UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

DOCTORADO EN CIENCIAS BIOMÉDICAS

CENTRO DE CIENCIAS GENÓMICAS

DETECCIÓN PRECISA DE VARIANTES *DE NOVO* DE UN SÓLO NUCLEÓTIDO EN GENOMAS HUMANOS

TESIS

QUE PARA OPTAR POR EL GRADO DE

DOCTORA EN CIENCIAS

PRESENTA

LCG LAURA LUCILA GÓMEZ ROMERO

DIRECTOR DE TESIS
DR RAFAEL PALACIOS DE LA LAMA

ENTIDAD DE ADSCRIPCIÓN:
CENTRO DE CIENCIAS GENÓMICAS

COMITÉ TUTOR
DR JOSÉ GUILLERMO DÁVILA RAMOS
DR FEDERICO SÁNCHEZ RODRÍGUEZ

ENTIDAD DE ADSCRIPCIÓN:
CENTRO DE CIENCIAS GENÓMICAS

CUERNAVACA, MORELOS. NOVIEMBRE 2018

Estar preparado es importante,
saber esperar lo es aún más,
pero aprovechar el momento adecuado
es la clave de la vida.
Arthur Schnitzler

Un pájaro posado en un árbol
nunca tiene miedo de que la rama se rompa,
porque la confianza no está en la rama
sino en sus propias alas.
Anónimo

ESTE TRABAJO ESTÁ DEDICADO A MI FAMILIA
POR SIEMPRE CREER EN MÍ

Agradecimientos

Agradezco a la UNAM y al CCG por brindarme la formación necesaria para culminar mis estudios de Posgrado. Agradezco al CONACYT y a la Unidad de Posgrado del Doctorado en Ciencias Biomédicas de la UNAM por el apoyo económico recibido durante todo este tiempo.

Agradezco al Dr Rafael Palacios de la Lama, quien me apoyó desde la Licenciatura y durante el doctorado. Le agradezco por ser parte de mi formación académica y por prepararme para enfrentar los retos de la vida profesional. Agradezco al Dr Michael Schatz de CSHL por aconsejarme en los aspectos técnicos de mi proyecto, también le agradezco enormemente el darme acceso a la infraestructura computacional de CSHL, la cual fue indispensable para realizar este proyecto.

Agradezco al Dr David Romero y al Dr Julio Collado por apoyarme cuando las cosas resultaron diferentes a lo planeado. Recuerdo con gran alegría las largas pláticas sobre mi futuro académico. Realmente aprecio la confianza que siempre mostraron en mis capacidades y habilidades. También agradezco al Dr Enrique Hernández y al Dr Hugo Tovar por confiar en mi capacidad.

Agradezco a mi familia. Especialmente a mi mamá, mi papá y mi hermana, quienes han sido un oasis para mí durante toda mi vida. Sin importar la situación, siempre he contado y sé que siempre contaré con su apoyo. Me han enseñado a ser la persona que soy. Le dedico esta tesis a mis padres que siempre han promovido en mí el amor por la ciencia, me han inculcado una cultura de respeto, me han enseñado a perseguir mis sueños y a nunca darme por vencida; y a mi hermana, quien es mi mejor amiga en el mundo y quien me enseña con su ejemplo. Los tres son un ejemplo para mí, los admiro y estoy tremendamente orgullosa de ser su hija y hermana. A todos mis primos, primas, tíos y tías, gracias por estar al pendiente y sentirse felices por cada logro pequeño o grande, esa felicidad siempre ha sido un empuje que me ayuda a seguir adelante. Abuelita, yo sé que estarías orgullosa de mí.

Agradezco a todos los que han formado parte de mi vida en sus distintas etapas. A todos los que compartieron momentos increíbles conmigo durante la licenciatura y durante el doctorado, gracias. A todos los que me acompañaron durante este viaje, saben que sin ustedes este viaje no hubiera resultado tan llevadero y agradable. Mariana y Daniel, gracias por escuchar pacientemente mis historias. Luis Pedro, Chiapas, Dani y Orli estoy feliz de haber compartido tantos momentos increíbles con ustedes, he aprendido lecciones valiosas de cada uno. Osam y Alex, esas clases de baile siempre fueron mi válvula de escape, fue un placer compartirlas con ustedes. Héctor, Claire, Omar, Andrei y Julio, el doctorado hubiera sido un poco triste y aburrido sin todos esos jueves de palapa juntos.

Índice general

Índice de figuras

Índice de cuadros

RESUMEN

La localización precisa de variantes genéticas *de novo* tiene grandes implicaciones a través de diferentes campos de la biología y de la medicina, particularmente en el campo de la medicina personalizada. Actualmente, las variantes *de novo* son identificadas mapeando las lecturas de un trio padres-descendencia hacia un genoma de referencia, permitiendo cierto grado de diferencias. Aunque este enfoque es ampliamente utilizado, puede generar Falsos Postivos (FP) debido al posicionamiento incorrecto de las lecturas o a la incorrecta caracterización de errores de secuenciación. En un estudio anterior, nuestro grupo desarrolló una propuesta alternativa para identificar con una alta precisión variantes de una sola base (SNVs) utilizando únicamente alineamientos perfectos. Sin embargo, este enfoque sólo puede ser aplicado a regiones haploides del genoma y es computacionalmente intensivo. En este estudio, presentamos una estrategia única, Identificación de Variantes de una sola base basado en la cobertura (COBASI), la cual permite la exploración del genoma completo usando lecturas cortas de segunda generación sin la necesidad de requisitos de cómputo extensivos. COBASI identifica SNVs a partir de cambios en la cobertura de subcadenas únicas utilizando únicamente alineamientos exactos, y es particularmente útil para determinar con precisión SNVs *de novo* . A diferencia de otros enfoques que utilizan frecuencias poblacionales obtenidas a partir de miles de muestras para filtrar cualquier sesgo metodológico, COBASI puede ser aplicado para detectar SNVs *de novo* en familias aisladas. En este estudio demostramos esta capacidad a través de estudios de simulación y al estudiar un trio utilizando lecturas cortas. La validación experimental de los 58 candidatos de SNVs *de novo* y una seleción de SNVs no *de novo* identificados en el trio, reveló la ausencia de FPs. COBASI está disponible para cualquier investigador como un proyecto de código abietto en https://github.com/Laura-Gomez/COBASI.

ABSTRACT

The precise determination of *de novo* genetic variants has enormous implications across different fields of biology and medicine, particularly personalized medicine. Currently, *de novo* variations are identified by mapping sample reads from a parent– offspring trio to a reference genome, allowing for a certain degree of differences. While widely used, this approach often introduces false-positive (FP) results due to misaligned reads and mischaracterized sequencing errors. In a previous study, we developed an alternative approach to accurately identify single nucleotide variants (SNVs) using only perfect matches. However, this approach could be applied only to haploid regions of the genome and was computationally intensive. In this study, we present a unique approach, coverage-based single nucleotide variant identification (COBASI), which allows the exploration of the entire genome using second-generation short sequence reads without extensive computing requirements. COBASI identifies SNVs using changes in coverage of exactly matching unique substrings, and is particularly suited for pinpointing *de novo* SNVs. Unlike other approaches that require population frequencies across hundreds of samples to filter out any methodological biases, COBASI can be applied to detect *de novo* SNVs within isolated families. We demonstrate this capability through extensive simulation studies and by studying a parent–offspring trio we sequenced using short reads. Experimental validation of all 58 candidate de novo SNVs and a selection of non-de novo SNVs found in the trio indicated the absence of FP calls. COBASI is available as open source at https://github.com/Laura-Gomez/COBASI for any researcher to use.

Capítulo 1

MARCO TEÓRICO

1.1. Importancia de la variación entre individuos

El genoma completo de cualquier organismo es la base para entender su composición genética y poder hacer comparaciones entre individuos. En 2001, el Consorcio Internacional para la Secuenciación del Genoma Humano hizo pública la primer versión del genoma humano [1]; en 2004, el mismo consorcio reportó el termino de la segunda fase del ensamblado del mismo [2]. Debido a los retos y dificultades encontradas en el camino de secuenciar el primer genoma de mamífero, se generaron nuevas tecnologías experimentales y herramientas de análisis. Como resultado se logró ensamblar un genoma de referencia con una tasa de error muy baja, la cual se estimó en alrededor de 1 error cada 100,000 nucleótidos. Además se encontró que el número de proteínas codificadas en el genoma humano está en el orden de 20,000, con un máximo de 25,000 proteínas [2].

Este genoma de referencia ha sido ampliamente utilizado para la comparación de genomas individuales y la generación de amplios catálogos de variación. Se han formado grandes consorcios alrededor del mundo con el objetivo de catalogar el universo de cambios que existen entre individuos de la misma o de distintas poblaciones. El Proyecto HapMap y el Proyecto de los 1000 Genomas son ejemplos exitosos de estos esfuerzos.

Como parte del Proyecto de los 1000 Genomas se han secuenciado 2,504 individuos [3]. Este proyecto descubrió que cada persona contiene en su genoma alrededor de 250 a 300 mutaciones que resultan en la pérdida de función de algún gen y alrededor de 50 a 100 mutaciones implicadas previamente en enfermedades hereditarias [4]. Un hallazgo importante fue que cada población tiene una huella genética única, compartiendo perfiles de variantes raras y comunes entre miembros de la misma población, perfiles que son distintos entre miembros de poblaciones diferentes [5]. Este descubrimiento llevó a ampliar el número de poblaciones muestreadas en el proyecto, generando al final un recurso con el catálogo de variación para 26 poblaciones diferentes, en el cual se incluyen más del 99 % de los polimorfismos con una frecuencia $> 1\,\%$ para al menos una de las poblaciones muestreadas. Con este recurso se pudo concluir que el número de variación promedio entre cualquier individuo y el genoma de referencia depende de la ancestría de dicho individuo, para personas con una ancestría africana se esperan aproximadamente 4.31 millones de polimorfismos de una solo base (SNPs) y alrededor de 625 mil inserciones-deleciones. En el caso de un individuo con ancestría americana, el número de variantes esperadas es menor, estimándose en alrededor de 3.64 millones

de SNPs y en el orden de 557 mil inserciones-deleciones [3].

Algunas de estas diferencias entre individuos se han asociado a rasgos fenotípicos como color de ojos, estatura, color de piel; a capacidades bioquímicas como la capacidad de degradar algún metabolito o compuesto particular; a síndromes o enfermedades complejas como cáncer, obesidad, diabetes o enfermedades cardiovasculares; e incluso a rasgos fenotípicos complejos como la capacidad intelectual.

1.2. Importancia de la variación *de novo*

La variación *de novo* se refiere a aquéllas mutaciones que aparecen por primera vez en un individuo, es decir, mutaciones que se generan en la línea germinal de los padres y que no forman parte del fondo genético somático de los padres. Estudios de secuenciación de genoma completo, han encontrado que la tasa de mutación para SNVs *de novo* en línea germinal se encuentra en el rango de 1,0 a 1,8x10^{-8}, lo cual se traduce en 44 a 82 posibles SNVs *de novo* por individuo, de los cuales únicamente 1 o 2 se encontrarán en regiones codificantes [6].

Algunos estudios han encontrado patrones en la herencia de este tipo de mutaciones. En un estudio en el que se analizaron 78 tríos se encontró que la edad del padre influye en la cantidad de mutaciones *de novo* encontradas en la descendencia, aumentando en 2 mutaciones por cada año adicional del progenitor, los individuos de este estudio tenían una edad mínima de 16 años y una edad máxima de 46 años al momento de la concepción del hijo afectado[7]. Adicionalmente, en un estudio posterior, en el cual se analizaron 693 tríos, se encontró que el número de mutaciones *de novo* observadas en la descendencia también se encuentra correlacionada con la edad de la madre [8], aumentando en 0.51 mutaciones adicionales por año; los individuos de este estudio tenían una edad mínima de 18 años y una edad máxima de 44 años al momento de la concepción. Además, se han descubierto firmas mutagénicas que difieren de acuerdo al padre de origen de la mutación [9]. Adicionalmente, se ha descubierto que existen características genómicas que influencian la frecuencia de las mutaciones *de novo*. Por un lado, existen factores como la conservación de la región y el tiempo durante la replicación, que están inversamente correlacionados con la predisposición de un determinado sitio a presentar una mutación *de novo*; por otro lado, existen factores como la hipersensibilidad a DNasa, la tasa de recombinación y el trinucléotido del sitio particular con los que se observa una correlación positiva; por último el contenido de GC se encuentra positivamente correlacionado si se mide en vantanas de 10 pares de bases y negativamente correlacionado si se mide en ventanas de más de 100 pares de bases [10].

Recientemente se le ha dado más importancia a este tipo de mutaciones debido a que distintos estudios han encontrado una asociación entre las mutaciones *de novo* y distintas enfermedades complejas, incluyendo algunos desórdenes en el desarrollo neurológico como autismo y esquizofrenia [11] [12], además de algunas enfermedades pediátricas como defectos congénitos del corazón [13].

