e incluirá cada una de las microesferas. De esta forma, el proceso es completamente paralelo e independiente para cada esfera, llevando cada una un fragmento de ADN diferente. Después, el proceso de secuenciación en sí se realiza utilizando oligonucleótidos modificados químicamente, de tal forma que contienen en un extremo un par de nucleótidos conocidos -1 de 16 combinaciones posibles de nucleótidos existentes- seguidos de una serie de moléculas universales que no tienen especificidad para el Secuencia de ADN. Cuando el oligonucleótido se hibrida a través de este par de bases, el sistema detecta el color fluorescente, tras lo cual se escinde y sigue el proceso de unión del oligonucleótido

hasta llegar al final del fragmento. Una vez finalizado un ciclo se inicia un nuevo proceso de secuenciación con el mismo fragmento, pero esta vez el oligonucleótido se unirá en la base n+1 para detectar el resto de bases que no han sido secuenciadas en el primer ciclo [8]. Tras finalizar la serie de ciclos el resultado es que el mismo fragmento de ADN ha sido secuenciado varias veces pero cambiando el orden del par de nucleótidos que se detecta, por lo que al final cada nucleótido se leerá varias veces, y por tanto reduciendo mucho el tasa de error del proceso de secuenciación. Es sin duda el método con mayor precisión tras la secuenciación de Sanger, con un valor del 99,94%, aunque esta ventaja se ve notablemente eclipsada por los numerosos inconvenientes que presenta, como la corta duración de sus lecturas o la larga duración por ejecución [9] .

Por otro lado, dentro de las tecnologías de segunda generación encontramos el SBS, cuyo método más conocido y más implementado actualmente es la terminación cíclica reversible con la plataforma Solexa/Illumina en el frontal (Illumina adquirió Solexa en 2007). Este método tiene una característica común respecto a la secuenciación de Sanger clásica, el uso de nucleótidos modificados químicamente que al añadirse a la cadena en elongación impiden la unión de otro nucleótido detrás. En cada ciclo de secuenciación se realizan los siguientes pasos: unión de los fragmentos de ADN y sus adaptadores a una superficie sólida -proceso de amplificación dentro de la preparación de la biblioteca, generando clusters con cada fragmento original-, adición de los componentes necesarios para la síntesis de nuevas cadenas de ADN, incluidos los nucleótidos modificados marcados con fluorescencia, hibridación del nucleótido complementario a la secuencia plantilla, lavado de las bases no incorporadas y detección de la molécula fluorescente, y finalmente separación de la parte terminal del nucleótido para que pueda comenzar un nuevo ciclo y secuenciar el fragmento completo [10]. La plataforma de secuenciación Illumina, con su amplia gama de secuenciadores con características y aplicaciones muy dispares, es actualmente líder en la industria de secuenciación de alto rendimiento y la mayoría de los protocolos de preparación de bibliotecas son compatibles con la tecnología Illumina. Con el lanzamiento de su nuevo secuenciador HiSeq X Ten, pudo lograr en gran medida las premisas del NHGRI; En concreto, es sin duda la tecnología con mayor rendimiento hasta la fecha -número de fragmentos secuenciados por ejecución-,

Las tecnologías de segunda generación trajeron también otro enfoque, el método de secuenciación por adición de un solo nucleótido -SNA-, cuyos buques insignia más importantes son el sistema de pirosecuenciación 454 y la tecnología IonTorrent; Ambos métodos se basan en una única señal para marcar la incorporación de una base a una cadena de ADN que se elonga, por lo que en esta ocasión cada uno de los nucleótidos se agregará uno a uno de forma secuencial. La tecnología de pirosecuenciación tiene el honor de ser la primera en ser lanzada en 2005 tras la finalización del HGP, convirtiéndose en el primer instrumento de secuenciación de próxima generación. En este método las bases siempre se incorporan en el mismo orden, y su detección se produce cuando se libera un pirofosfato en la formación del enlace químico entre nucleótidos -los pirofosfatos son moléculas compuestas por dos grupos fosfato que se liberan cuando el nucleótido original pasa a formar parte del ADN-, tras lo cual sufren una reacción química y se transforman. en otro compuesto, la luciferasa, que es capaz de emitir una señal bioluminiscente. Dependiendo de dónde se haya detectado esta señal luminosa se podrá saber en qué fragmento se ha añadido dicho nucleótido, eso lo sabemos de antemano, así como la intensidad nos dirá si se han añadido múltiples bases del mismo tipo. La tecnología de Ion Torrent, por otro lado, fue el primer instrumento NGS sin utilizar señales ópticas para la detección de bases, sustituido en su lugar por un sistema semiconductor que detectará la unión de un nucleótido mediante un cambio de pH en el medio. Esta variación se debe a la liberación de protones en el proceso de síntesis del ADN, lo que permitirá que una señal química en este caso se transforme en información digital [8]. Este tipo de sistemas surgieron a partir de la metodología de pirosecuenciación, teniendo como principales ventajas la alta velocidad de ejecución, un menor coste y una instrumentación más compacta. Sin embargo, este tipo de secuenciación SNA, especialmente la tecnología de pirosecuenciación 454, se ha quedado atrás con respecto a Illumina y otras tecnologías más nuevas, principalmente debido a la gran diferencia en el rendimiento de

ejecución y otros problemas técnicos que veremos a continuación [6]. lo que permitirá transformar una señal química en este caso en información digital [8]. Este tipo de sistemas surgieron a partir de la metodología de pirosecuenciación, teniendo como principales ventajas la alta velocidad de ejecución, un menor coste y una instrumentación más compacta. Sin embargo, este tipo de secuenciación SNA, especialmente la tecnología de pirosecuenciación 454, se ha quedado atrás con respecto a Illumina y otras tecnologías más nuevas, principalmente debido a la gran diferencia en el rendimiento de ejecución y otros problemas técnicos que veremos a continuación [6]. lo que permitirá transformar una señal química en este caso en información digital [8]. Este tipo de sistemas surgieron a partir de la metodología de pirosecuenciación, teniendo como principales ventajas la alta velocidad de ejecución, un menor coste y una instrumentación más compacta. Sin embargo, este tipo de secuenciación SNA, especialmente la tecnología de pirosecuenciación 454, se ha quedado atrás con respecto a Illumina y otras tecnologías más nuevas, principalmente debido a la gran diferencia en el rendimiento de ejecución y otros problemas técnicos que veremos a continuación [6].

2.2 Secuenciación de tercera generación

Uno de los principales inconvenientes de las técnicas de secuenciación de lectura corta está relacionado con una de las últimas etapas del análisis de datos bioinformáticos, es decir, el alineamiento o mapeo de fragmentos de ADN con un genoma de referencia. Como se describirá a continuación, la mayoría de las aplicaciones de secuenciación de alto rendimiento requieren que estas lecturas secuenciadas se alineen con un genoma de referencia, un proceso en el que se utilizan algoritmos de búsqueda para mapear o conocer la posición específica dentro de un genoma de los fragmentos secuenciados. En este sentido, hay dos tipos de problemas a tener en cuenta con este paradigma de secuenciación, ya sea por las propias lecturas secuenciadas o por el genoma de referencia utilizado. En primer lugar, la alta complejidad de los genomas hace que tengan regiones que complican en gran medida el correcto mapeo de fragmentos, como zonas muy repetidas o variaciones estructurales; los primeros pueden medir varios cientos de pares de bases, por lo que nuestras lecturas mucho más cortas tienen la opción de alinearse en varias posiciones diferentes sin llegar a una alineación única, algo que también ocurre con las variaciones estructurales. Este tipo de variantes genómicas, a diferencia del polimorfismo de un solo nucleótido, consisten en fragmentos de una determinada longitud que se duplican o eliminan -una deleción es la eliminación de un fragmento de ADN- en diferentes posiciones del genoma, a menudo incluso distribuidos entre varios cromosomas. Se ha demostrado que este tipo de variantes genómicas están implicadas en un gran número de enfermedades [11], por lo que su mapeo y/o interpretación incorrecta puede tener graves implicaciones clínicas. Además de esta cuestión relativa a la longitud de los fragmentos secuenciados, Otra posible mejora que podría realizarse en los métodos de segunda generación es el uso de PCR para la amplificación clonal de fragmentos. Este proceso se evitará en la tecnología de tercera generación, ya que es una técnica relativamente sensible a errores en zonas de ADN con alto contenido de bases de GC, y daría como resultado un ahorro considerable de tiempo en la preparación de bibliotecas [12].