1.3. Secuenciación, la alternativa para conocer el contenido genómico de un individuo

Existe una gran variedad de técnicas que han sido utilizadas históricamente para caracterizar la variación genómica de un individuo en posiciones definidas y previamente conocidas. Desde técnicas moleculares como la identificación de polimorfismos que provocan un cambio en la longitud de fragmentos de restricción (RFLP, por sus siglas en inglés), la amplificación específica de un alelo utilizando PCR en tiempo real, la interrogación de un número definido de posiciones genómicas utilizando microarreglos con sondas de DNA, entre otras.

La primer técnica utilizada para secuenciar un genoma completo fue la Secuenciación por Sanger, este método fue utilizado por el Proyecto del Genoma Humano, el cual logró generar un primer borrador del genoma completo en 2001. Durante este proceso la técnica de secuenciación fue mejorada ampliamente en cada uno de sus pasos, sin embargo, al finalizar el proyecto aún existían dudas sobre su aplicabilidad para la secuenciación masiva dado su bajo rendimiento y alto costo [14]. Sin embargo, con la creación de técnicas de secuenciación masiva que se basan en la paralelización de cada uno de los pasos del proceso de secuenciación se logró disminuir abruptamente los costos y aumentar el rendimiento. Como resultado, cada vez se hace más común secuenciar para identificar variantes genómicas, ya sea a nivel de una fracción definida del genoma, como el exoma o, incluso, a nivel de genoma completo [15] [16].

En términos de tecnologías de secuenciación existen muchas posibles alternativas en el mercado actual las cuales se basan en una gran variedad de principios químicos y diseños experimentales. Las tecnologías más comúnmente utilizadas generan lecturas cortas (desde 35 pares de bases hasta 700 pares de bases) con una tasa de error que va desde el 0.5 % hasta el 0.01 %.

En sentido amplio, las tecnologías que generan lecturas cortas, se clasifican en tres tipos: secuenciación por ligación, secuenciación por síntesis con terminación reversible y secuenciación por síntesis por adición de un solo nucleótido. SOLiD y Complete Genomics utilizan un método de secuenciación por ligación, el cual está basado en la ligación entre una sonda con una o dos bases conocidas y el resto compuesto por bases degeneradas y el templado de DNA; cuando ocurre la ligación un fluoróforo es liberado y de esta manera, se identifican la o las bases siguientes del fragmento de DNA que está siendo secuenciado. En el caso de Illumina y Qiagen el método de secuenciación utilizado es secuenciación por síntesis en la categoría de terminación reversible por ciclo, es decir, cada ciclo de secuenciación se agregan los cuatro nucleótidos marcados cada uno con un fluoróforo diferente y bloqueados en el extremo 3', el nucleótido complementario a la siguiente base en la secuencia blanco es añadido por la polimerasa a la secuencia naciente y el resto es lavado, cada ciclo se toma una foto utilizando las cuatro longitudes de onda diferentes para determinar cuál nucleótido ha sido agregado; en ambas tecnologías, en un paso previo a la secuenciación se generan conglomerados de secuencias idénticas, es importante notar que cada uno de los conglomerados puede añadir un nucleótido diferente en cada ciclo, lo cual permite que estas tecnologías sean sumamente escalables [17] [18]. Por último, 454 y Ion Torrent utilizan secuenciación por síntesis en la

categoría de adición de un sólo nucleótido, en este tipo de secuenciación se utilizan nucleótidos que no están bloqueados en el extremo 3'; durante cada ciclo de secuenciación se añade únicamente un tipo de nucleótido y dependiendo de la tecnología utilizada la incorporación de dicho nucleótido se mide como presencia de luz o como un cambio en pH; una limitante de esta tecnología es que varios nucleótidos del mismo tipo podrían ser incorporados durante un ciclo, tomando en cuenta que se tiene una capacidad restringida de detección, la longitud de algunos homopolímeros podría ser asignada incorrectamente [19] [20].

Recientemente se han desarrollado otro tipo de tecnologías, con las cuales se ha logrado aumentar considerablemente el tamaño de las lecturas a expensas de la calidad de las mismas, generando tasas de error de hasta un 12 % [16], lo cual resulta en que este tipo de secuenciación no se pueda utilizar directamente en la identificación de variantes.

Las tecnologías que generan lecturas largas se pueden clasificar en dos tipos: secuenciación en tiempo real de una sola molécula y secuenciación por enfoques sintéticos; las primeras pueden generar lecturas de hasta 20 mil nucleótidos, mientras que las segundas pueden llegar a producir lecturas de hasta 200 mil nucleótidos, con lo cual se abre la posibilidad de secuenciar un genoma completo pequeño en una única lectura. PacBio y Oxford Nanopore Technology (ONT) utilizan el principio de secuenciación en tiempo real de una sola molécula; en el caso de PacBio, esto se logra al fijar la polimerasa al fondo de un pozo con fondo transparente, a través del cual un lector óptico mide las emisiones de fluorescencia cuando distintos nucleótidos son agregados a la molécula de DNA que está siendo sintetizada. En el caso de ONT, la secuenciación de la molécula de DNA se basa en los cambios de voltaje característicos de cada nucleótido al pasar a través de un poro proteico. Por otro lado, los enfoques sintéticos siguen el siguiente principio: primero, el DNA genómico es fragmentado y el mismo código (barcode, en inglés) es agregado, experimentalmente, a moléculas de DNA que pertenecen a regiones genómicas contiguas; después, todas las moléculas de DNA de un genoma son secuenciadas con tecnologías que generan lecturas cortas, incluyendo en la muestra moléculas de DNA con códigos diferentes; por último, las lecturas que presentan el mismo código son ensambladas generando lecturas sintéticas largas. Las dos tecnologías que existen que utilizan este sistema son la plataforma de Illumina para la secuenciación de lecturas largas sintéticas, y el sistema basado en emulsión 10X genomics [21] [22].

El uso generalizado de las distintas técnicas de secuenciación ha generado una cantidad enorme de datos en los últimos años, un análisis reveló que la cantidad de datos generada es comparable con dominios como la astronomía, Youtube o Twitter, los cuales han sido clasificados como dominios "BigData"[23]. Esto implica que en los tiempos por venir deberán ser desarrolladas tecnologías más eficientes para el almacenamiento, análisis y distribución de los datos genómicos. En mi opinión, esto también implica que los algoritmos de análisis deberán ser cada vez más precisos y menos demandantes de la capacidad de procesamiento.

Cada tecnología de secuenciación es diferente y está basada en un diseño experimental y química distintos, esto se traduce en sesgos de error característicos para cada una de las tecnologías de secuenciación. Sesgos que, además, dependen del contenido de GC del genoma o de la región genómica secuenciada. Por ejemplo, al utilizar Illumina la tasa de error, con

respecto a inserciones y deleciones (indeles) es bastante baja; sin embargo, aumenta considerablemente para regiones con contenido de GC extremos. Por otro lado, la tasa de indeles para Ion Torrent es consistente para distintos contenidos de GC, aunque también es consistentemente mayor que la tasa de indeles de Illumina; sin embargo, la tasa de sustituciones es más elevada en regiones con contenidos extremos de GC. En el caso de Pacific Biosciences, la tasa de deleciones aumenta en regiones con alto GC, mientras la tasa de inserciones disminuye. Complete Genomics, por otro lado, mantiene una tasa alta, aunque consistente, de sustituciones al variar el contenido de GC; y una tasa baja, aunque consistente, de deleciones e inserciones [24].

Además de variaciones sistemáticas en las tasas de error, cada tecnología de secuenciación muestra variaciones sistemáticas en cuanto a la cobertura de distintas regiones genómicas, sesgos que dependen de las características locales de la secuencia y que son diferentes dependiendo de la tecnología utilizada [25].

1.4. El proceso de identificación de variantes

Los métodos que se utilizan para identificar las variantes genómicas a partir de datos de secuenciación pertenecen a una de las siguientes clases. Por un lado están aquéllos que requieren el alineamiento de cada una de las lecturas contra el genoma de referencia durante el cual se le asigna una posición de origen a cada lectura; por otro lado, existe otra clase de algoritmos los cuales no requieren generar estos alineamientos.

1.4.1. Métodos basados en alineamientos

Todos estos métodos utilizan la misma serie de pasos. Para empezar, las lecturas se alinean contra el genoma de referencia asignando una posición a cada una de ellas, este paso se conoce como "mapeo". Una vez que las lecturas han sido mapeadas, se eliminan los duplicados de PCR, lo que significa que las lecturas que pertenecen a exactamente la misma posición genómica son filtradas, este paso es recomendable únicamente si la cobertura promedio del experimento es menor al tamaño de las lecturas. Para mejorar la calidad de las variantes identificadas se realinean las lecturas alrededor de los indeles y se recalibran las puntuaciones de calidad por base; el método utilizado por GATK para estos pasos de recalibración necesita de la existencia de muestras poblacionales para la generación de haplotipos de referencia. Finalmente, se asigna el genotipo de cada variante [26] [27].

Mapeo de las lecturas

La primera etapa en el proceso de identificación de variantes, el mapeo de las lecturas, es uno de los pasos más exhaustivos, y por lo tanto, uno de los más tardados. Dos tipos de algoritmos son populares entre las múltiples herramientas que existen: unos basados en "hashes" y otros basados en "sufijos" en conjunto con la transformada de Burrows-Wheeler

[28].

El principio común a todos los algoritmos basados en "hashes" consiste en buscar una subcadena de tamaño k (kmer) que exista tanto en la lectura como en el genoma de referencia y extender el alineamiento a partir de dicha semilla. El software Bowtie fue uno de los primeros en implementar este principio [29] [30]. A partir de esta base, muchos desarrolladores han implementado modificaciones tratando ya sea de optimizar el algoritmo, reducir los tiempos de cómputo o aumentar la sensibilidad del método. Algunos algoritmos (SSAHA y Stampy) tratan de alinear el resto de la lectura utilizando estrategias de alineamiento global [31] [32]. Mientras que otros (SOAP2, SeqMap) buscan la existencia de un número definido de subcadenas adyacentes en cada lectura, estas subcadenas se consideran semillas espaciadas que serán utilizadas como anclas para fijar el resto del alineamiento, permitiendo recuperar algunas lecturas con errores [33] [34]. Otros algoritmos (SHRiMP2) cortan cada una de las lecturas en todas las posibles subcadenas sobrelapantes de tamaño k y se quedan únicamente con aquellas que contienen más de cierto número de subcadenas que pertenecen a la región genómica que está siendo interrogada [35]. Otros algoritmos se han concentrado en reducir el número de secuencias que deben ser interrogadas en la fase de extensión; por ejemplo, el algoritmo GASST elige inteligentemente las lecturas en las que se llevará a cabo dicha fase; esto lo logra al hacer una comparación del contenido de cada uno de los cuatro nucleótidos de la lectura contra el contenido nucleótidico de la región blanco del genoma de referencia, con lo cual se determina si esa lectura es de interés y, por lo tanto, si se debe proseguir con la fase de extensión [36]. La mayor desventaja de este tipo de algoritmos es que si la semilla utilizada como ancla para determinar la posición de una lectura cae en una región altamente repetida del genoma, entonces la fase de extensión requerirá un gran número de intentos para ser completada.

Los sufijos de una palabra son todas aquéllas subcadenas que empiezan en cualquier posición de la palabra original y terminan en el final de la misma. En un árbol de sufijos para una palabra dada, todos los sufijos de dicha palabra son representados como un camino de la raíz hacia alguna hoja de este árbol. De esta forma, la existencia de una palabra específica (la palabra blanco) se puede buscar rápidamente en la palabra original. Si en el árbol de sufijos de la palabra original existe un camino raíz-hoja que contenga la palabra blanco de interés entonces este blanco existe en la palabra original. La principal ventaja de esta estructura es que colapsa todas las repeticiones de una misma subcadena a un único camino raíz-hoja; sin embargo, la principal desventaja es que sólo permite encontrar alineamientos exactos, es decir, sin sustituciones o errores. El algoritmo MPScan está basado en este algoritmo [37]. Aunque esta estrategia resuelve el problema de las repeticiones, se ha demostrado que los árboles de sufijos son estructuras extremadamente grandes que no pueden ser almacenadas en RAM, y por lo tanto, no pueden ser interrogadas rápidamente, al menos en el caso de la mayoría de los genomas grandes, como el caso del genoma humano.