Las tecnologías de secuenciación de tercera generación o métodos basados en lectura larga se caracterizarán, como su nombre indica, por obtener fragmentos de ADN secuenciados con una longitud mucho mayor que las técnicas descritas anteriormente, del orden de kilobases -1 kb corresponde a 1000 pares de bases-. . Además, como se mencionó anteriormente, los fragmentos a secuenciar no se amplifican mediante PCR, sino que las bases se detectan a partir de la única molécula original obtenida en la preparación de la biblioteca; esto también hace que el proceso de secuenciación sea en tiempo real, es decir, sin ciclos de lavado y escaneo como se hacía en los métodos de segunda generación. El primer acercamiento a este tipo de tecnología nació a principios de 2011 con el lanzamiento del primer secuenciador PacBio RS por parte de Pacific Biosciences. que utiliza la tecnología conocida como SMRT sequencing -secuenciación en tiempo real de una sola molécula-, un método muy similar al utilizado por la plataforma Illumina. A diferencia del sistema Illumina, los fragmentos a secuenciar forman una estructura conocida como SMRTbell -resultado de unir

secuencias adaptativas abiertas en ambos extremos que conforman una molécula lineal y circular- que se cargarán individualmente en pocillos que actuarán como detectores. Dentro de estas estructuras se colocará una ADN polimerasa que estará preparada para replicar la secuencia lineal de ADN que ha entrado, haciendo uso de los correspondientes nucleótidos marcados fluorescentemente, cuyas señales ópticas serán detectadas por un sistema de cámaras en tiempo real. Una particularidad de esta plataforma es que la polimerasa, una vez que ha terminado de replicar la cadena original, puede iniciar el proceso nuevamente con las moléculas resultantes, secuenciando así la misma secuencia varias veces; esto dará como resultado un grado de precisión sin precedentes, alcanzando un nivel del 99,999% con aproximadamente 25 ciclos de secuenciación. Esta gran ventaja se ve limitada por el alto coste que supone durante los primeros años y el bajo rendimiento obtenido; aunque los avances recientes los han mejorado considerablemente, estos siguen siendo los principales obstáculos para implementar esta tecnología en diversos proyectos de secuenciación masiva [12]. Esta gran ventaja se ve limitada por el alto coste que supone durante los primeros años y el bajo rendimiento obtenido; aunque los avances recientes los han mejorado considerablemente, estos siguen siendo los principales obstáculos para implementar esta tecnología en diversos proyectos de secuenciación masiva [12]. Esta gran ventaja se ve limitada por el alto coste que supone durante los primeros años y el bajo rendimiento obtenido; aunque los avances recientes los han mejorado considerablemente, estos siguen siendo los principales obstáculos para implementar esta tecnología en diversos proyectos de secuenciación masiva [12].

El segundo gran enfoque de secuenciación de tercera generación es la tecnología basada en nanoporos conocida como Oxford Nanopore Technology -ONT, nombre dado por la empresa que la desarrolló-, que surgió en 2014 con la aparición de su primer secuenciador MinION. Este novedoso método de secuenciación permite detectar cada nucleótido de la cadena de ADN a su paso por un nanoporo, gracias a los cambios de voltaje que experimenta una corriente eléctrica provocados por el paso de estas moléculas. Los fragmentos de ADN a secuenciar llevan consigo secuencias adaptativas que permitirán la unión de proteínas motoras, cuya función es transportar la cadena de ADN a través del nanoporo; Esta diminuta abertura forma parte de un gran complejo proteico, y tiene en su interior una zona más sensible al paso de los nucleótidos con un detector que mide los cambios en el voltaje de la corriente. en diferentes grados dependiendo de la naturaleza del propio nucleótido. Esta tecnología tiene características que la convierten en la más prometedora en la actualidad, siendo perfectamente válida en múltiples proyectos de secuenciación. Con él se han conseguido las lecturas más largas hasta la fecha -llegando incluso a 1 Mb [13]-, y no hay límite superior si la calidad y cantidad del ADN de origen es buena; el costo de la secuenciación es relativamente bajo, lo que hace posible que pequeños laboratorios la adquieran y utilicen; y por último, su gran escalabilidad -cuenta desde secuenciadores portátiles que caben en la palma de la mano como el MinION hasta dispositivos de alto rendimiento como el PromethION- la convierte en una tecnología verdaderamente flexible a las necesidades que requiere el proyecto. Por otro lado, todavía tiene una tasa de error promedio relativamente alta en comparación con otras plataformas.

Además de estas dos tecnologías principales, en los últimos años han surgido nuevos métodos de secuenciación de lectura larga que es interesante destacar. Estos nuevos sistemas se conocen como SLR, o secuenciación sintética de lectura larga, ya que los fragmentos secuenciados no son realmente de esta longitud, sino lecturas cortas ensambladas in silico para generar un fragmento más grande. Se basan en la plataforma de secuenciación de generación de fragmentos cortos Illumina, pero presentan una serie de cambios respecto a la preparación de la biblioteca. La propia empresa de Illumina adquirió el sistema de secuenciación Moleculo, en el que el ADN inicial se fragmenta en moléculas largas, de hasta 10 kb, y luego se introduce en micropocillos donde se marcarán especialmente con adaptadores a modo de sistema de códigos de barras. Posteriormente se vuelven a fragmentar para poder ser secuenciados por Illumina y al final, Como cada fragmento está etiquetado según su molécula de origen, se pueden reconstruir sintéticamente para generar lecturas largas. La tecnología 10X Genomics, en cambio, no clasifica los fragmentos largos originales en

micropocillos, sino en un sistema de micelas en emulsión, similar al sistema de secuenciación por ligación con microesferas. En cuanto a las ventajas de este tipo de plataformas, su principal característica es que está basada en la secuenciación de Illumina, aprovechando su bajo nivel de error y su enorme rendimiento por ejecución; sin embargo, requiere la adquisición de nuevos equipos para preparar las bibliotecas, aumentando relativamente su costo, además de que dependen de un proceso de amplificación por PCR, por lo que en ocasiones estas tecnologías no son consideradas como métodos basados en lecturas realmente largas [12].

3. Análisis de datos NGS

Los rápidos avances en la secuenciación de alto rendimiento tras la finalización del Proyecto Genoma Humano han permitido que esta tecnología se asiente como una herramienta rutinaria en múltiples laboratorios de investigación y centros genéticos, independientemente de su área de trabajo o de su capacidad para abordar proyectos grandes o pequeños. Se ha descubierto que en nuestro genoma se encuentran las respuestas biológicas a muchas de las preguntas que continuamente se plantea la humanidad respecto a todo tipo de problemas médicos, desde la base de cualquier enfermedad hasta el por qué de nuestra inteligencia. Sin embargo, el nuevo paradigma ha traído consigo una enorme cantidad de datos que los enfoques computacionales clásicos no han logrado manejar, lo que ha llevado a la aparición de múltiples herramientas y algoritmos para intentar analizar y gestionar todos estos datos desde diferentes plataformas de secuenciación. Este conjunto de herramientas se clasificará según la etapa en la que interviene en el procesamiento de los datos generados por los métodos de secuenciación de última generación, desde el análisis de las lecturas del secuenciador hasta la obtención de información biológica relevante según el proyecto en cuestión. . Las aplicaciones donde se puede explotar todo el potencial de la secuenciación de alto rendimiento son muy diversas: la secuenciación de ADN genómico, que incluye también la obtención de nuevos genomas desconocidos, el estudio de variantes genéticas a nivel poblacional o el diagnóstico clínico de enfermedades mendelianas y más. síndromes complejos como el cáncer; Secuenciación de ARN, también conocida como RNA-seq, donde podremos analizar un transcriptoma completo como complemento al uso de microarrays clásicos - el ARN es un polímero de nucleótidos, al igual que el ADN.

En este artículo nos centraremos en analizar la línea existente para el análisis de datos de secuenciación masiva de estudios de identificación de variantes genómicas basados en aplicaciones clínicas, es decir, variaciones del ADN relacionadas con diversas enfermedades humanas; cuando comparamos el ADN de diferentes individuos dentro de la misma especie, -Vemos que es exactamente igual salvo en determinadas posiciones del genoma, cuya variación es responsable de que determinadas proteínas celulares no funcionen, generando una enfermedad, o de las diferencias entre individuos-. Las etapas que componen este tipo de análisis, y que se describirán a continuación, son las siguientes: evaluación de la calidad de las lecturas, alineamiento frente a un genoma de referencia, identificación de las variantes y, finalmente, su anotación para dar significado biológico a los datos. Además de este flujo de trabajo principal,

3.1. Evaluación de calidad

Los archivos resultantes de cualquiera de las tecnologías de secuenciación descritas anteriormente contienen todas las lecturas detectadas por el secuenciador en un formato estandarizado conocido como FASTQ. Este formato presenta un input para cada lectura secuenciada, en el que cada nucleótido trae consigo un valor asociado de su calidad generado por el propio secuenciador. El motivo de introducir una medición de la calidad de las bases es que, como comentamos anteriormente, cada tecnología de secuenciación tiene un determinado valor de precisión a la hora de detectar verdaderos positivos, por lo que los errores de secuenciación son inherentes a todas las técnicas, más aún teniendo en cuenta tener en cuenta tanto el factor humano como los fallos relacionados con la instrumentación o los reactivos químicos utilizados (es común, por ejemplo, en determinadas plataformas, como Illumina, que a medida que avanza la secuenciación

a lo largo del fragmento aumenta la probabilidad de error por desgaste de los componentes moleculares utilizados -. Por tanto, nuestro fichero FASTQ tendrá asociado un valor de probabilidad de error para cada nucleótido de las lecturas, siendo necesario realizar este primer paso para asegurar la calidad de los fragmentos.