Finalmente, se ha utilizado otra estructura de datos que colapsa las repeticiones de un genoma y es lo suficientemente compacta, permitiendo su almacenamiento y rápida interrogación, la transformada de Burrows-Wheeler. Para entender la transformada de Burrows-Wheeler es necesario entender primero cómo se forma un arreglo de sufijos. Si todos los sufijos

de una palabra se ordenan alfabéticamente (tomando en cuenta que el carácter que indica el final de la palabra es el último alfabéticamente), y la posición de inicio de cada sufijo es almacenada, entonces se ha creado un arreglo de sufijos. Este arreglo no es más que una lista de posiciones, que va desde 1 hasta el tamaño de la palabra que se está analizando (alrededor de 3mil millones, en el caso del genoma humano) ordenada de acuerdo al orden alfabético de todos los sufijos de esta palabra. Para construir la transformada de Burrows-Wheeler, se debe generar un arreglo circular de sufijos, esto es, para cada índice del arreglo de sufijos se escribe la palabra completa, empezando por la posición indicada en el arreglo, llegando al final de la palabra, continuando por la primer letra y finalizando una antes de la posición indicada en el arreglo: la transformada de Burrows-Wheeler se forma por la última letra de este arreglo circular de sufijos. Aunque no es fácil de ver, utilizando la transformada de Burrows-Wheeler y la palabra original, el arreglo de sufijos inicial se puede navegar igual que se navegaría un árbol de sufijos, permitiendo decir si una palabra blanco está o no presente en la palabra original. Esta estrategia ha sido adoptada por algoritmos ampliamente utilizados como el Alineador basado en la Transformada de Burrows-Wheeler (BWA, "Burrows-Wheeler Aligner"). Además, el algoritmo utilizado por BWA ha propuesto la adaptación de los arreglos de sufijos para poder hacer búsquedas de palabras con errores [38].

Uno de los problemas a los que se enfrenta cualquier algoritmo durante esta etapa del proceso de identificación de variantes, son las lecturas mal mapeadas. La mayoría de los mapeadores resuelven este problema al establecer límites en la calidad de mapeo, lo cual puede llegar a significar que sólo lecturas que mapean a una única posición del genoma sean tomadas en cuenta en pasos posteriores del análisis. Lo anterior, puede producir una baja visibilidad en ciertas zonas del genoma, Por lo tanto, algunos otros algoritmos especializados se han propuesto mejorar la identificación de variantes, tomando en cuenta que algunas lecturas pueden mapear a múltiples posiciones genómicas, y utilizando un modelo probabilístico Bayesiano que explícitamente toma en cuenta la posibilidad de variación multi-locus [39].

Eliminación de lecturas duplicadas

Cualquier muestra que va a ser secuenciada debe pasar por un proceso de preparación, normalmente conocido como preparación de la librería. El primer paso en esta preparación es la fragmentación del DNA, generalmente por sonicación; después, los fragmentos son seleccionados por tamaño y sólo aquéllos fragmentos seleccionados son ligados con adaptadores universales por ambos extremos; estos adaptadores se utilizan como sondas para amplificar todos los fragmentos mediante PCR. Para iniciar el proceso de secuenciación, todos los productos de PCR se unen al dispositivo en el cual sucederá la reacción de secuenciación; este dispositivo podría estar formado por celdas de flujo, perlas o pozos, dependiendo de la tecnología utilizada.

Al final se espera que sólo una copia de cada uno de los fragmentos de DNA que fueron amplificados por PCR se una al dispositivo. Sin embargo, en algunas ocasiones, más de una copia del mismo fragmento se une en varias posiciones y, por lo tanto, es secuenciado múltiples ocasiones. Esto puede resultar en una identificación incorrecta de la variación porque un error producido por la polimerasa durante la amplificación por PCR de este fragmento

se contará varias veces; de la misma manera, un alelo encontrado en este fragmento de DNA duplicado se contará proporcionalmente más veces comparado con cualquier otro alelo que exista en la muestra. Para evitar este tipo de errores, se han incluido módulos para eliminar los duplicados de PCR en distintos algoritmos como el módulo "Markduplicates" en la herramienta Picard o el módulo "rmdup" en el caso de SAMTools [40] [41].

La verdadera contribución de los duplicados de PCR a la asignación de genotipo final, se ha puesto en duda durante algún tiempo. En un estudio reciente se encontró que el 92 % de las variantes son obtenidas sin importar si se eliminan los duplicados o no; además, en cualquier caso, 99 % de las variantes obtenidas son verdaderas [42]. Esto implica que, posiblemente, este paso podría ser omitido de los protocolos actuales resultando en un ahorro de tiempo.

Realineamiento alrededor de indeles

La existencia de inserciones y deleciones en las muestras genómicas puede producir errores en varios pasos del proceso de identificación de variantes. Recordemos que, los algoritmos utilizados para mapear las lecturas se especializan en encontrar ocurrencias de subcadenas idénticas o similares, sin embargo, estos algoritmos no se especializan en hacer el mejor alineamiento posible. Debido a lo anterior, frecuentemente, los alineamientos entre las lecturas que contienen algún indel y el genoma de referencia contienen un gran número de bases no apareadas cerca del indel, las cuales se podrían identificar incorrectamente como SNPs, en vez de atribuirse a un alineamiento incorrecto.

El módulo "realineamiento alrededor de indeles" fue introducido por los creadores del conjunto de herramientas llamado "Genome Analysis Toolkit" (GATK) [43] [27]. En este módulo se busca minimizar el número de cambios totales identificados tomando en cuenta todas las lecturas. De esta manera, aquéllas lecturas que contienen al indel correctamente mapeado ayudan a corregir los alineamientos del resto de las lecturas. Por lo tanto, al finalizar este paso, los alineamientos de la mayoría de las lecturas apoyarán la existencia de un indel consenso, libre de múltiples sustituciones a ambos lados . Recientemente, se han incluido a este conjunto de herramientas algoritmos más sofisticados para el llamado de variantes, así que los desarrolladores en su sitio web indican que este paso "realineamiento alrededor de indeles" ya no será necesario si el llamado de variantes se hace utilizando la técnica de "ensamble de haplotipos".

La identificación correcta de indeles es en sí un problema bajo investigación. Algoritmos especializados se han desarrollado con el único e importante propósito de identificar correctamente inserciones y deleciones. Uno de ellos, Scalpel, examina un conjunto de regiones genómicas determinadas. Para cada región este algoritmo obtiene las lecturas mapeadas y las ensambla utilizando gráficas *de Brujin* independientemente del genoma de referencia utilizado. Al final, estos ensambles *de novo* se mapean con respecto a la referencia utilizando un alineamiento sensible a gaps basado en el algoritmo de Smith-Waterman [44].

Recalibración de puntajes de calidad

Cada secuenciador asigna una calidad a cada una de las bases que llama. Este puntaje de calidad está relacionado con la probabilidad de que esa base se haya llamado incorrectamente. Los valores de calidad se asignan de acuerdo con una escala de Phred, la cual se define como $Q = -10 \log_{10} P$, donde Q es el valor de calidad de cierta base y P es la probabilidad de que la base haya sido llamada incorrectamente; esto significa que un valor de calidad de 10 implica que esa base tiene una probabilidad de 1 en 10 de ser un error, una calidad de 20 implica que esa base tiene una probabilidad de 1 en 100 de ser un error, y así sucesivamente [45].

Otro de los pasos necesarios para mejorar la calidad de las variantes identificadas es el módulo "recalibración de puntajes de calidad por base", también desarrollado por los creadores de GATK [43] [27]. El módulo "recalibración de puntajes de calidad por base" se encarga de buscar sesgos técnicos que el secuenciador no haya tomado en cuenta al momento de asignar calidades. Para lograr esto, el algoritmo necesita de un catálogo de variación esperada, el cual debe de ser provisto por el usuario; en el caso de humano, el proyecto de los 1000 genomas ha generado un catálogo con millones de mutaciones comunes en distintas poblaciones. Tomando en cuenta este catálogo, el algoritmo enmascara todas aquéllas posiciones con variación esperada. Posteriormente, se asume que todas aquéllas posiciones que varían y no han sido enmascaradas se deben a un error por parte del secuenciador. A partir de estos sitios, se entrena un modelo que toma en cuenta la calidad actual de cada base, la posición dentro de la lectura de cada base y la naturaleza específica de la base anterior y la actual (contexto de dinucléotidos). Utilizando estos covariados, el modelo busca patrones en los que el secuenciador comúnmente asigna calidades más altas o más bajas de las que debería. En estas regiones, se recalibran los valores de calidad. Al hacer esto, se disminuye la calidad de las bases donde ésta se ha sobre estimado en un principio, esto es importante porque para estas bases se hubiera tenido más confianza de la que se debería tener; al mismo tiempo, se aumenta la calidad de las bases donde se ha subestimado la calidad, las cuales se hubieran descartado incorrectamente [43].

Asignación de genotipo

Durante este paso del proceso de identificación de variantes, se asigna el genotipo para aquellos sitios en los que se ha detectado variación, es decir, se identifica cuáles alelos existen para cada posición genómica mutante. La asignación de genotipo no es un proceso directo debido a que la cobertura de secuenciación para cada uno de los posibles alelos que forman un genotipo puede ser dispareja; además, el proceso de secuenciación no es perfecto, resultando en cierta fracción de errores. En teoría, estos errores se posicionan al azar, sin embargo, se ha demostrado que existe una predisposición de acumulación de errores en secuencias con cierto contexto genómico [24].

Los primeros algoritmos diseñados para la asignación de genotipo utilizaban una premisa simple basada en que un genotipo heterocigoto debía de tener un porcentaje para el alelo de referencia ubicado en cierto rango; todos aquéllos sitios en los que el alelo de referencia se

encontraba en un mayor porcentaje de lecturas se clasificaban como sitios homocigoto referencia y el resto como sitios homocigoto no referencia [25] [46]. Sin embargo, este método ha demostrado dos limitaciones importantes: primero, tiende a equivocarse en zonas con baja cobertura y; segundo, no genera una medida de la calidad del genotipo asignado.

Varios grupos de investigación se han dedicado a generar algoritmos que resuelven este problema traduciéndolo en un problema de cálculo de probabilidades posteriores. Utilizando este enfoque, se calcula cuál es el probabilidad de que el verdadero genotipo de la muestra para una posición específica sea X dados los datos de secuenciación que se tienen (a este tipo de probabilidad se le conoce como probabilidad posterior o "likelihood", en inglés), es decir, cuál es la probabilidad posterior de que la posición 5935 del cromosoma 1 sea homocigota no referencia, heterocigota o homocigota referencia dado que el alelo de referencia está en 5 lecturas con calidades 30, 35, y 40 y el alelo no referencia está en 3 lecturas con calidades 20, 25, 30 [47]. Además de identificar el genotipo más probable para cierta posición, estos métodos ofrecen la ventaja de generar un valor de calidad para el genotipo asignado, este valor de calidad generalmente corresponde a la división de la probabilidad posterior del genotipo más probable entre la probabilidad posterior del segundo genotipo más probable; de esta forma, el valor de calidad es un indicador de qué tantas veces es más probable el genotipo asignado a cualquier otro genotipo [41].

Existen distintas formas de aumentar la confianza en el genotipo asignado. Por ejemplo, se pueden utilizar las frecuencias alélicas de cierta variante para obtener las probabilidades anteriores de que exista la variante observada. La probabilidad anterior de cada genotipo se puede incluir en el cálculo de la probabilidad posterior. En algunos casos esto podría ayudar, por ejemplo, a elegir entre genotipos equiprobables. Las frecuencias alélicas poblacionales se pueden obtener de bases de datos de variantes conocidas como dbSNP [47], [33], o se pueden obtener interrogando muchas muestras durante el mismo experimento [48]. Otro algoritmo desarrollado recientemente utiliza un método de reconstrucción de haplotipos. En este enfoque, las regiones para las que existe variación son identificadas, las lecturas que fueron mapeadas a estas regiones se ensamblan utilizando gráficos de Brujin, los cuales se utilizan para identificar los posibles haplotipos para cada región. Cada haplotipo se alinea contra el genoma de referencia y los sitios variantes son determinados. Finalmente, se calcula la probabilidad posterior de cada posible haplotipo dados los datos de secuenciación observados, con este propósito cada lectura se alinea contra cada haplotipo. Las probabilidades posteriores de cada haplotipo se utilizan para asignar probabilidades posteriores para cada posible alelo para cada uno de los sitios variantes. A partir de estos datos y utilizando el Teorema de Bayes, se infiere el genotipo más probable para cada sitio variante en cada muestra [27].

Además, datos de ligamiento, como los bloques de desequilibrio de ligamiento y los haplotipos existentes en una población, pueden utilizarse para "adivinar" el genotipo más probable en sitios con genotipo desconocido, a través de un proceso conocido como imputación [49] [50].

1.4.2. Métodos no basados en alineamientos

Cuando se quiere encontrar la variación genómica de una muestra problema, pero no se cuenta con un genoma de referencia de alta calidad o cuando la comparación contra éste se considera innecesaria, las etapas descritas en la sección anterior no pueden ser aplicadas. Para resolver este problema se han desarrollado una batería de algoritmos.

Una primera aproximación consiste en ensamblar *de novo* el genoma problema, hay que notar que si únicamente se tiene el genoma de un individuo, al final de este procedimiento no se podrá hablar de variación *per se*, pero sí de la secuencia genómica ensamblada de dicho individuo. Uno de los algoritmos más ampliamente utilizados para ensamble *de novo* son las gráficas de Brujin. Un grafo está formado por una serie de nodos unidos por vértices. Cuando este tipo de grafos representa un genoma, entonces los nodos representan subcadenas y la existencia de un vértice uniendo dos nodos específicos indica que las dos subcadenas existen como palabras contiguas en el genoma. De esta forma, un camino a través de una gráfica de Brujin representa una palabra que existe en el genoma secuenciado. La mayoría de las estrategias de ensamble *de novo* siguen la siguiente lógica: todas las posibles subcadenas de tamaño k-1 se representan como un nodo, dos nodos se conectan únicamente si en las lecturas existe una subcadena, de tamaño k, formada por el prefijo del primer nodo y el sufijo del segundo nodo, es decir, supongamos dos nodos que representan a las subcadenas ATGC y TGCT, estos dos nodos estarán conectados si en las lecturas existe la subcadena ATGCT. [51]. Al final del procedimiento esperamos obtener un grafo que represente la secuencia genómica completa del organismo secuenciado.

Los ensambles *de novo* se pueden utilizar para identificar variantes entre distintos individuos para especies donde no existe un genoma de referencia o para regiones en las que el genoma es altamente variable, ya sea que existan variantes concatenadas en una región pequeña, como en el caso de la región codificante para el gene del antígeno de leucocitos humanos; o pocas variantes de un gran tamaño, como en el caso de algunas regiones que contienen variación estructural. Incluso, experimentos que han aplicado este método en un gran número de muestras de humano, han encontrado pedazos del genoma, de tamaño considerable, que no existen en la referencia [52].

Otros métodos han utilizado patrones en las frecuencias de subcadenas para identificar mutaciones a partir de las lecturas de un proyecto de secuenciación. Algunos de estos métodos no consideran el genoma de referencia en ningún momento y otros sólo lo utilizan en el paso final para la identificación de la naturaleza específica de la variación.

Varios métodos se han desarrollado bajo la premisa de que existe un catálogo de variación. En uno de ellos, a partir de este catálogo se crea un diccionario formado por subcadenas cortas que contienen la variación y algunas subcadenas aledañas a las primeras. Las lecturas son representadas utilizando la transformada de Burrrows-Wheeler. Utilizando esta representación se cuenta el número de apariciones de las subcadenas que forman parte del diccionario. Se identifican aquellas regiones con un coeficiente de unicidad cercano a 1 y que contengan patrones específicos en la frecuencia de las subcadenas analizadas. Las subcadenas que se espera que sean iguales a la referencia se utilizan como semillas para extender el alineamiento

de esa región contra el genoma de referencia e identificar la variación [53].

En un método alternativo, a partir del catálogo de variación se crea un diccionario compuesto por pares de subcadenas que sobrelapan cada variante a interrogar. Es decir, cada variante está cubierta por k pares de subcadenas (k es el tamaño de las subcadenas), y cada par corresponde al par de alelos alternativos; es importante notar, que dado lo anterior, únicamente se pueden identificar variantes bialélicas. Para cada variante sólo aquellos pares que existen en una posición única del genoma son conservados. El genotipo para cada sitio se obtiene a partir del número de lecturas en las que existe la subcadena que contiene a cada alelo. El número de copias más probable para cada uno de los alelos se identifica utilizando un clasificador empírico de Bayes; los parámetros del modelo se estiman a partir de 100,000 marcadores autosomales para cada individuo analizado [54].

Existen algoritmos que se han enfocado en reconstruir la variación para regiones que son muy distintas a la referencia. Una vez más, los cambios en la frecuencia de subcadenas consecutivas en el genoma de referencia se usan como marcadores de las regiones en las que se debe buscar variación. Para estas regiones, se hace una reconstrucción del haplotipo base por base, de la siguiente forma: a partir de la última subcadena de referencia encontrada en las lecturas en alta frecuencia se agrega cada uno de los posibles cuatro nucleótidos, el nucleótido que produzca la subcadena con el mayor número de apariciones en las lecturas se asigna como la siguiente base en el haplotipo; si existen varias bases que produzcan subcadenas con la misma cuenta, entonces se crean múltiples posibles haplotipos a partir de ese punto. Este procedimiento se repite hasta llegar a la siguiente subcadena idéntica a alguna región del genoma de referencia [55].

Por último, en nuestro laboratorio se han utilizado cadenas únicas del genoma de referencia humano para identificar variantes en la región no pseudoautosómica del cromosoma X de los genomas de Craig Venter y James Watson [56]. En el presente trabajo se extenderá este enfoque para ser aplicado al genoma completo utilizando lecturas cortas producidas por aparatos de última generación.

1.5. Alternativa para optimizar el proceso de identificación de variantes: Cómputo en paralelo

Como se puede observar claramente en la sección anterior, el proceso de identificación de variantes es un proceso largo y complejo, requiriendo gran cantidad de algoritmos y recursos computacionales. Distintos proyectos se han interesado en las alternativas para agilizar este proceso. Una propuesta es la recodificación de los algoritmos existentes en una arquitectura de software que aproveche las ventajas del cómputo en paralelo; bajo esta premisa se han logrado desarrollar algoritmos que paralelizan el proceso de asignación de genotipo o el proceso de análisis de cobertura a partir de las lecturas mapeadas [57]. Como una solución más general, se ha propuesto utilizar una arquitectura de software que permita utilizar los recursos computacionales más eficientemente y así disminuir el tiempo de cómputo necesario; la solución propuesta consiste en utilizar procesamiento en la nube, utilizando los algoritmos existentes pero dividiendo el genoma en regiones genómicas pequeñas y fijas, dividiendo cada

proceso en cientos o, incluso, en miles de subprocesos más pequeños y rápidos [58].

1.6. Métodos utilizados para identificar variación *de novo*

Debido al bajo número de mutaciones *de novo* que existen en un genoma, el problema de identificar este tipo de mutaciones se podría comparar con "encontrar la aguja en el pajar". La existencia de errores de secuenciación y la cobertura de secuenciación variable a lo largo del genoma, hacen que el problema sea aún mas complicado. Una de las primeras aproximaciones utilizadas para resolver este problema estaba basada en establecer probabilidades posteriores mínimas asociadas al genotipo asignado para cada uno de los individuos de la familia problema; por ejemplo, para la descendencia el genotipo heterocigoto debía ser al menos diez órdenes de magnitud más probable que el genotipo homocigoto referencia, y para ambos padres el genotipo homocigoto referencia debía ser al menos dos órdenes de magnitud más probable que el genotipo heterocigoto [7]. Aunque esta solución funciona, es demasiado simple para resolver la complejidad del problema, tendiendo a equivocarse en regiones de baja cobertura.

Algunos algoritmos especializados han sido creados para resolver específicamente este problema. El algoritmo *PhaseByTransmission*, integrado en GATK, calcula la probabilidad posterior para cada posible combinación de genotipos entre todos los miembros de una familia, esto lo hace tomando en cuenta los datos de secuenciación de todos los individuos al mismo tiempo. Para esto, el algoritmo considera: los datos de secuenciación de todos los individuos, las relaciones familiares conocidas y una tasa de mutación conocida; además, este método se beneficia al utilizar probabilidades anteriores para cada genotipo, calculadas a partir de las frecuencias alélicas observadas en todas aquellas muestras no relacionadas [59]. Por otra parte, *SNVSniffer*, utiliza un modelo probabilístico Bayesiano para identificar mutaciones germinales, este modelo asume que las cuentas alélicas para los sitios con este tipo de mutaciones siguen una distribución condicional multinomial, el modelo identifica el genotipo más probable por medio del cálculo de probabilidades posteriores [60].

En aquellos análisis en los que se buscan específicamente las diferencias entre dos individuos o entre tipos celulares del mismo individuo, las comparaciones con la referencia obtendrán muchas diferencias que, sin embargo, son compartidas entre las muestras de interés. Por lo tanto, algunos algoritmos se han enfocado en encontrar la variación entre muestras sin tener que usar la referencia en el análisis. Un método desarrollado recientemente se centró en identificar las subcadenas de un tamaño definido que existen en la muestra de interés pero que no existen en la o las muestras relacionadas. Una vez identificadas estas subcadenas, se ensamblan las lecturas que las contienen y se generan contigs que contienen a la variación. Finalmente, estos contigs son alineados contra el genoma de referencia para identificar la naturaleza de la variación. Esta estrategia ha sido aplicada para encontrar los SNVs *de novo* en un trío familiar y para localizar mutaciones entre células somáticas y cancerosas obtenidas del mismo paciente [61].

1.7. Comparación del rendimiento de distintos métodos utilizados para la identificación de variantes genéticas

Derivado de la gran explosión en el uso de la secuenciación, ha habido un incremento importante en el número de técnicas experimentales que utilizan la secuenciación para generar datos y, por lo tanto, en el número de métodos que se encargan de analizar los datos generados. Consecuentemente, también ha ido en aumento el interés por medir el rendimiento de los métodos existentes.

Respecto a la identificación de variantes, un estudio encontró una gran falta de correspondencia entre los resultados obtenidos con distintos métodos. En este estudio se analizaron los resultados para 5 combinaciones de mapeadores y genotipificadores. El estudio encontró que había una intersección de únicamente el 57.4 % de los SNVs identificados, mientras que entre un 0.5 a 5.1 % de los SNVs identificados fueron encontrados únicamente por una de las combinaciones de software probadas. En el caso de los indeles los resultados fueron aún más impactantes, encontrando una concordancia de únicamente el 26.8 % entre las técnicas de análisis utilizadas [62].

Debido a los resultados obtenidos por el trabajo previamente descrito, y a otros trabajos con resultados similares, se determinó la importancia de la generación de estándares, de métricas cuantitativas para la comparación de los resultados, y de la existencia de una infraestructura que permitiera la generación de resultados reproducibles y comparables. Como parte de estos esfuerzos, el Instituto Nacional de Estándares y Tecnología (NIST, The National Institute of Standards and Technology) se dio a la tarea de generar dicho recurso, el cual se llamó “Genome in a Bottle” (GiaB). La muestra piloto agregada a este proyecto, corresponde a un individuo secuenciado, en un principio, por el Proyecto de los 1000 genomas (1000HGP): el individuo NA12878. El proyecto GiaB recibió más de 8000 alícuotas de esta muestra en 2013. El proyecto GiaB también almacena muestras correspondientes a un trío familiar de la población de judíos Ashkenazi y a un trío familiar con ascendencia china, muestras derivadas de 1000HGP [63]. Como parte del esfuerzo por generar materiales de referencia, el proyecto GiaB analizó la muestra piloto, NA12878, utilizando cinco tecnologías de secuenciación, siete alineadores y tres genotipificadores; a partir de estos datos se generó un conjunto de variantes de alta calidad que se ha utilizado como “gold standard” en varios estudios comparativos [64].

Un primer estudio dio a conocer algunas métricas de rendimiento comparando algoritmos comúnmente utilizados para la identificación de variantes. En este estudio se secuenció el exoma de la muestra NA12878, correspondiente al proyecto GiaB, con cinco teconologías Illumina diferentes. Las variantes se llamaron utilizando seis alineadores (Bowtie2, BWA mem, BWA sampe, CUSHAW3, MOSAIK y Novoalign) y cinco genotipificadores (FreeBayes, GATK HaplotypeCAller, GATK Unified Genotyper, SAMtools mpileup, SNPSVM). Las variantes identificadas se compararon contra el conjunto de variantes “gold standard” definido en el proyecto GiaB y se calcularon dos métricas de rendimiento: el Valor Positivo Predictivo (PPV) o precisión, el cual describe la fracción de las variantes identificadas que son reales y se calcula utilizando la siguiente expresión: [TP/(TP + FP)], donde TP denota el número de

Verdaderos Positivos y FP denota el número de Falsos Positivos; y, por otro lado, la sensibilidad, la cual describe cuántas de las variantes que existen fueron identificadas, y se calcula de la siguiente manera: [TP/(TP + FN)] donde FN denota los Falsos Negativos. Los resultados mostraron valores altos de precisión, los cuales iban desde un 80.69 % hasta un 99.92 %; sin embargo, la sensibilidad se mantuvo en valores bajos alcanzando un mínimo de 35.79 % y un máximo de 50.85 %. El número esperado de SNVs para NA12878 es de 34,886, los resultados de este estudios mostraron que, independientemente de la combinación de algoritmos utilizada, el número máximo de SNVs recuperado fue de 22,324. Además, en este estudio se reporta una concordancia del 70 % entre los distintos algoritmos utilizados [65]. Por último, los autores identificaron a *BAW-mem* como el mejor alineador y *GATK UnifiedGenotyper* como el mejor genotipificador . Estos resultados son altamente controversiales, debido a que en un estudio más reciente se obtuvieron rendimientos muchos más optimistas.