En la actualidad existen numerosas herramientas que nos permiten tanto evaluar la calidad general de las lecturas como realizar un recorte de las mismas en función de determinados parámetros para filtrar, por ejemplo áreas que no alcanzan un determinado umbral de calidad. Se han desarrollado herramientas que permiten realizar ambos procesos de forma conjunta, como NGSQC Toolkit [14], PRINSEQ [15] o el entorno Galaxy [16], que producen informes generales de las lecturas y son capaces de filtrarlas. Sin embargo, lo más utilizado hoy en día en los pipelines de análisis de datos de secuenciación es la evaluación de lecturas mediante FASTQC [17] y su posterior filtrado con Trimmomatic [18], dos herramientas totalmente independientes pero con una amplia gama de módulos muy interesantes. FASTQC es un software que proporciona una gran cantidad de gráficos y estadísticas que muestran la calidad promedio de las lecturas, la calidad promedio por base, la distribución de nucleótidos (Ns) indeterminados o contenido de GC, etc.; Por otro lado, Trimmomatic es una herramienta muy potente para filtrar secuencias en base a múltiples parámetros, bastante optimizada para los datos de la plataforma de secuenciación Illumina, además de permitir la eliminación sistemática de adaptadores, secuencias sin valor biológico real que provienen de la preparación de bibliotecas. en cualquiera de las plataformas de secuenciación, cuya eliminación evita la entrada de una gran señal de ruido en etapas posteriores [19]. Aunque no es la etapa más crítica de todo el proceso, es necesario limpiar las lecturas y facilitar el trabajo de las herramientas siguientes, pero actualmente no existen revisiones exhaustivas que comparen el rendimiento de diferentes software de control de calidad. Aún así,

3.2. Alineación

Una vez que las lecturas se han procesado adecuadamente, es necesario realizar un mapeo o alineación con un genoma de referencia existente, para lo cual existen dos fuentes principales, la Universidad de Santa Cruz (UCSC) y el Genome Reference Consortium. Ambas instituciones ponen a disposición de la comunidad científica un conjunto de referencia del genoma humano, sobre el que aplican continuamente mejoras y diversas optimizaciones para conocer la posición genómica específica de millones de lecturas de un secuenciador. En cuanto al proceso de alineación en sí, hay que destacar la enorme complejidad computacional que implica tener que colocar con precisión las lecturas en su posición correcta dentro del genoma, algo que no es tan sencillo como podría parecer. El genoma humano posee una enorme complejidad, con regiones tan extrañas que aún no se han podido caracterizar, como repeticiones de uno o varios nucleótidos en las regiones intergénicas o duplicaciones de un mismo gen en diferentes cromosomas, por lo que el proceso que lleva a cabo un software de alineación es increíblemente costoso. Para facilitar el trabajo de este tipo de herramientas existen varios conceptos que es interesante aclarar. Por un lado, cuanto más largos son los fragmentos, más fácil es mapearlos, del mismo modo que es más fácil encontrar en un libro una coincidencia única de una frase concreta que de una sola palabra, ya que en este último caso es más probable que se encontrarán varias opciones donde colocar la palabra. Por otro lado, un modo de secuenciación muy utilizado hoy en día es la secuenciación de extremos pares, en la que cada fragmento de ADN secuenciado tiene un par conocido y etiquetado de antemano, entonces sabemos exactamente el grado que separa las dos lecturas. Por tanto, son fragmentos que van de la mano, por lo que su mapeo es más preciso porque las coordenadas genómicas de uno pueden ayudar a localizar al otro, si cae en una zona comprometida.

Los diferentes software de alineación se clasificarán según el tipo de algoritmo que implementen para el mapeo de las lecturas [24]. En primer lugar existen algoritmos basados en hash -resultado de la función hash que genera claves para representar de forma inequívoca un conjunto de datos- que elaboran un índice para encontrar rápidamente la posición de cada lectura, pero a cambio de

mapearlas muy rápidamente son muy sensible a los errores; dentro de esta categoría encontraríamos RMAP [24], SOAP [26], Novoalign [27] o SHRiMP [28]. En segundo lugar encontramos los basados en el algoritmo de Smith-Waterman, que aplica métodos de programación dinámica para asegurar que el alineamiento local sea óptimo respecto a un determinado sistema de puntuación, por lo que serán más precisos y menos sensibles a errores pero consumirán más tiempo. ; un ejemplo de este tipo de software es BFAST [29], cuya peculiaridad es que implementa exclusivamente este algoritmo. Finalmente, los algoritmos basados en la transformada de Burrows-Wheeler optimizan el uso de la memoria, siendo actualmente los preferidos para lecturas cortas por ofrecer un equilibrio entre eficiencia, sensibilidad y especificidad; actualmente existen muchas herramientas que implementan este algoritmo, como BWA [30], Bowtie [31], o SOAP2 [32] y SOAP3 [33]. Numerosos estudios han evaluado el rendimiento de diversos alineadores para la identificación de variantes, aunque al final los más utilizados hoy en día son los que mejor rendimiento ofrecen: BWA, Bowtie, Novoalign y SOAP. Generalmente, la mayoría de los proyectos de análisis de datos NGS utilizan BWA o Bowtie2, la versión mejorada de su predecesor, aunque varias revisiones parecen indicar que BWA ofrece resultados ligeramente mejores [34] a una velocidad significativamente más rápida [35]; sin embargo, parece que este software no es tan preciso incluso con tasas de error bajas: una característica de un buen alineador es que puede mapear lecturas correctamente, incluso cuando contienen errores de secuenciación o polimorfismos genéticos que disminuyen las coincidencias con respecto a la genoma de referencia, por lo que es una herramienta que no deja escapar ningún alineamiento potencial, con el compromiso de generar muchas lecturas mapeadas incorrectamente [36]. Novoalign, por su parte, también ha mostrado un muy buen comportamiento respecto al resto de herramientas cuando posteriormente se utiliza GATK para la identificación de variantes [37]. además de presentar una mayor sensibilidad o proporción de verdaderos positivos cuando las lecturas son muy cortas [35]. Finalmente, SOAP y sus versiones mejoradas tienen una alta precisión incluso con altas tasas de error en el mapeo, por lo que parece la mejor opción para la identificación de SNPs o polimorfismos de un solo nucleótido en etapas posteriores [36].

3.3. Procesamiento de llamadas posteriores a la alineación y previas a la variante

Los resultados obtenidos de los algoritmos de mapeo contienen las lecturas alineadas contra el genoma de referencia en un formato cuasi estándar conocido como SAM, o formato Sequence Alignment/Map, en el que se presenta información diversa sobre cada lectura alineada, como la posición específica en el referencia, su orientación -recordemos que la molécula de ADN es bicatenaria, por lo que se dice que tiene una cadena positiva y una cadena negativa, también llamadas directa e inversa, respectivamente- o la calidad de dicha alineación. Esta información se almacena en etiquetas conocidas como banderas, cuyo valor resultante va a ser el resultado de la suma de todas las etiquetas individuales, representando cada una de ellas un tipo de información que nos servirá posteriormente para gestionarlas y filtrarlas. Una vez obtenidos estos archivos SAM, casi siempre es necesario realizar una serie de preprocesos antes de identificar las variantes propiamente dichas, ya sea porque son requisitos imprescindibles para herramientas posteriores o porque facilitan mucho su trabajo. La mayoría de proyectos de análisis de datos NGS centrarán la atención de este preprocesamiento en tres herramientas fundamentales, como SAMtools [38], GATK [39] - un conjunto de herramientas de análisis que luego serán fundamentales para la identificación de variantes - y Picard [40]. . A continuación describiremos el flujo de trabajo seguido en la mayoría de los estudios de identificación de variantes de estas herramientas, ya que tienden a ser procesos más estandarizados y revisados dentro de la comunidad bioinformática [41]. La mayoría de proyectos de análisis de datos NGS centrarán la atención de este preprocesamiento en tres herramientas fundamentales, como SAMtools [38], GATK [39] - un conjunto de herramientas de análisis que luego serán fundamentales para la identificación de variantes - y Picard [40]. . A continuación describiremos el flujo de trabajo seguido en la mayoría de los estudios de identificación de variantes de estas herramientas, ya que tienden a ser procesos más estandarizados y revisados dentro de la comunidad bioinformática [41]. La mayoría de proyectos de análisis de datos NGS centrarán la atención de este preprocesamiento en tres herramientas fundamentales, como SAMtools [38], GATK [39] - un

conjunto de herramientas de análisis que luego serán fundamentales para la identificación de variantes - y Picard [40]. . A continuación describiremos el flujo de trabajo seguido en la mayoría de los estudios de identificación de variantes de estas herramientas, ya que tienden a ser procesos más estandarizados y revisados dentro de la comunidad bioinformática [41].