En el estudio mencionado al final del párrafo anterior se comparó el desempeño de tres alineadores (BWA mem, Bowtie2 y novoalign) y cuatro genotipificadores (FreeBayes, GATK HaplotypeCAller, SAMtools mpileup y el software de Ion Proton (TVC)]; además, en este estudio se incluyó una variedad de muestras secuenciadas con distintas plataformas de secuenciación (Illumina2000, Illumina2500 y Ion Proton) y generadas a partir de distintas técnicas de captura. En este estudio se utilizó el área bajo la curva de una curva de precisión-sensibilidad (APR) para medir el desempeño de cada combinación de softwares utilizada. Varios resultados interesantes se derivan de este estudio: se observó que la mejor combinación de algoritmos varía dependiendo de la técnica utilizada para generar los datos; se concluyó que el genotipificador utilizado tiene un mayor efecto en el desempeño global que el alineador; se observó un nivel de concordancia del 92 % entre los tres genotipificadores, sin embargo este nivel de concordancia varía dependiendo de los datos analizados alcanzando un mínimo del 82 % y un máximo del 97 %; además, se determinó que cada genotipificador tiene sesgos característicos hacia un tipo de error específico, por ejemplo, *Freebayes* se equivoca más comúnmente en SNPs genotipificados como homocigotos, por otro lado *GATK HaplotypeCaller* y *Samtools* se equivocan más frecuentemente cuando determinan que un sitio es heterocigoto; por último, en el caso de la identificación de SNVs, los valores de APR que reportan varían desde 0.93 hasta 0.99 [66]. Los autores aseguran que la diferencia entre los resultados obtenidos por ellos y los obtenidos en el estudio anterior [65] se debe a las distintas versiones de los software utilizados y a que en este estudio se incluyeron muestras obtenidas con distintas plataformas de secuenciación y técnicas de captura.

Como se ha descrito en los párrafos anteriores, los recursos del proyecto GiaB han sido ampliamente utilizados por la comunidad científica. Sin embargo, en este trabajo no fue posible descargar el conjunto de datos “gold standard”. Aunque se intentó descargar varias veces y contactar a los autores, el número de variantes descargadas no correspondía con lo reportado en la literatura. Por lo tanto, se utilizaron experimentos de simulación como principal herramienta para medir el rendimiento de nuestro software y compararlo con los software comúnmente utilizados.

Capítulo 2

OBJETIVOS

2.1. Objetivo general.

El objetivo de este trabajo es desarrollar un método que identifique SNVs *de novo*, reduciendo al máximo la cantidad de FPs basándose únicamente en alineamientos perfectos y mapas de cobertura generados a partir de estos alineamientos.

2.2. Objetivos particulares.

- Determinar la factibilidad de utilizar mapas de cobertura, generados a partir de datos de secuenciación de última generación, para identificar SNVs en organismos diploides.
- Desarrollar un método capaz de identificar SNVs en un individuo a partir de los mapas de cobertura generados.
- Desarrollar un método, basado en el algoritmo anterior, para identificar la variación *de novo* en un trío padres-descendencia.
- Cuantificar el rendimiento del algoritmo desarrollado y comparar con algoritmos comúnmente utilizados en la comunidad científica.

Capítulo 3

RESULTADOS

La mayoría de los métodos que buscan encontrar variantes genómicas a partir de datos de secuenciación, inician por asignarle a cada una de las lecturas obtenidas una ubicación dentro del genoma de interés. Para lograr esto, se debe confiar en alineamientos que contienen cierto grado de discrepancias entre la lectura y la región genómica asignada. En este trabajo se demuestra que alineamientos de secuencias totalmente idénticas son suficientes para identificar variantes entre dos genomas, y para detectar precisamente SNVs *de novo*.

Los resultados principales de este trabajo se describen en el artículo "Precise detection of de novo Single Nucleotide Variants in human genomes" publicado en la revista PNAS el 7 de Mayo de 2018. El artículo completo se puede consultar en la presente sección. El Material Suplementario asociado a dicho artículo se encuentra en el Apéndice A.

3.1. Gómez-Romero, Laura, et al. "Precise detection of de novo single nucleotide variants in human genomes."Proceedings of the National Academy of Sciences 115.21 (2018): 5516-5521

PNAS

Precise detection of de novo single nucleotide variants in human genomes

Laura Gómez-Romero[a,1], Kim Palacios-Flores[a,b], José Reyes[a], Delfino García[a], Margareta Boege[a,b], Guillermo Dávila[a,b], Margarita Flores[a,b], Michael C. Schatz[c,d], and Rafael Palacios[a,b,1]

[a]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, 62210 Morelos, México; [b]Laboratorio Internacional de Investigación Sobre el Genoma Humano, Universidad Nacional Autónoma De México, Juriquilla, 76230 Querétaro, Mexico; [c]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; and [d]Departments of Computer Science and Biology, Johns Hopkins University, Baltimore, MD 21211

Contributed by Rafael Palacios, April 10, 2018 (sent for review February 7, 2018; reviewed by Ludmil B. Alexandrov and George Weinstock)

The precise determination of de novo genetic variants has enormous implications across different fields of biology and medicine, particularly personalized medicine. Currently, de novo variations are identified by mapping sample reads from a parent–offspring trio to a reference genome, allowing for a certain degree of differences. While widely used, this approach often introduces false-positive (FP) results due to misaligned reads and mischaracterized sequencing errors. In a previous study, we developed an alternative approach to accurately identify single nucleotide variants (SNVs) using only perfect matches. However, this approach could be applied only to haploid regions of the genome and was computationally intensive. In this study, we present a unique approach, coverage-based single nucleotide variant identification (COBASI), which allows the exploration of the entire genome using second-generation short sequence reads without extensive computing requirements. COBASI identifies SNVs using changes in coverage of exactly matching unique substrings, and is particularly suited for pinpointing de novo SNVs. Unlike other approaches that require population frequencies across hundreds of samples to filter out any methodological biases, COBASI can be applied to detect de novo SNVs within isolated families. We demonstrate this capability through extensive simulation studies and by studying a parent–offspring trio we sequenced using short reads. Experimental validation of all 58 candidate de novo SNVs and a selection of non-de novo SNVs found in the trio confirmed zero FP calls. COBASI is available as open source at https://github.com/Laura-Gomez/COBASI for any researcher to use.

human genome variation | genomic algorithms | de novo mutations | genomic landscape | coverage map

The identification of variations among genomes is the starting point for a diversity of projects to understand human health and disease. It is such an important step that several large international consortia have been established, such as the HapMap Project (1, 2) and the 1000 Genomes Project (3, 4), to catalog variations among different healthy human populations, as well as several large consortia to examine genetic variations associated with different diseases, such as the International Cancer Genome Consortium (5) and the Cancer Genome Atlas Project (6) to identify variations between normal versus cancer cells. A particularly important type of variation, de novo variants, are those variants that occur spontaneously between parents and children, and have been implicated in a variety of diseases, such as autism, intellectual disabilities, and schizophrenia (7–9).

Several bioinformatic pipelines have been developed to identify single nucleotide variants (SNVs). Most of these begin by mapping sequencing reads from the sample to the reference genome (RG), allowing some number of mismatches or indels using one of a number of short-read aligners [Burrows–Wheeler aligner (BWA), Bowtie, etc.] (10). A mapping quality score is reported to reflect the probability of the read being correctly mapped. The mapped reads are then used to make genotype assignments using computational tools, such as SAMtools (11) or Genome Analysis Toolkit (GATK) (12), which evaluate the alignment of reads at every position along the genome and assign a confidence score to indicate the probability of the existence of a variant. This is achieved using statistical inference algorithms, which are necessary because imperfect alignments create uncertainty about the position assigned to each read and sequencing errors can induce false variants (11, 12). Various correction steps, such as around-indel realignment or quality recalibration, have been proposed to correct for common artifacts. However, most of these steps require a database of known variants (13). Finally, to correctly assign each genotype, the likelihoods for each possible genotype are calculated based on the observed data, modeling both alignment accuracy and sequencing accuracy. Different scoring schemes have been used to compute the probability that the read has been correctly mapped (14) and the genotype has been correctly assigned to ultimately indicate the overall confidence in the results. Additionally, some pipelines specialized for finding de novo variants incorporate stringent filtering based on each individual genotype likelihood (15–17). These pipelines also often use population-specific samples to identify and filter out any methodological bias (15–17, 18) or they require a predetermined de novo mutation rate and population-specific allelic frequencies

Significance

The precise location of variants in the human genome is of utmost importance. We present a unique approach, coverage-based single nucleotide variant (SNV) identification (COBASI), which uses only perfect matches between the reads of a sequence project and a reference genome to detect and accurately identify de novo SNVs. From the perfect matches, a representation of the read coverage per nucleotide along the genome, the variation landscape, is generated. SNVs are then pinpointed as significant changes in coverage and de novo SNVs can be identified with high precision. The performance of COBASI was analyzed using simulations and experimentally validated by sequencing de novo SNVs identified from a parent–offspring trio. We propose this pipeline as a useful tool for different genomic applications.

Author contributions: L.G.-R., K.P.-F., J.R., M.B., G.D., M.F., M.C.S., and R.P. designed research; L.G.-R., K.P.-F., and M.F. performed research; L.G.-R. contributed new reagents/analytic tools; L.G.-R. and D.G. analyzed data; and L.G.-R., M.C.S., and R.P. wrote the paper.

Reviewers: L.B.A., Los Alamos National Laboratory; and G.W., The Jackson Laboratory.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: lgomez@lcg.unam.mx or palacios@liigh.unam.mx.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1802244115/-/DCSupplemental.

to calculate the probability of the called de novo variant being a false positive (FP) (19, 20).

These methods are needed to overcome an apparent paradox: when sequence reads are aligned to a reference genome, some degree of mismatch must be tolerated, since variation would not be detected by using only perfect alignments. On the other hand, because of the highly repetitive and complex structure of the human genome, the tolerance of mismatches could result in the misplacement of some reads, introducing false variants. Our group has addressed this paradox by applying a different approach to the problem of detecting SNV's in human genomes called context-dependent individualization of nucleotides and virtual genomic hybridization (COIN-VGH) (21). It is based on perfect alignments of unique substrings of a specific size (k; kmers) of the sequencing reads to the reference genome. As a proof of concept, the COIN-VGH approach was previously used to identify SNVs in a haploid region (nonpseudoautosomal region of the chromosome X) of Craig Venter's and James Watson's genomes using the same Sanger or 454 sequencing data as in the original studies (22, 23). Despite the success in eliminating false-positive calls over alternative approaches, COIN-VGH has important limitations for its widespread use: (*i*) it can only be used in haploid regions of the genome, (*ii*) it requires relatively long reads, and (*iii*) the algorithm is time consuming and utilizes a large amount of random-access memory (RAM) and disk storage.

Addressing these issues, we have developed a unique approach, called coverage-based single nucleotide variant identification (COBASI). COBASI builds on the original COIN-VGH approach but can be used to call variants from both haploid and diploid regions of the human genome and works with 30× or greater fold coverage (it has been used in datasets with as much as 100× fold coverage) of second-generation short sequence reads. In addition to circumventing the previous limitations of COIN-VGH, the approach is particularly suited to identify de novo SNVs through the joint analysis of a parent–offspring trio sequencing data. To evaluate COBASI, we first apply it to a diverse collection of simulated sequencing data and show that its performance is similar or superior to alternative approaches. We next apply it to the whole genomes of a parent–offspring trio we sequenced using Illumina sequencing and identified de novo SNVs across the entire child genome. From this, we discover 58 de novo SNVs, and all predicted de novo SNVs were experimentally confirmed as correct (zero false positives). Furthermore, the computing time and resources required for the bioinformatics pipeline have been significantly reduced, allowing for its routine application over many human datasets or other large mammalian datasets with a high-quality reference genome. Thus, COBASI is a powerful tool to systematically scan genomes for regions of interest for a broad range of applications.

Results

Rationale of the COBASI Approach. When a single specific nucleotide is searched along the genome, the position to which it belongs cannot be unambiguously determined. If two adjacent nucleotides are incorporated into the search, the set of possible locations is reduced, although it remains quite large. At some point, however, the context of the target nucleotide will contain enough information to unambiguously determine its unique origin position (Fig. 1*A*). In our previous research, we defined COIN-Strings (CSs) as the set of all overlapping sequences (with a one-nucleotide sliding window) from the reference genome of a specific size (k) that are uniquely localized. Thus, each nucleotide along the reference genome is contained in, at most, k CSs.

COBASI extends this analysis of CSs to robustly find variations in the sample across the entire genome. When a SNV is present in a sample at a particular position X, it is expected that about half the reads for heterozygous SNVs, or nearly all of the reads in homozygous SNVs that overlap with X will contain th SNV. Accordingly, the CSs that include X will be present only i the reads that do not contain the alternative allele. This can b translated into specific patterns that are designated as variatio signature regions (VSRs) (Figs. 1 *C* and *D* and 2*A*). Once can didate regions are identified, local alignments between the read and the genome at the regions of interest will uncover the natur of the specific variants.

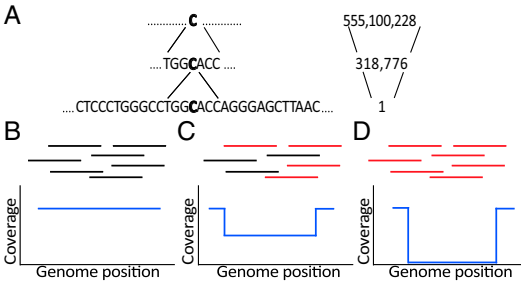


Fig. 1. Rationale of the COBASI approach. (*A*) A specific nucleotide (larg bold C) cannot be uniquely localized along the genome until its context i included in the search. (*Left*) The string to be searched; (*Right*) the numbe of positions at which such a string is found. The bottom string is a COIN String (CS) of 30 nt. (*B–D*) (*Upper*) Schematic representation of sequenc reads. (*Lower*) Specific regions of variation landscapes (VLs) for three sce narios. (*B*) No variation signal. (*C*) A heterozygous SNV variation signal. (*D*) A homozygous SNV variation signal. Black lines in *B*, *C*, and *D* represent read from the genome project that contain the reference allele. Red lines rep resent reads from the genome project that contain the SNV allele. Th sections of the VL in ref. 2 are represented by blue lines. The *x* axis indicate the genome position for every CS start. The *y* axis indicates the number c reads containing the CS sequence starting at that position.

De Novo SNV Discovery Using the COBASI Pipeline. Based on th rationale presented, we designed and implemented a strategy t detect de novo SNVs from a parent–offspring trio. First, all o the CS positions from the reference genome are computed. W define the COBASI-accessible genome as regions at least 100 b long for which at least 50% of the kmers starting inside the re gion are CSs using k = 30 bp. Even though more than 50% of th human genome is classified as repetitive sequences (24), the vas majority (around 84%) of the genome can be interrogated usin COBASI (*SI Appendix*, Table S1).

Next, all of the SNVs from the child individual are identifie by analyzing the variation landscape (VL). The VL is a repre sentation of the number of reads that contain each CS sequenc (coverage) along the whole genome (Fig. 2*A*). To magnify th difference in coverage between two adjacent CSs, the VL wa transformed into a relative variation landscape (RVL) using relative coverage index (RCI), measured on a scale from −1 t +1 (Fig. 2*B*). Under this formulation, the RCI is close to zer when there is little to no difference in coverage, and its absolut value approaches 1 when abrupt differences occur, most ofte because of underlying genetic variation (Fig. 2*C*). Since the RV is variable in low-coverage regions, a coverage threshold wa established to avoid noise in the VSR identification proces (*Materials and Methods*).

From the RVL, the VSRs can be identified spanning an candidate mutations. We define the last CS before the start of VSR as PrevCS, and define the first CS after the end of a VSR a PostCS, and both of these CSs we call signature CSs. Next, read containing perfect matches to the signature CSs are identifie and global alignments between the corresponding region in th reads and the genome are computed. Finally, the variant nucleotide

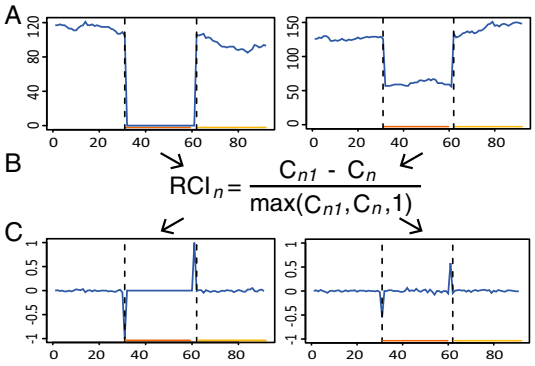


Fig. 2. Variation landscape transformation into a relative coverage landscape. (*Left*) A homozygous SNV is shown. (*Right*) A heterozygous SNV is shown. (*A*) The VL for a region composed of 30 nt upstream and 30 nt downstream of each VSR is shown. The plots show the start position of each CS in that genomic region (*x* axis) and the coverage for each CS (*y* axis). (*B*) The VL is turned into the RVL using the RCI. RCI_n refers to the relative coverage index for nucleotide n. C_n and C_{n1} denote the number of reads that contain the CS starting at nucleotide n and the next downstream CS, respectively. (*C*) The RVL for the same regions shown in *A*. The plots show the start position of each CS (*x* axis) and RCI values associated with each CS (*y* axis). The VL and the RVL are represented by blue lines. The PrevCS and PostCS are shown as orange and yellow lines at the *Bottom* of each plot, and their start positions are highlighted with dashed black vertical lines (*SI Appendix*, Fig. S1).

in the reads are highlighted in the local alignment to identify the specific SNV (Fig. 3). Since CSs are guaranteed to be unique in the genome, and only perfect matches are considered, no other quality filters are required.

To discover the de novo SNVs, variable positions in the child are next interrogated in the parents. For each SNV in the child, its signature CSs were used as anchors to retrieve the reads of interest in the parents. Those reads from the parents are then aligned to the RG using the above procedure. A catalog containing all of the child SNVs and the alleles found in each parent for the same positions is then generated. The genotypes for each individual are assigned and compared, so that candidate de novo SNVs can be identified (Fig. 4). We considered as bona fide de novo variants those not found in either parent in more than one alignment containing both signature CSs, which are considered as high-quality alignments.

Performance of COBASI by Simulation Experiments. We first evaluated COBASI relative to the most commonly used pipelines through simulation experiments considering several different sequencing depths, kmer sizes, and other internal parameters (*SI Appendix*, *SI Materials and Methods*). Mutations were introduced into one human diploid chromosome (chromosome 12), simulated reads were produced, and SNVs were called using COBASI. We quantified the performance using the widely used area under the precision-recall (AUPR) curve statistic.

The best performing parameters were derived from the simulation experiments. Over all of the tested sequencing depths, the best kmer size was 30, and the best ratio between the coverage of both signature CSs was 2.0. This maintained a low number of FPs while not significantly increasing the false negatives (FNs). Values of 0.2 or 0.3 for the RCI threshold had very similar AUPR scores. In contrast, the best value for other key parameters depended on the sequencing depth. If the sequencing depth was 35×, the minimum coverage for the signature CSs was 5, the optimal extension for alignments that contain only the PrevCS was 5 bp, and the minimum number of alignments with both CSs was 2. If the sequencing depth was 100×, the minimum coverage for the signature CSs was 10, the optimal extension for alignments that contain only the PrevCS was 5 bp or 10 bp, and the minimum number of total alignments with both CSs was 3 or 4. Once the best performing parameters were identified, the AUPR ranged from 0.94 to 0.96. To compare COBASI performance with the performance of the most commonly used variant-calling pipeline, the SNVs were also called from the simulation experiment with a sequencing depth of 100× using a combination of BWA, Picard Tools, and GATK. The AUPR was 0.99, while the AUPR obtained for COBASI was 0.96. However, the time required to obtain a list of SNVs from raw sequencing data was incredibly reduced, from more than 30 h in the case of the standard alignment-based pipeline to less than 6 h required by COBASI.

Besides, in a previous study, Hwang et al. measured the performance for any combination of three different mappers and

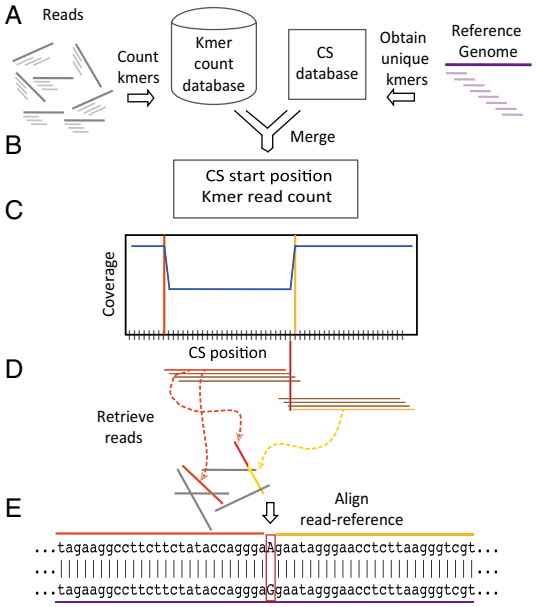


Fig. 3. The COBASI experimental pipeline for SNV discovery in one individual. (*A*, *Left*) Every overlapping 30-nt kmer (with a sliding window of 1 nt) along each of the reads of the sequencing project is obtained (only 3 kmers are shown per read). The counts for every kmer are stored in a database. Reads and read kmers are shown as gray and light gray lines, respectively. (*A*, *Right*) CS along the RG is obtained, and the start and end positions of all nonoverlapping unique regions is stored. RG and RG kmers are shown as purple and light purple lines. (*B*) The two virtual products are merged and the variation landscape (VL) is generated. (*C*) A region of the VL containing one heterozygous SNV is presented. The plot shows the start position of each CS along the genome (*x* axis) and each CS coverage (*y* axis). The VL is represented as a blue line. The VL is transformed into the RVL. Only the VL is depicted. The start position of the PrevCS and the PostCS are indicated by vertical orange and yellow lines, respectively. The PrevCS and PostCS are represented by horizontal orange and yellow lines, respectively. Some interCSs are shown as horizontal brown lines. The position of the SNV is shown as a red vertical line. All CSs located between the Prev- and PostCSs (interCSs) contain the SNV position. (*D*) The Prev- and PostCSs (signature CSs) are used as anchors to retrieve all of the reads of interest (*Materials and Methods*). (*E*) Each of the retrieved reads is then aligned with the corresponding region of the RG. An aligned read-RG region is shown. The SNV position and specific nucleotide is highlighted in a red rectangle.

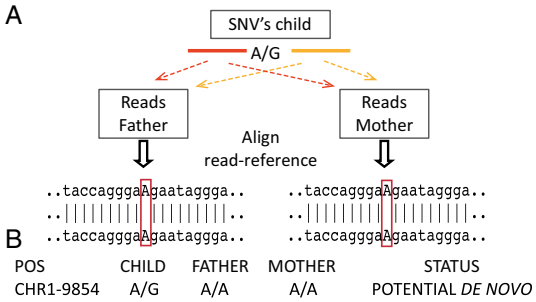


Fig. 4. The COBASI experimental pipeline for SNV discovery in a family-based framework. (*A*) For each SNV in the child, its signature CSs are used as anchors to retrieve the corresponding reads in the parents. The reads are then aligned to the RG. (*B*) A catalog containing all child SNVs and the alleles found in each parent at the same positions is generated. The three genotypes are then compared, and the possible de novo SNVs are identified.

three different callers for any of 11 datasets (10). In most cases, the AUPR for COBASI was similar to previously reported AUPRs, even though Hwang et al. used only exome data (about 2% of the genome) and COBASI was tested on the whole callable genome (about 84% of the genome) (*SI Appendix*, Tables S2 and S3).

We next measured the performance of de novo SNV discovery by COBASI using parent–offspring trio simulations. A trio of parent–offspring genomes was created following Mendelian inheritance along with a limited number of de novo variants (with a median of 35 de novo SNVs per simulation) (*Materials and Methods*), from which sequencing data were simulated. The sequencing depth was chosen to resemble our experimental sequencing data: 35× for the parents and 100× for the child. The de novo SNVs were then called using COBASI. The experiment was repeated five times, so that robust median accuracy values could be computed. The median precision obtained was 1.0 and the median recall was 0.91 with a median of 32 true positives (TPs), 3 FNs, and 0 FPs.

As with any variant detection pipeline, sufficient sequencing coverage is required to accurately detect mutations. To examine this for COBASI, we plotted the precision-recall curve ordered by the available coverage, defined as the number of alignments that contain the variant. The median AUPR across all coverage values was 0.86. However, most of the errors were found in low coverage variants, and with a reasonable coverage level (>10 reads), the median precision and recall for de novo simulations were 1.0 and 0.91, respectively. In one individual experiment, the precision and recall at the same coverage threshold were 0.9999 and 0.9613, respectively. Thus, the de novo discovery pipeline was more precise than the whole-genome pipeline at the expense of a small degree of sensitivity. Using the same simulated data, the de novo SNVs were called using the standard practices of the most commonly used alignment-based pipeline, resulting in an AUPR of 0.91. Thus, the COBASI performance can be compared with state of the art pipelines reducing the time required to complete the variant-calling process.

COBASI Application in a Family-Based Framework. We next applied the de novo discovery COBASI pipeline to find genome-wide SNVs in a parent–offspring trio we sequenced using Illumina sequencing (*Materials and Methods*). Here we used the best performing parameters determined from the simulation experiments. Additionally, we considered as bona fide de novo variants those not previously reported in public databases, such as dbSNP, since the probability of two independent individuals having a de novo mutation event at the same nucleotide is very low (*SI Appendix*, *SI Materials and Methods*). Using these parameters, we found 2,912,889 SNVs in the discovery individual and 58 de novo variants (Fig. 5).