En primer lugar, la mayoría de los algoritmos de llamada de variantes requieren que las lecturas mapeadas se ordenen por posición genómica y se indexen, es decir, se crea un archivo de índice para facilitar la búsqueda de información sobre las lecturas alineadas. Además, también es común que el archivo SAM se transforme a su versión binaria BAM, que contiene exactamente la misma información pero de forma comprimida para facilitar la gestión de los datos; todo esto se puede hacer a través del paquete SAMtools, que incluso proporciona funciones para ofrecer resúmenes de las principales estadísticas de alineación, como el porcentaje de lecturas correctamente mapeadas o la proporción de pares correctamente alineados. Estas estadísticas nos darán la posibilidad de eliminar ciertos sesgos del propio software de alineación, por ejemplo mantener todas aquellas lecturas mapeadas de forma correcta o única en el genoma de referencia, todo de SAMtools [41]. Posteriormente, y debido a que el uso de GATK para la identificación de variantes está muy estandarizado, es común seguir el protocolo o pipeline de mejores prácticas diseñado por los creadores de este software [42], en el que se desarrollan varias etapas de preprocesamiento de Se detallan las lecturas alineadas antes de la identificación de variantes, utilizando herramientas del propio GATK y de Picard. Por tanto, utilizando como input los archivos de SAMtools, se llevarán a cabo los siguientes procesos: creación de un diccionario de secuencias de referencia y preparación de la información adecuada a partir de las lecturas, que actualizará la información de nuestros archivos; marcar o etiquetar secuencias duplicadas utilizando Picard, ya que se trata de fragmentos de ADN que se han secuenciado varias veces durante el proceso de secuenciación, dando lugar a lecturas que no aportan ningún tipo de información y pueden falsear los valores de cobertura de determinadas regiones del genoma; realineamiento local alrededor de los indeles -inserciones y eliminaciones-, ya que este tipo de variaciones estructurales provocan que las zonas adyacentes sean mapeadas incorrectamente, problema típico en la mayoría de alineadores existentes en la actualidad; y finalmente, se lleva a cabo un proceso GATK específico conocido como BQSR, o Base Quality Score Recalibration, que como su propio nombre indica determinará el valor real de probabilidad de error asociado a cada base secuenciada, que en ocasiones no son del todo precisos. Esto será fundamental porque los algoritmos de identificación de variantes utilizarán posteriormente estos valores de calidad, junto con otro conjunto de parámetros, para obtener el grado de fiabilidad de cada variante identificada.

3.4. llamada variante

Hasta ahora, la identificación de variantes y SNP se realizaba normalmente en microarrays, pero su densidad limitaba hasta cierto punto la detección de polimorfismos genéticos; sin embargo, la aparición de técnicas de secuenciación masiva ha hecho posible un nuevo enfoque de identificación exhaustiva de variantes, cubriendo todos los puntos posibles de un genoma donde existe una variación respecto al de referencia, y pudiendo además obtener variantes que llamamos raras -debido a su baja proporción en la población, cuyo papel en enfermedades complejas ha sido demostrado recientemente [24] [43]. Por tanto, y gracias a numerosas herramientas desarrolladas en los últimos años, podremos obtener un mapa completo de las variantes genómicas de cualquier individuo de una forma mucho más precisa y fiable.

Las variantes genómicas se pueden clasificar en varios grupos, dependiendo tanto de su naturaleza genética como del tipo de algoritmo necesario para identificarlas. En primer lugar hay un primer gran grupo constituido por variantes de pequeña longitud, desde un único nucleótido -lo que conocemos como SNP, polimorfismo de un solo nucleótido, o SNV, variación de un solo nucleótido- hasta varios pares de bases -llamados indeles-. por la conjunción de inserciones y eliminaciones -. Los polimorfismos de un solo nucleótido son las variantes genómicas más comunes y por tanto más

conocidas, basadas simplemente en la sustitución de una base de nucleótido por otra; la maquinaria celular, como se sabe, traduce esta secuencia de nucleótidos en una secuencia de otro tipo de moléculas, los aminoácidos, constituyendo lo que conocemos como proteínas. De este modo, el cambio de un nucleótido a otro provocará a su vez una variación en la secuencia de aminoácidos, que puede tener efectos tanto negativos -la proteína se trunca y deja de realizar su función, dando paso a una enfermedad- como neutros -el cambio de aminoácido no afecta a la proteína en su conjunto y puede seguir realizando su función -, o incluso positiva-, el nuevo aminoácido potencia la proteína existente, ya sea añadiendo una nueva función u optimizando la que ya tenía, lo que finalmente provoca la nueva secuencia a mantener en la evolución por el principio básico de la selección natural -. Por otro lado, los llamados indeles son pequeñas inserciones o deleciones de varios nucleótidos en una posición determinada, que comúnmente provocarán un efecto negativo al alterar la lectura secuencial de la cadena de ADN. Ambos tipos de variantes a su vez se dividirán en dos grupos por motivos técnicos, ya que los algoritmos que las detectan serán diferentes: por un lado, las variantes de línea germinal son aquellas que se producen en las células germinales de un organismo -óvulos y espermatozoides- y son por tanto aquellas que se heredan de la descendencia y estarán presentes en todas las células de tu cuerpo; Por otro lado, las variantes somáticas son aquellas que surgen como su nombre indica en las células somáticas -el resto de células de un organismo- durante la vida adulta de cualquier ser vivo, pero que no transmitirán a la descendencia. Estas últimas variantes, sin embargo, son clave para comprender la aparición y desarrollo de enfermedades complejas como el cáncer. En segundo lugar se encuentran las CNV o variantes de número de copias, basadas en fragmentos repetidos de tamaño relativo que se distribuyen a lo largo del genoma. cuya diferencia entre individuos radica en el número de repeticiones que presenta cada uno. Se ha demostrado que este tipo de variantes representan hasta el 9,5% de todo nuestro genoma [44], y al igual que el resto de variantes pueden ser causa de determinadas enfermedades o no tener ningún efecto visible en el organismo, representando simplemente una variación genética entre individuos. . Finalmente, las variantes estructurales o SV se basan en reordenamientos genéticos de grandes áreas de nuestro genoma, que pueden pasar de un cromosoma a otro o incluso eliminarse por completo, provocando claramente graves problemas en el individuo. y como el resto de variantes puede ser la causa de determinadas enfermedades o no tener ningún efecto visible en el organismo, representando simplemente una variación genética entre individuos. Finalmente, las variantes estructurales o SV se basan en reordenamientos genéticos de grandes áreas de nuestro genoma, que pueden pasar de un cromosoma a otro o incluso eliminarse por completo, provocando claramente graves problemas en el individuo. y como el resto de variantes puede ser la causa de determinadas enfermedades o no tener ningún efecto visible en el organismo, representando simplemente una variación genética entre individuos. Finalmente, las variantes estructurales o SV se basan en reordenamientos genéticos de grandes áreas de nuestro genoma, que pueden pasar de un cromosoma a otro o incluso eliminarse por completo, provocando claramente graves problemas en el individuo.