The 58 de novo SNVs and a selection of two randomly chosen SNVs per chromosome (46 random variants total) identified in the child were selected for experimental validation via PCR and Sanger sequencing. In the case of the de novo variants, for five cases no PCR product could be obtained and one case could not be properly sequenced. For all 52 de novo mutations that could be sequenced, the Sanger sequencing confirmed that each predicted SNV represented a real de novo variant. *SI Appendix* Table S4 presents the genomic coordinates, the genotype for each individual, and the results of the experimental validation for every de novo SNV. *SI Appendix*, Fig. S2 presents the experimental validation for each individual of the family trio for 10 de novo variants, chosen at random. All of the 46 Mendelian variants were successfully validated (*SI Appendix*, Fig. S3 and Table S5) (five examples).

Discussion

To find de novo SNVs in sequenced genomes, the COBAS approach represents a fast and precise solution to the variant calling problem. It is based on the concept that by using only perfect matches of unique substrings to a reference genome variation can nevertheless be found with great precision. In this study, we used unique DNA strings of 30 nucleotides, which can interrogate about 84% of all of the base pairs of the complete reference genome. Importantly, this percentage was calculated to include all repetitive sequences, such as low-complexity regions and segmental duplications of high identity. Larger strings would identify a greater percentage of the genome, although this will become more sensitive to any sequencing errors in the reads.

The VL constructed in the first stages of our approach represents a powerful tool to pinpoint regions of polymorphisms by identifying abrupt changes in local coverage. Moreover, these sharp differences were proven to be robust to noisy coverage fluctuations found in any sequencing project. The VL is generated in a fast, computationally efficient process and represents a comprehensive description of the read coverage across the genome at a single-nucleotide resolution.

The identification of de novo variants is a particularly challenging task because any false-positive calls in the child or any false-negative calls in the parents result in a variant incorrectly identified as de novo. To address this challenge, several specialized algorithms that analyze sequence data for all family individuals have been proposed. These algorithms rely on a prior probability of de novo mutations that is used to compute a posterior probability for each de novo mutation being correctly identified (11, 25). These algorithms therefore must be trained with a set of quality metrics obtained from a previously validated positive and negative set of variants (26). In addition, in previous reports, large populations are needed to remove the artifacts produced by the sequencing process, along with stringent quality filters to identify bona fide de novo variants (15–17, 27).

The strategy presented in this work is based on the most reliable types of alignments: perfect matches of unique strings of the genome followed by an analysis of the resulting alignment coverage. Other algorithms rely on less reliable alignments of imperfect matches spanning repetitive sequences and establishing probability thresholds to measure the quality of the findings. The performance of COBASI was assessed by simulation experiments, and for SNV discovery in one individual, we obtained an AUPR of 0.94 and 0.96 for a sequencing depth of 35× and 100×, respectively. In most cases, the AUPR for COBASI was similar to previously reported AUPRs (10), even though previous reports only used exome data, which represents about 2% of the genome. For de novo SNV discovery, we obtained

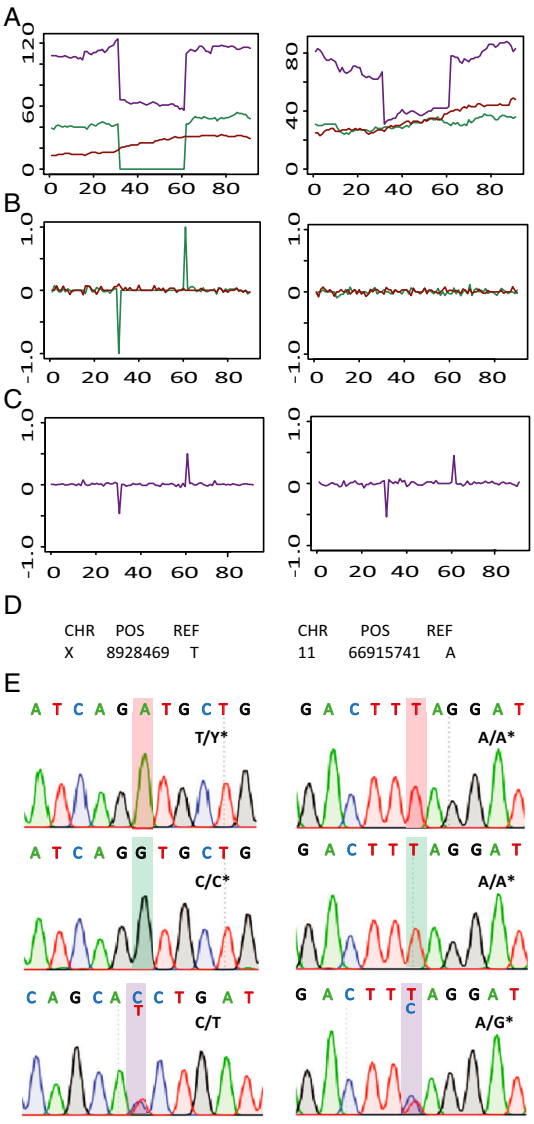


Fig. 5. Experimental example of the COBASI strategy in the family-based framework. (*Left*) A Mendelian SNV is shown. Position 1 in the plots corresponds to chrX position 8928409. (*Right*) A de novo SNV is shown. Position 1 in the plots corresponds to chr11 position 66915681. (*A*) The corresponding section of the VL is shown for each parent–offspring trio individual: the red, green, and purple lines correspond to the VL for the father, mother, and child, respectively. Since the Mendelian SNV is located in the chrX, the father has around half the coverage of the mother. (*B*) The RVL is shown for both parents. (*C*) The RVL is shown for the child. (*D*) The nucleotide present at the RG is shown. (*E*) The chromatograms obtained by Sanger sequencing for these regions are shown. The genotypes obtained for each individual by the COBASI approach are shown in bold letters. An asterisk next to the individual genotype indicates that the chromatogram is in the reverse orientation. The SNV position is shadowed according to the individual color code.

precision of 1.0 and a recall of 0.91 using COBASI, while a precision of 0.89 and a recall of 1 were obtained if the de novo SNV discovery was done by alignment-based approaches. COBASI achieves a good compromise between the increase of precision at the expense of a small decrease in recall. Furthermore, COBASI was tested on the whole callable genome, which constituted about 84% of the genome. It is also much faster than alignment-based approaches to achieve similar accuracy.

The precise identification of variant sites by COBASI relies on global alignments that include the variant site and two unique strings, one string located at each side of the variant site. Due to the small size of the reads, only small insertions or deletions would generate these high-quality alignments. Furthermore, in such cases, specialized aligners and detection algorithms would be required to pinpoint the variant positions. Incorporation of these specialized algorithms could be an extension of COBASI's scope.

The computing resources and time required by COBASI enable its routine utilization. Generating a whole-genome SNV list from 35× raw sequencing data requires around 40 h on a computer server with 12 cores and 64 Gb of RAM. Moreover, the whole-genome variation landscape can be generated in only 8 h. Furthermore, if only some regions of interest are chosen to be investigated, the time required to generate a list of SNVs can be greatly reduced (*SI Appendix*, Table S6).

In this work, we analyzed the whole-genome sequencing of a parent–offspring trio sequenced to a genome coverage of 35× for the parents and 100× for the child. We did not assume any a priori de novo mutation rate. We applied coverage filters, but not quality filters on the reads. Regardless, we found no false positives in either our de novo SNV predictions or in the randomly selected Mendelian SNVs. Moreover, we found 58 de novo SNVs, and this number is consistent with the number of de novo SNVs expected from the previously reported germline mutation rate, $1.0–1.8 \times 10^{-8}$ per nucleotide per generation, which translates into 44–82 de novo SNVs per individual (9, 28). This was accomplished because our approach combines a highly sensitive discovery in the child genome with an exhaustive validation in both parents. The number of discovered variants could be an underestimate, given that we can only interrogate 84% of the genome. However, with a world-wide sequencing capacity tending toward hundreds of thousands of genomes each year (29), our main interest is in maximizing the precision in the called variants to diminish as much as possible the extent of experimental validation that is required.

Recently, some publications have addressed the issue of calling SNVs by implementing mapping-free strategies. Known SNVs have been identified from sequencing reads if unique kmers containing the alternative allele are present in the reads (30). A Burrows–Wheeler transform of the reads was used to localize SNVs based on differences in kmer frequency (31). Changes in kmer frequency have been used to reconstruct haplotypes from genomic regions harboring long variants, this strategy focused on specific regions of the genome (32). A recently published work from our group used kmer frequency changes to identify variants along natural genomes and synthetic chromosomes of haploid yeast strains (33). However, no previous work has focused on finding de novo SNVs in human whole genomes.

COBASI could be used to identify SNVs from different organisms, since the successful application of COBASI is only limited by the ploidy of the organism and the fraction of its genome that can be represented by unique strings. Within a single genome this approach can also be used to analyze CSs from particular regions of interest, such as a cancer gene panel or other sets of genes, thus speeding up the analysis time. We propose that the general principle underlying COBASI can be used in a broad range of applications, including personalized

genomics, family studies, population genetics, ancient DNA studies, and metagenomics. It could also be used for general correlations between genotype and phenotype, such as different disorders characterized by the presence of de novo mutations, such as intellectual disability, autism, and schizophrenia (7–9).

Materials and Methods

COBASI Pipeline. The program Jellyfish (34) was used to count the number of occurrences of each kmer (k = 30) along the reads. To eliminate possible sequencing errors, all unique kmers were discarded. From the Jellyfish database, the count for every kmer along the RG was retrieved using the covplot script from the AMOS repository (35), and the read-based kmer counts associated with CSs were kept to generate the VL. The VL contained the start position for every CS along the genome and its number of occurrences in the reads (coverage). To identify CSs with abnormal coverage for each simulation or sequencing experiment, a coverage threshold was calculated. It corresponded to the median of the coverage [+/−]10 interquartile range (IQR), and ∼99.99% of the CSs had coverage values inside this rank. The VL was transformed into the RVL using the RCI. All CSs with an abnormal coverage were not taken into account.

In the child, the VSRs were identified from the RVL. Specifically, COBASI searches for regions with an abrupt drop in coverage followed by an abrupt rise in coverage. These partial VSRs were extended at most k nucleotides upstream and k nucleotides downstream. To characterize drastic changes in coverage, we required a minimum coverage as well as a minimum absolute value for the RCI for each of the signature CSs. Additionally, to extend the partial VSRs, a maximum ratio between the coverage of both signature CSs was established. The reference sequence for each signature CS was obtained, and all of the reads containing a signature CS were retrieved. A file containing the read identifier, the start reference position for the signature CS, and the position in the read for the match between the CS and the read and its orientation was created. Some inconsistent reads were filtered out (*SI Appendix, SI Materials and Methods*). For the case of the parents, the signature CSs obtained in the child were used to retrieve the reads of interest.

From reads containing both signature CSs, whole-VSR alignments were computed using a modified C++ align function from the AMOS repository. For each read, the region from the start of the PrevCS to the end of the PostCS was aligned to the corresponding RG region. These alignments were considered high-quality alignments, and only variants found in at least a certain number of these were further analyzed. For reads containing onl the PrevCS, the alignment between the RG and the read was done from th start of the PrevCS to 5 nt downstream of the last variant nucleotid obtained from the high-quality alignments. In the case of the parents, i there was no variation in the whole-VSR alignments, the default extensio was 5 bp. For all complete alignments, SNVs were identified.

The genotype of every SNV was assigned based on the algorithm describe by Li (11), modified as described in *SI Appendix, SI Materials and Methods*. T identify the possible de novo SNVs, the genotypes for each of the individual of the family trio were compared, and the potential de novo SNVs wer identified. We defined criteria to establish a possible variant, such as a bon fide de novo variant (*SI Appendix, SI Materials and Methods*). Low-coverag sequencing experiments are prone to a higher number of both FN and FP call Therefore, COBASI includes additional quality requirements to avoid incorrec de novo SNV calls. Regions prone to incorrect genotype assignment wer identified and excluded: (*i*) regions with low CS density, (*ii*) regions with mor than one CS with a coverage higher than expected, (*iii*) regions with lo coverage for any of the signature CSs in any individual, (*iv*) regions wit additional significant changes in coverage inside the region correspondin to the child VSR: in the case of the child if there is any additional drop or ris it should correspond to a region with almost no coverage; in the case of th parents there should not exist any drop or rise corresponding to the chil SNV position, and (*v*) regions with unequal coverage in both sides of the VS for the child.

Additional Methods. Additional methods are found in *SI Appendix, SI Materia and Methods*: *Definition of CS Regions from the RG*, *Definition of Accessibl Genome*, *Simulation Experiments*, *Variant Calling Using Alignment-Base Pipelines*, *TRIO Sequencing* and *COBASI Application*, and *Experimental Val dation of de Novo SNVs*.