Por lo general, las diferentes herramientas de llamada de variantes se agruparán según su capacidad para detectar un tipo particular de variante, aunque algunas tienen módulos específicos que permiten la identificación de diferentes tipos de variantes de la misma muestra. En cuanto al primer gran grupo de SNPs e indeles, debido a que aparecen con mayor frecuencia y son más conocidos que las variantes estructurales, podemos decir que se han desarrollado numerosas herramientas basadas fundamentalmente en dos enfoques: por un lado, los métodos heurísticos asignan variantes en función de múltiples fuentes de información relacionadas con la calidad de los datos, como VarScan2 [45], que también implementa métodos estadísticos como la prueba de Fisher para comparar variantes con distribuciones teóricas [24]; por otro lado, los métodos probabilísticos se basan en enfoques bayesianos para optimizar la probabilidad de los genotipos identificados, donde encontramos más herramientas actualmente y muy utilizadas como SAMtools o GATK. Hablando concretamente de llamadores de línea germinal, cuya detección es la más estandarizada de todas, encontramos diversos software como el ya mencionado GATK, SAMtools o VarScan2, además de otros como SNVer [46] o FreeBayes [47]; De todos ellos, el algoritmo GATK suele ser el que siempre

ofrece resultados más fiables y precisos [37], además de contar con módulos para detectar otras variantes y diversas funciones de filtrado y recalibración de los resultados, por lo que parece ser el la mejor opción en la mayoría de los estudios. Sin embargo, otras revisiones han destacado el buen papel de FreeBayes a la hora de detectar un buen número de variantes de verdadera calidad, por lo que puede ser una buena opción en los casos en los que se necesite una mayor precisión en detrimento del número de variantes obtenidas [34]. Por otro lado, a la hora de detectar variantes somáticas sólo se observaron resultados aceptables en las herramientas mencionadas anteriormente, como GATK, SAMtools y VarScan2; aun así, se intentó probar la eficacia de otro software, SomaticSniper [48], que ofreció resultados aceptables al identificar SNP entre muestras tumorales y controles. Para la identificación de CNVs también se han desarrollado algunas herramientas específicas, como CNVnator [49], CONTRA [50], ExomeCNV [51], o RDXplorer [52], mientras que para variantes estructurales disponemos de varios software como Breakpointer [53] , INTELIGENTE [54] o SVMerge [55]. En conclusión, según numerosos estudios y revisiones, es muy recomendable abordar el problema de la identificación de variantes con un enfoque múltiple, es decir, aplicar un conjunto de algoritmos a nuestro conjunto de datos para maximizar el conjunto de variantes potenciales y luego llevar a cabo una serie de filtros para retener la proporción más alta posible de verdaderos positivos; Estos procesos de filtrado se pueden realizar utilizando módulos de herramientas específicos como GATK o SAMtools, o también se puede hacer un filtrado más manual en el que mantenemos aquellas variantes que están presentes en un número determinado de herramientas [19].

3.5. Anotación variante

El proceso de análisis de datos de secuenciación de próxima generación culmina con el proceso de anotación de variantes para aportar cierta importancia biológica a los resultados obtenidos. Gracias a determinadas aplicaciones y herramientas es posible realizar lo que se conoce como anotación biológica o funcional de variantes, en la que se busca una gran cantidad de información sobre dichas variantes en función de múltiples parámetros, como la región genómica donde se encuentra, el gen y la proteína a la que afecta, su efecto según la naturaleza de la variante, etc. Todo esto es posible gracias a toda la información disponible en diferentes bases de datos y recursos online, como dbSNP [56] o el proyecto 1000 genomas [57], lo que a su vez nos proporcionará métricas para evaluar el posible impacto clínico de la variante en cuestión, algo imprescindible si hablamos de proyectos de secuenciación para investigación clínica, donde es necesario conocer la potencial relación o causalidad entre la enfermedad de un paciente y sus variantes genómicas. Estas métricas, como Condel [58], PolyPhen [59] o SIFT [60], proporcionan una puntuación de predicción basada en la anotación de variante que la clasifica según su potencial impacto clínico, desde variantes con gran certeza de ser patógenas, hasta variantes neutras. o variantes posiblemente benignas, e incluso variantes con función desconocida o VUS - Variante de Significado Incierto -. Esta clasificación está actualmente estandarizada y existen pautas de consenso para su evaluación y aplicación en diferentes líneas de análisis de datos NGS [61]. donde es necesario conocer la potencial relación o causalidad entre la enfermedad de un paciente y sus variantes genómicas. Estas métricas, como Condel [58], PolyPhen [59] o SIFT [60], proporcionan una puntuación de predicción basada en la anotación de variante que la clasifica según su potencial impacto clínico, desde variantes con gran certeza de ser patógenas, hasta variantes neutras. o variantes posiblemente benignas, e incluso variantes con función desconocida o VUS - Variante de Significado Incierto -. Esta clasificación está actualmente estandarizada y existen pautas de consenso para su evaluación y aplicación en diferentes líneas de análisis de datos NGS [61]. donde es necesario conocer la potencial relación o causalidad entre la enfermedad de un paciente y sus variantes genómicas. Estas métricas, como Condel [58], PolyPhen [59] o SIFT [60], proporcionan una puntuación de predicción basada en la anotación de variante que la clasifica según su potencial impacto clínico, desde variantes con gran certeza de ser patógenas, hasta variantes neutras. o variantes posiblemente benignas, e incluso variantes con función desconocida o VUS - Variante de Significado Incierto -. Esta clasificación está actualmente estandarizada y existen pautas de consenso para su evaluación y aplicación en diferentes líneas de análisis de datos NGS [61]. proporcionan una puntuación de predicción basada en la anotación de

variante que la clasifica según su potencial impacto clínico, desde variantes con gran certeza de ser patógenas, hasta variantes neutras o posiblemente benignas, e incluso variantes con función desconocida o VUS - Variante de significado incierto -. Esta clasificación está actualmente estandarizada y existen pautas de consenso para su evaluación y aplicación en diferentes líneas de análisis de datos NGS [61]. proporcionan una puntuación de predicción basada en la anotación de variante que la clasifica según su potencial impacto clínico, desde variantes con gran certeza de ser patógenas, hasta variantes neutras o posiblemente benignas, e incluso variantes con función desconocida o VUS - Variante de significado incierto -. Esta clasificación está actualmente estandarizada y existen pautas de consenso para su evaluación y aplicación en diferentes líneas de análisis de datos NGS [61].

Para este proceso también disponemos de numerosas herramientas, cuya principal diferencia con el resto de software de análisis NGS es que muchos ofrecen una interfaz gráfica o una plataforma web que permiten que la anotación funcional sea más intuitiva y no requiera tantos conocimientos computacionales; sin embargo, en la mayoría de los casos los proyectos de secuenciación ofrecen una cantidad tan grande de datos y variantes que este tipo de plataforma no puede soportarlo, por lo que las herramientas de línea de comandos van a ser ampliamente utilizadas cuando se requiera una alta paralelización o cómputo en el proceso. Existen muchas herramientas para realizar este paso, como ANNOVAR [62], NGS-SNP [63], snpEff [64] o VEP [65]; de todos ellos, los más revisados y utilizados actualmente son ANNOVAR y VEP - Variant Effect Predictor -,

4. Marcos de análisis de datos de secuenciación de próxima generación

El análisis de datos de secuenciación de próxima generación implica, como hemos visto, numerosas etapas en las que la salida generalmente se convierte en la entrada del siguiente paso, dando lugar a lo que se conoce como un pipeline, un flujo compuesto por una serie de etapas de análisis. hasta llegar finalmente al resultado que necesitamos, la información biológica y clínica de las variantes genómicas detectadas. Es por esto que el proceso de análisis bioinformático de los datos provenientes de NGS es una tarea que requiere de un mínimo conocimiento informático para saber manejar todos los archivos generados, implementar todo el software de terceros que se ha mencionado anteriormente para cada etapa y, en la mayoría de los casos, En estos casos, cree un script que pueda ejecutarse en la línea de comandos para automatizar más el proceso. Por ello, durante los últimos años la comunidad bioinformática ha estado desarrollando líneas analíticas para afrontar este problema. generando herramientas en las que la única tarea es importar las lecturas crudas provenientes del secuenciador y dejarlo trabajar para finalmente obtener un conjunto de variantes identificadas con información biológica relevante, permitiendo su uso y aplicación por parte de investigadores sin ningún conocimiento computacional. Muchas veces este tipo de software trae consigo una interfaz gráfica para que el usuario pueda modificar parámetros y la interpretación de los resultados sea mucho más intuitiva, evitando lo que en informática se conoce como caja negra, un sistema en el que sólo se estudian las entradas y salidas. sin poder conocer ni tener en cuenta su funcionamiento interno. permitiendo su uso y aplicación por parte de investigadores sin ningún conocimiento computacional. Muchas veces este tipo de software trae consigo una interfaz gráfica para que el usuario pueda modificar parámetros y la interpretación de los resultados sea mucho más intuitiva, evitando lo que en informática se conoce como caja negra, un sistema en el que sólo se estudian las entradas y salidas. sin poder conocer ni tener en cuenta su funcionamiento interno. permitiendo su uso y aplicación por parte de investigadores sin ningún conocimiento computacional. Muchas veces este tipo de software trae consigo una interfaz gráfica para que el usuario pueda modificar parámetros y la interpretación de los resultados sea mucho más intuitiva, evitando lo que en informática se conoce como caja negra, un sistema en el que sólo se estudian las entradas y salidas. sin poder conocer ni tener en cuenta su funcionamiento interno.

La necesidad de desarrollar estos canales y flujos de trabajo también surge de la gran cantidad de desafíos que plantea el nuevo paradigma del análisis de datos genómicos. Las numerosas

aplicaciones que se están descubriendo en esta era de la secuenciación masiva está provocando la constante aparición de nuevas herramientas, la evolución y optimización de plataformas existentes o el desarrollo de algoritmos cada vez más innovadores para abordar nuevos problemas que van surgiendo. Todo ello se traduce en un aumento de la complejidad del análisis y una dificultad cada vez mayor a la hora de seleccionar las herramientas adecuadas para cada subproceso del pipeline, ya que para cada paso surgen nuevos algoritmos, cada vez más sofisticados y optimizados, hasta el punto de que en 2017 Ya existían más de 11.000 herramientas para el análisis de datos ómicos catalogadas en la plataforma OMICtools [66]. Esto se ve agravado por la constatación de que esta alta complejidad muchas veces se deja en manos de los investigadores, quienes además de su propia línea de investigación deben ser capaces, con sus escasos conocimientos de TI, de ensamblar estos pipelines y elegir la herramienta adecuada en cada paso. , dejando clara la urgencia de estandarizar los análisis y aumentar la reproducibilidad en biología computacional [67]. Finalmente, esta complejidad de uso que hemos mencionado hace aún más necesario el desarrollo de tecnologías que no requieran un alto nivel técnico, sin tener que aplicar intrincadas instrucciones de línea de comandos, de modo que el grupo de usuarios que puedan aplicar este tipo de análisis sea muy ampliado [68]. con sus escasos conocimientos de TI, para ensamblar estos canales y elegir la herramienta adecuada en cada paso, dejando clara la urgencia de estandarizar los análisis y aumentar la reproducibilidad en biología computacional [67]. Finalmente, esta complejidad de uso que hemos mencionado hace aún más necesario el desarrollo de tecnologías que no requieran un alto nivel técnico, sin tener que aplicar intrincadas instrucciones de línea de comandos, de modo que el grupo de usuarios que puedan aplicar este tipo de análisis sea muy ampliado [68]. con sus escasos conocimientos de TI, para ensamblar estos canales y elegir la herramienta adecuada en cada paso, dejando clara la urgencia de estandarizar los análisis y aumentar la reproducibilidad en biología computacional [67]. Finalmente, esta complejidad de uso que hemos mencionado hace aún más necesario el desarrollo de tecnologías que no requieran un alto nivel técnico, sin tener que aplicar intrincadas instrucciones de línea de comandos, de modo que el grupo de usuarios que puedan aplicar este tipo de análisis sea muy ampliado [68].

Algunos análisis bioinformáticos en proceso de análisis existentes a menudo ofrecen un orden predefinido de pasos y procesos a llevar a cabo, no permitiendo una gran flexibilidad para modificar o reemplazar ciertos módulos; es el caso de pipelines como HugeSeq [69], SIMPLEX [70], TREAT [71], bcbio-nextgen [72] o Sam2bam [73], que implementan un análisis automático de los datos NGS desde la recepción de las lecturas hasta el identificar diferentes tipos de variantes, tener la capacidad de recibir diferentes formatos, usarse en plataformas en la nube, realizar secciones específicas de todo el proceso y ofrecer a los investigadores resultados completos en forma de informes resumidos. Sin embargo, suelen ser herramientas poco flexibles a la hora de insertar nuevos módulos o modificar determinadas etapas para adaptarlo a las necesidades del proyecto en cuestión. por lo que pueden quedar rezagados especialmente para la comunidad bioinformática debido a su gran rigidez. Para solucionar esto surgen nuevas plataformas conocidas como sistemas de gestión de flujos de trabajo o pipeline frameworks, herramientas que ofrecen mayor apertura y flexibilidad para dar cabida a diferentes pipelines, tanto en serie como en paralelo, dependencias complejas, software variado o parámetros modificados por el usuario, además de más características avanzadas como la visualización del proceso en tiempo real, la posibilidad de trabajar en la nube y con interfaz gráfica de usuario o la capacidad de contenerizar varias herramientas [74]. Actualmente existe una gran cantidad de sistemas de gestión de flujo de trabajo, algunos más estandarizados y otros más novedosos e innovadores. Galaxy [75] es una plataforma web ampliamente utilizada en análisis bioinformático con más de 100 herramientas disponibles para las diferentes etapas del análisis NGS, con la posibilidad de crear pipelines personalizados, reproducirlos y compartirlos posteriormente con la comunidad. Al ser una plataforma web, la interfaz gráfica permite que su uso sea sencillo e intuitivo incluso en la creación y personalización de scripts, por lo que se ha convertido en un sistema de benchmarking para el resto de frameworks de flujo de trabajo por su amplio uso en la comunidad científica. SEQprocess [76] es un framework para realizar análisis de datos NGS que ya ofrece varios

pipelines preinstalados, así como la posibilidad de generarlos de forma personalizada. Es un paquete R cuya principal característica es que implementa análisis específicos para nuevas aplicaciones oncológicas basados en las TGCA, El Atlas del Genoma del Cáncer, aunque en el caso de realizar tu propio pipeline requiere de cierta base computacional para instalar el software específico y modificar los parámetros en los archivos de configuración. Closha [77], otro framework de flujo de trabajo desarrollado recientemente, es un sistema optimizado para su uso en la nube a través de clusters informáticos de alto rendimiento, también con una interfaz gráfica y la posibilidad de ejecutar tanto pipelines existentes como personalizados por el usuario. Presenta ciertas ventajas técnicas, como la implementación de un nuevo sistema conocido como KoDS para la transferencia rápida de archivos o la escalabilidad de los recursos -aumenta su rendimiento a medida que aumentan los requisitos computacionales-, lo que hace que su velocidad de ejecución sea ligeramente superior a la de Galaxy. NGS-pipe [78] es otro marco de análisis que permite diseñar tuberías personalizadas de forma automática y fácil de usar, garantiza la reproducibilidad en aplicaciones clínicas y permite la paralelización en grupos; sin embargo, también requiere la instalación manual del software y la modificación de un archivo de configuración para ajustar los parámetros necesarios. Finalmente, otro framework más innovador en este sentido es Bio-Docklet [79], una herramienta que permite gestionar pipelines de otros sistemas como Galaxy en contenedores Docker, encapsulando todo el software preconfigurado necesario y siendo un enfoque muy interesante en la actualidad. escenario para que el investigador no tenga que preocuparse por instalar manualmente todo el software requerido. también requiere la instalación manual del software y la modificación de un archivo de configuración para ajustar los parámetros necesarios. Finalmente, otro framework más innovador en este sentido es Bio-Docklet [79], una herramienta que permite gestionar pipelines de otros sistemas como Galaxy en contenedores Docker, encapsulando todo el software preconfigurado necesario y siendo un enfoque muy interesante en la actualidad. escenario para que el investigador no tenga que preocuparse por instalar manualmente todo el software requerido. también requiere la instalación manual del software y la modificación de un archivo de configuración para ajustar los parámetros necesarios. Finalmente, otro framework más innovador en este sentido es Bio-Docklet [79], una herramienta que permite gestionar pipelines de otros sistemas como Galaxy en contenedores Docker, encapsulando todo el software preconfigurado necesario y siendo un enfoque muy interesante en la actualidad. escenario para que el investigador no tenga que preocuparse por instalar manualmente todo el software requerido.

Como hemos visto, actualmente existen múltiples herramientas y plataformas para afrontar la ardua tarea de analizar datos de un proyecto de secuenciación masiva, cada una de ellas más sofisticada que la anterior; esto lo convierte en un campo de investigación computacional que evoluciona muy rápidamente, por lo que carece de gran parte de la estandarización y reproducibilidad que presentan otros campos científicos, como puede ser el caso de la investigación clínica y biomédica, en las que suelen existir protocolos bien estructurados y acordados. pautas a seguir antes de un experimento en particular. Por tanto, ante esta situación nuestra recomendación sería realizar un estudio exhaustivo de las aplicaciones en las que está presente la secuenciación de última generación, evaluando cada caso concreto y optimizando los diferentes parámetros requeridos. Por lo tanto, Se puede decir que la elaboración de directrices y canales para cada una de las aplicaciones sería un gran paso hacia la mejora de la transparencia y la reproducibilidad entre diferentes proyectos de secuenciación. En segundo lugar, en relación a los diferentes sistemas framework existentes, se ha visto que cuentan con numerosas herramientas para convertirse en instrumentos habituales en cualquier laboratorio biomédico, como su facilidad de uso, la posibilidad de ejecutar trabajos paralelos en la nube o la implementación de gráficos. interfaces de usuario; sin embargo, estas interfaces sólo se contemplan durante el proceso de análisis y creación del pipeline personalizado, ya que se ha visto que en la mayoría de los casos la información final no se ofrece de forma tan gráfica e intuitiva. En casos como la investigación clínica donde las conclusiones de un experimento o incluso la salud de un paciente dependen de la información final obtenida, la interpretación de los resultados es fundamental, por lo que sería necesario centrar esfuerzos en

desarrollar un pipeline robusto y optimizado para clínicos en los que uno de los pilares básicos es la anotación biológica de las variantes y una óptima interpretación de las mismas, generando informes gráficos e intuitivos donde aparece la información clínicamente más relevante de un paciente. Finalmente, un enfoque muy interesante para aplicar en estos pipelines de análisis de datos NGS sería la implementación de algoritmos de aprendizaje automático, algo bastante escaso en los frameworks estándar actuales. La inteligencia artificial se está abriendo camino en los últimos años en el campo de la genómica,

Agradecimientos
El presente estudio ha sido financiado por el proyecto AIR Genomics (con número de expediente CCTT3/20/SA/0003), mediante la convocatoria 2020 PROYECTOS I+D ORIENTADOS A LA EXCELENCIA Y MEJORA COMPETITIVA DE LOS CCTT por el Instituto de Competitividad Empresarial de Castilla y León y fondos FEDER.

Contribuciones de autor: Conceptualización, AC y MS; metodología, AC; redacción: preparación del borrador original, VMD; redacción: revisión y edición, AC, JMC y MS; supervisión, EM; administración de proyectos, JMC; adquisición de financiación, JMC

Conflictos de interés: Los autores declaran no tener ningún conflicto de intereses.

Referencias

1. Watson, JD y Crick, FH La estructura del ADN. Puerto de primavera fría. Síntoma. Cuant. Biol. 18, 123-131 (1953).
2. F. Sanger, et al. Secuenciación de ADN con inhibidores terminadores de cadena. Proc. Nacional. Acad. Ciencia. EE.UU., 74 (1977), págs. 5463-5467.
3. AM Maxam, W. Gilbert. Un nuevo método para secuenciar el ADN. Proc. Nacional. Acad. Ciencia. EE.UU., 74 (1977), págs. 560-564.
4. Consorcio Internacional de Secuenciación del Genoma Humano. Terminando la secuencia eucromática del genoma humano. Naturaleza, 431 (2004), págs. 931-945.
5. JA Schloss. Cómo obtener genomas a una diezmilésima parte del coste. Nat. Biotechnol., 26 (2008), págs. 1113-1115.
6. Erwin L. van Dijk, et al. Diez años para la secuenciación de próxima generación. Tendencias en genética, volumen 30, número 9 (2014), págs. 418-426.
7. Steven R. Head, et al. Construcción de bibliotecas para secuenciación de próxima generación: descripciones generales y desafíos. Biotecnicas. 2014; 56(2): 61–pasim.
8. Sara Goodwin, et al. Mayoría de edad: diez años de tecnologías de secuenciación de próxima generación. Nature Reviews Genetics 17, 333-351 (2016).
9. Lin Liu, et al. Comparación de sistemas de secuenciación de próxima generación. J Biomed Biotecnología. 2012.
10. Michael L. Metzker. Tecnologías de secuenciación: la próxima generación. Nature Reviews Genetics 11, 31-46 (2010).
11. J. Weischenfeldt, et al. Impacto fenotípico de la variación estructural genómica: conocimientos desde y para las enfermedades humanas. Nat. Rev. Genet., 14 (2013), págs. 125-138.
12. Erwin L. van Dijk, et al. La tercera revolución en la tecnología de secuenciación. Volumen 34, Número 9, septiembre de 2018, págs. 666-681.
13. M. Jain, et al. Secuenciación de nanoporos y ensamblaje de un genoma humano con lecturas ultralargas. Nat. Biotechnol., 36 (2018), págs. 338-345.
14. Dai M, Thompson RC, Maher C, et al. NGSQC: proceso de análisis de calidad multiplataforma para datos de secuenciación profunda. Genómica BMC. 2010;11(Suplemento 4):S7.
15. Schmieder R, Edwards R. Control de calidad y preprocesamiento de conjuntos de datos metagenómicos. Bioinformática. 2011;27:863–4.
16. Blankenberg D, Gordon A, Von Kuster G, et al. Manipulación de datos FASTQ con Galaxy. Bioinformática. 2010;26:1783–5.

17. Babraham Bioinformatics, FastQC: una herramienta de control de calidad para datos de secuencia de alto rendimiento. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
18. Anthony M. Bolger, et al. Trimmomatic: un recortador flexible para datos de secuencia de Illumina. Bioinformática. 1 de agosto de 2014; 30(15): 2114–2120.
19. Un estudio de herramientas para el análisis de variantes de datos de secuenciación del genoma de próxima generación. Breve Bioinformación. marzo de 2014; 15(2): 256-278.
20. Hong C, et al. PathoQC: preprocesamiento de lectura computacionalmente eficiente y control de calidad para conjuntos de datos de secuenciación de alto rendimiento. Informe sobre el cáncer. 12 de mayo de 2015; 13 (suplemento 1): 167-76.
21. Shifu Chen, et al. AfterQC: filtrado, recorte, eliminación de errores y control de calidad automáticos para datos fatsq. BMC Bioinformatics 18, número de artículo: 80 (2017).
22. Shifu Chen, et al. fastp: un preprocesador FASTQ todo en uno ultrarrápido. Bioinformática, volumen 34, número 17, septiembre de 2018, págs. i884–i890.
23. Xiaoshuang Liu, et al. FastProNGS: procesamiento rápido de lecturas de secuenciación de próxima generación. BMC Bioinformática 2019; 20: 345.
24. M. Mielczarek, J. Szyda. Revisión de algoritmos de alineación y llamada de SNP para datos de secuenciación de próxima generación. Journal of Applied Genetics, volumen 57, número 1 (2016), págs. 71-79.
25. Smith AD, et al. Actualizaciones del software de mapeo de lectura corta RMAP. Bioinformática 25 (2009): 2841–2842.
26. Li R, Li Y, Kristiansen K, Wang J (2008b) SOAP: programa corto de alineación de oligonucleótidos. Bioinformática 24:713–714.
27. Novocraft (2010), http://www.novocraft.com/.
28. Rumble SM, et al. Camarones: mapeo preciso de lecturas cortas de espacios de color, PLoS Comput. Biol., 2009, vol. 5 pág. e1000386.
29. Homer N, Merriman B, Nelson SF (2009) BFAST: una herramienta de alineación para la resecuenciación del genoma a gran escala. PLoS One 4(11):e7767.
30. Li H, Durbin R. Alineación de lectura corta rápida y precisa con la transformada de Burrows-Wheeler. Bioinformática. 15 de julio de 2009; 25(14):1754-60.
31. Langmead B, Trapnell C, Pop M, Salzberg SL. Alineamiento ultrarrápido y con memoria eficiente de secuencias cortas de ADN con el genoma humano. Genoma Biol. 2009; 10(3):R25.
32. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: una herramienta ultrarrápida mejorada para alineación de lecturas cortas. Bioinformática. 1 de agosto de 2009; 25(15):1966-7.
33. Liu CM, et al. SOAP3: herramienta de alineación paralela ultrarrápida basada en GPU para lecturas breves. Bioinformática. 15 de marzo de 2012; 28 (6): 878-9.
34. Sohyun Hwang, et al. Comparación sistemática de canales de llamadas de variantes utilizando variantes de exoma personal estándar de oro. Representante de ciencia 2015; 5: 17875.
35. Subazini Thankaswamy-Kosalai, et al. Evaluación y valoración del mapeo de lectura mediante múltiples alineadores de secuenciación de próxima generación basados en características de todo el genoma. Genomics, volumen 109, números 3 a 4, julio de 2017, págs. 186-191.
36. Matthew Ruffalo, et al. Análisis comparativo de algoritmos para alineación de lectura de próxima generación. Bioinformática, volumen 27, número 20 (2011), págs. 2790-2796.
37. Adam Cornish y Chittibabu Guda. Una comparación de canalizaciones de llamadas de variantes utilizando Genome in A Bottle como referencia. Biomed Res Int. 2015; 2015: 456479.
38. Li H, et al. El formato de alineación/mapa de secuencia y SAMtools. Bioinformática 25 (2009): 2078–2079.
39. McKenna A, et al. El kit de herramientas de análisis del genoma: un marco MapReduce para analizar datos de secuenciación de ADN de próxima generación. Genoma Res 20 (2010): 1297–1303.
40. http://picard.sourceforge.net/
41. André Altman, et al. Una guía para principiantes sobre llamadas SNP a partir de datos de secuenciación de ADN de alto rendimiento. Genética humana, volumen 131, número 10 (2012), págs. 1541-1554.
42. Geraldine A. Van der Auwera, et al. Desde datos FastQ hasta llamadas de variantes de alta confianza: el canal de mejores prácticas del kit de herramientas de análisis del genoma. Bioinformática del Protocolo Curr. 15 de octubre de 2013; 11(1110): 11.10.1–11.10.33.

43. Handel AE, Disanto G, Ramagopalan SV. Secuenciación de próxima generación para comprender enfermedades neurológicas complejas. Experto Rev Neurother 13(2):215–227 (2013).
44. Zrrei M, et al. Un mapa de variación del número de copias del genoma humano. Nat Rev Genet. Marzo de 2015; 16(3): 172-83.
45. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: descubrimiento de mutación somática y alteración del número de copias en cáncer mediante secuenciación del exoma. Genoma Res 22(3):568–576 (2012).
46. Wei Z, Wang W, Hu P, et al. SNVer: una herramienta estadística para la llamada de variantes en el análisis de datos de secuenciación de próxima generación individuales o agrupados. Ácidos nucleicos res. 2011;39:e132.
47. Garrison E, Marth G. Detección de variantes basada en haplotipos a partir de secuenciación de lectura corta. Preimpresión de arXiv arXiv:1207.3907 [q-bio.GN] 2012.
48. Larson DE, Harris CC, Chen K, et al. SomaticSniper: identificación de mutaciones puntuales somáticas en datos de secuenciación del genoma completo. Bioinformática. 2012;28:311–7.
49. Abyzov A, Urban AE, Snyder M, et al. CNVnator: un enfoque para descubrir, genotipar y caracterizar CNV típicas y atípicas a partir de la secuenciación del genoma familiar y poblacional. Genoma Res. 2011;21:974–84.
50. Li J, Lupat R, Amarasinghe KC, et al. CONTRA: análisis del número de copias para resecuenciación dirigida. Bioinformática. 2012;28:1307–13.
51. Sathirapongsasuti JF, Lee H, Horst BAJ, et al. Variación del número de copias basada en la secuenciación del exoma y detección de pérdida de heterocigosidad: ExomeCNV. Bioinformática. 2011;27:2648–54.
52. Yoon S, Xuan Z, Makarov V, et al. Detección sensible y precisa de variantes del número de copias utilizando la profundidad de cobertura de lectura. Genoma Res. 2009;19:1586–92.
53. Sun R, Love MI, Zemojtel T, et al. Punto de interrupción: uso de artefactos de mapeo local para admitir el descubrimiento de puntos de interrupción de secuencia a partir de lecturas de un solo extremo. Bioinformática. 2012;28:1024–5.
54. Marschall T, Costa I, Canzar S, et al. INTELIGENTE: buscador de variantes que enumera camarillas. Bioinformática. 2012;28(22):2875–288.
55. Wong K, Keane TM, Stalker J, et al. Detección mejorada de variantes estructurales y puntos de interrupción mediante SVMerge mediante la integración de múltiples métodos de detección y ensamblaje local. Genoma Biol. 2010;11:R128.
56. https://www.ncbi.nlm.nih.gov/snp/
57. El Consorcio del Proyecto 1000 Genomas. Una referencia mundial para la variación genética humana. Nature, volumen 526, páginas 68 a 74 (1 de octubre de 2015).
58. González-Pérez A, López-Bigas N. Mejora de la evaluación del resultado de SNV no sinónimos con una puntuación deletérea de consenso, Condel. Soy J Hum Genet. 2011;88:440–9.
59. Adzhubei IA, Schmidt S, Peshkin L, et al. Un método y servidor para predecir mutaciones sin sentido dañinas. Métodos Nat. 2010;7:248–9.
60. Kumar P, Henikoff S, Ng PC. Predecir los efectos de codificar variantes no sinónimas sobre la función de las proteínas utilizando el algoritmo SIFT. Protocolo Nacional. 2009;4:1073–81.
61. Sue Richards, et al. Estándares y pautas para la interpretación de variantes de secuencia: una recomendación de consenso conjunto del Colegio Americano de Genética y Genómica Médica y la Asociación de Patología Molecular. Genética en Medicina, volumen 17, páginas 405–423 (2015).
62. Wang K, Li M, Hakonarson H. ANNOVAR: anotación funcional de variantes genéticas a partir de datos de secuenciación de alto rendimiento. Ácidos nucleicos res. 2010;38:e164.
63. Grant JR, Arantes AS, Liao X, et al. Anotación detallada de los SNP que surgen de proyectos de resecuenciación que utilizan NGS-SNP. Bioinformática. 2011;27:2300–1.
64. Cingolani P, Patel VM, Coon M, et al. Utilizando Drosophila melanogaster como modelo para estudios mutacionales químicos genotóxicos con un nuevo programa, SnpSift. Genet delantero. 2012;3:35.
65. McLaren W, Pritchard B, Ríos D, et al. Derivar las consecuencias de variantes genómicas con el predictor de efectos Ensembl API y SNP. Bioinformática. 2010;26:2069–70.
66. https://omictools.com/
67. Geir Kjetil Sandve, et al. Diez reglas simples para una investigación computacional reproducible. PLoS Comput Biol. Octubre de 2013; 9 (10).

68. Jeremy Davis-Turak, et al. Canalizaciones de genómica e integración de datos: desafíos y oportunidades en el entorno de la investigación. Experto Rev Mol Diagn. marzo de 2017; 17(3): 225–237.
69. Lam HYK, Pan C, Clark MJ, et al. Detectar y anotar variaciones genéticas utilizando el canal HugeSeq. Nat Biotecnología. 2012;30:226–9.
70. Fischer M, Snajder R, Pabinger S, et al. SIMPLEX: canal habilitado en la nube para el análisis integral de datos de secuenciación del exoma. Más uno. 2012;7:e41948.
71. Asmann YW, Middha S, Hossain A, et al. TREAT: una herramienta bioinformática para anotaciones y visualizaciones de variantes en datos de secuenciación de exomas y específicos. Bioinformática. 2012;28:277–8.
72. https://github.com/bcbio/bcbio-nextgen
73. Ogasawara T, Cheng Y, Tzeng TK. Sam2bam: marco de alto rendimiento para herramientas de preprocesamiento de datos NGS. Más uno. 18 de noviembre de 2016;11(11).
74. Leipzig J. Una revisión de los marcos de tuberías bioinformáticas. Breve Bioinformación. 1 de mayo de 2017; 18 (3): 530-536.
75. Goecks J, Nekrutenko A, Taylor J; Equipo Galaxia. Galaxy: un enfoque integral para respaldar la investigación computacional accesible, reproducible y transparente en las ciencias de la vida. Genoma Biol. 2010;11(8):R86.
76. Joo T, Choi JH, Lee JH, Park SE, Jeon Y, Jung SH, Woo HG. SEQprocess: un marco de canalización modularizado y personalizable para el procesamiento NGS en el paquete R. Bioinformática BMC. 20 de febrero de 2019;20(1):90.
77. Ko G, Kim PG, Yoon J, Han G, Park SJ, Song W, Lee B. Closha: sistema de flujo de trabajo bioinformático para el análisis de datos de secuenciación masiva. Bioinformática BMC. 19 de febrero de 2018; 19 (suplemento 1): 43.
78. Singer J, Ruscheweyh HJ, Hofmann AL, Thurnherr T, Singer F, Toussaint NC, Ng CKY, Piscuoglio S, Beisel C, Christofori G, Dummer R, Hall MN, Krek W, Levesque MP, Manz MG, Moch H, Papassotiropoulos A , Stekhoven DJ, Wild P, Wüst T, Rinn B, Beerenwinkel N. NGS-pipe: un marco flexible, fácilmente ampliable y altamente configurable para el análisis NGS.
79. Kim B, Ali T, Lijeron C, Afgan E, Krampis K. Bio-Docklets: contenedores de virtualización para la ejecución en un solo paso de canalizaciones NGS. Gigaciencia. 1 de agosto de 2017; 6 (8): 1-7.