ACKNOWLEDGMENTS. We thank James Gurtowski and Giuseppe Narzi (Cold Spring Harbor Laboratory) and Jair García Sotelo [Laboratorio Inte nacional de Investigación Sobre el Genoma Humano, Universidad Nacion Autónoma de México (UNAM)] for their technical support. L.G.-R. is a doc toral student from Programa de Doctorado en Ciencias Biomédicas, UNA and received Fellowship 275908 from Consejo Nacional de Ciencia y Tecno logía. This work was supported, in part, by US National Science Foundatio Award DBI-1350041 and US National Institutes of Health Award R01-HG00667 (to M.C.S.).

1. International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437:1299–1320.
2. Frazer KA, et al.; International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.
3. Abecasis GR, et al.; 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.
4. Auton A, et al.; 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526:68–74.
5. Hudson TJ, et al.; International Cancer Genome Consortium (2010) International network of cancer genome projects. *Nature* 464:993–998, and erratum (2010) 465:966.
6. The Cancer Genome Atlas. National Cancer Institute and National Human Genome Research Institute. Available at https://cancergenome.nih.gov/. Accessed April 22, 2018.
7. Acuna-Hidalgo R, Veltman JA, Hoischen A (2016) New insights into the generation and role of *de novo* mutations in health and disease. *Genome Biol* 17:241.
8. Lupski JR (2010) New mutations and intellectual function. *Nat Genet* 42:1036–1038.
9. Veltman JA, Brunner HG (2012) *De novo* mutations in human genetic disease. *Nat Rev Genet* 13:565–575.
10. Hwang S, Kim E, Lee I, Marcotte EM (2015) Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci Rep* 5:17875.
11. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27:2987–2993.
12. McKenna A, et al. (2010) The genome analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303.
13. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498.
14. Brockman W, et al. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res* 18:763–770.
15. Kong A, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488:471–475.
16. Girard SL, et al. (2011) Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet* 43:860–863.
17. Besenbacher S, et al. (2015) Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun* 6:5969.
18. Jin SC, et al. (2017) Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat Genet* 49:1593–1601.
19. Francioli LC, et al.; Genome of the Netherlands consortium (2017) A framework fo the detection of de novo mutations in family-based sequencing data. *Eur J Hu Genet* 25:227–233.
20. Peng G, et al. (2013) Rare variant detection using family-based sequencing analysi *Proc Natl Acad Sci USA* 110:3985–3990.
21. Reyes J, et al. (2011) Context-dependent individualization of nucleotides and virtu genomic hybridization allow the precise location of human SNPs. *Proc Natl Acad S USA* 108:15294–15299.
22. Levy S, et al. (2007) The diploid genome sequence of an individual human. *PLoS Biol* e254.
23. Wheeler DA, et al. (2008) The complete genome of an individual by massively paralle DNA sequencing. *Nature* 452:872–876.
24. Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing Computational challenges and solutions. *Nat Rev Genet* 13:36–46.
25. Li B, et al. (2012) A likelihood-based framework for variant calling and *de nov* mutation detection in families. *PLoS Genet* 8:e1002944.
26. Michaelson JJ, et al. (2012) Whole genome sequencing in autism identifies hotspo for *de novo* germline mutation. *Cell* 151:1431–1442.
27. Sanders SJ, et al. (2012) *De novo* mutations revealed by whole-exome sequencing ar strongly associated with autism. *Nature* 485:237–241.
28. Campbell CD, Eichler EE (2013) Properties and rates of germline mutations in human *Trends Genet* 29:575–584.
29. Stephens ZD, et al. (2015) Big Data: Astronomical or genomical? *PLoS Biol* 1 e1002195.
30. Pajuste FD, et al. (2017) FastGT: An alignment-free method for calling common SN directly from raw sequencing reads. *Sci Rep* 7:2537.
31. Kimura K, Koike A (2015) Ultrafast SNP analysis using the Burrows-Wheeler transfor of short-read data. *Bioinformatics* 31:1577–1583.
32. Audano PA, Ravishankar S, Vannberg FO (2017) Mapping-free variant callin using haplotype reconstruction from k-mer frequencies. *Bioinformatics*, 10.109 bioinformatics/btx753.
33. Palacios-Flores K, et al. (2018) A perfect match genomic landscape provides a unifie framework for the precise detection of variation in natural and synthetic haploi genomes. *Genetics* 208:1631–1641.
34. Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel countin of occurrences of k-mers. *Bioinformatics* 27:764–770.
35. Schatz MC, et al. (2013) Hawkeye and AMOS: Visualizing and assessing the quality c genome assemblies. *Brief Bioinform* 14:213–224.

Capítulo 4

RESULTADOS ADICIONALES

Los resultados descritos en las siguientes secciones incluyen extensiones de los resultados principales y algunos resultados adicionales.

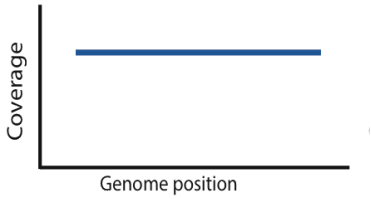
4.1. Descripción del principio fundador del método COBASI

Supongamos que tenemos un genoma de referencia: ATGGCACTAA. Asumiendo un tamaño de palabra k=4 es claro que todas las palabras que conforman nuestro genoma son únicas. Supongamos que secuenciamos tres individuo de esta especie, la cual es diploide. Sabemos que uno de ellos tiene dos cromosomas idénticos a la referencia (ATGG**C**ACTA/ATGG**C**ACTA), otro tiene dos cromosomas idénticos entre ellos pero que cambian respecto al genoma de referencia en una posición (ATGG**T**ACTA/ATGG**T**ACTA), y el tercero tiene un cromosoma de cada tipo (ATGG**C**ACTA/ATGG**T**ACTA). Definamos que la cobertura de una palabra particular es igual al número de ocurrencias de esa palabra en las lecturas de un proyecto de secuenciación. Ahora, supongamos que para cada palabra del genoma de referencia se obtiene la cobertura en cada uno de los proyectos de secuenciación individuales. Los resultados esperados para cada individuo se ilustran en la Figura 4.1. En cada tabla existe un renglón para cada palabra de referencia, la primer columna es la posición de inicio de esa palabra, la segunda columna es la secuencia y la tercer columna es la cobertura. Claramente, podemos observar como los organismos con al menos una copia del genoma mutante, tienen una reducción en la cobertura para todas las palabras que contienen la posición mutante. Estos cambios en cobertura generan firmas. Estas firmas son utilizadas por muestro método como la guía para identificar posiciones con variación en el genoma. Es importante mencionar que los resultados anteriores cambiarían dramáticamente si se incluyeran en el análisis palabras no únicas en el genoma de referencia. Este procedimiento de identificación de firmas es extrapolable al genoma completo y es totalmente dependiente del tamaño de palabra elegido.

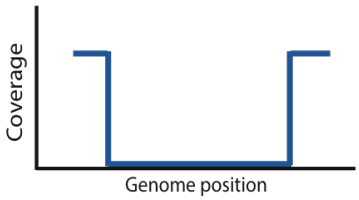
Las palabras únicas, de un tamaño definido, son muy importantes para nuestro método y se han definido como COIN-Strings (CSs) [56]. La ventaja de este tipo de palabras, es que "marcan" un contexto genómico. Por lo tanto, todas las lecturas (o la mayoría) que contengan un CS específico, deben pertenecer a la misma ubicación genómica. Por lo tanto, si en un genoma haploide un nucleótido cambia, todos los CSs que contengan este nucleótido particular dejarán de existir. En el caso de un genoma diploide, la magnitud de los cambios

en cobertura dependerán de la ploidía del cambio.

POSICIÓN	CS	COBERTURA
1	ATGG	30
2	TGGC	30
3	GGCA	30
4	GCAC	30
5	CACT	30
6	ACTA	30



POSICIÓN	CS	COBERTURA
1	ATGG	30
2	TGGC	0
3	GGCA	0
4	GCAC	0
5	CACT	0
6	ACTA	30



POSICIÓN	CS	COBERTURA
1	ATGG	30
2	TGGC	15
3	GGCA	15
4	GCAC	15
5	CACT	15
6	ACTA	30

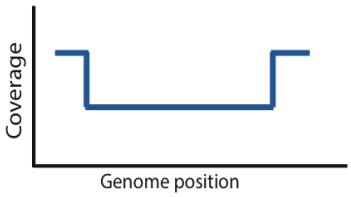


FIGURA 4.1: Cambios drásticos en la cobertura de CSs consecutivos revelan posiciones en las que existe una variante entre el genoma de referencia y el genoma blanco

4.2. Determinación del valor de corte para el índice de cobertura relativo

Las gráficas de cobertura mostradas en la Figura 4.1 son únicamente ilustrativas. Quien ha trabajado con datos de secuenciación producidos por aparatos de última generación se habrá dado cuenta de que la cobertura de cualquier experimento fluctúa a lo largo del genoma [67] [25]. En la Figura 4.2 se muestra una gráfica de cobertura de una región pequeña de un experimento de secuenciación real. Algunas de las causas conocidas de estas fluctuaciones son: diferencias en la afinidad de la amplificación por PCR entre distintas zonas del genoma,

o diferencias en la afinidad de secuenciación dependiente del contenido de GC. Además, los errores de secuenciación tampoco presentan una distribución uniforme, su distribución varía de acuerdo al contenido de GC del contexto genómico de cada lectura y a la posición de la base secuenciada dentro de la lectura [24].

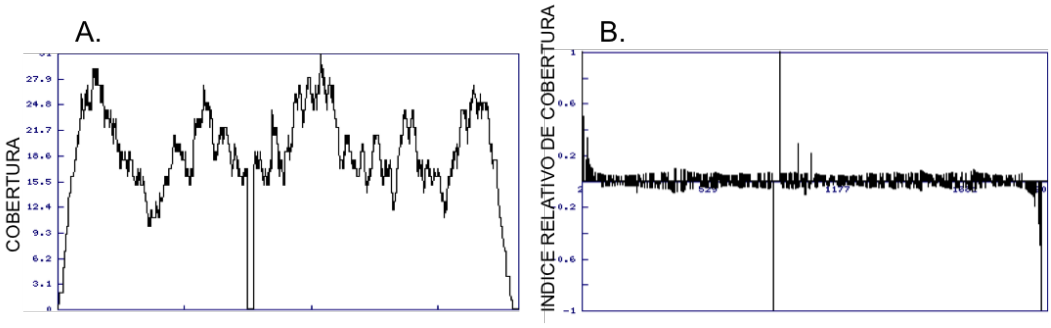


FIGURA 4.2: Una gráfica de cobertura y su transformación utilizando el Índice de Cobertura Relativo

Tomando en cuenta que el principio fundador de nuestro método se basa en cambios en la cobertura para identificar las posiciones variables de un genoma problema, diseñamos un índice y un experimento para determinar si los cambios en la cobertura producidos por la existencia de una variante genética podían ser diferenciados de las fluctuaciones "normales" en la cobertura. El índice utilizado, llamado Índice Relativo de Cobertura o RCI (por sus siglas en inglés, Relative Coverage Index), se define de la siguiente manera:

$$RCI_n = \frac{X_{n1} - X_n}{max(X_{n1}, X_n, 1)} \tag{4.1}$$

Donde RCI_n es el valor del índice para la posición n del genoma, X_n es la cobertura para el CS que inicia en la posición n y X_{n1} es la cobertura para el CS río abajo más cercano al CS_n. Debido a que X_n y X_{n1} pueden ser iguales a cero, el factor 1 en el denominador ayuda a que la ecuación no se indefina en estos casos límite.

El comportamiento de este índice para algunos valores ejemplo se muestra en la Tabla 4.1. Se puede observar que el valor de este índice es cercano a cero cuando existen ligeros cambios de cobertura, su valor absoluto es cercano a 0.5 cuando la cobertura cambia en un 50 % y su valor absoluto es cercano a 1 cuando la cobertura cambia en un 100 %. Además, el signo de este índice indica la dirección del cambio, siendo negativo cuando la cobertura baja o positivo cuando la cobertura aumenta.

Para demostrar que el valor del RCI nos permite diferenciar entre fluctuaciones al azar en la cobertura y cambios drásticos provocados por la existencia de una variante genética se

Cuadro 4.1: Comportamiento del Índice Relativo de Cobertura.

Posición inicio CS	**Cobertura**	**RCI**
1	22	-0.09
2	20	-1
3	0	0
4	0	1
5	25	-0.4
6	15	

utilizaron los datos de la secuenciación de una familia. Se identificaron dos tipos de regiones, aquéllas para las que la cobertura presenta una bajada drástica a cero en ambos padres y aquéllas para las que el RCI permanece prácticamente idéntico para uno de los padres y presenta una bajada drástica a cero en el otro [Figura 4.3 (panel superior)], es decir, regiones homocigotas no referencia en ambos padres [Figura 4.3(panel izquierdo)] o regiones homocigotas referencia en uno de los padres y homocigotas no referencia en el otro [Figura 4.3(panel derecho)]. En la descendencia estas regiones deben ser homocigotas no referencia o heterocigotas, respectivamente.

Para ambos tipos de regiones se calculó el RCI a partir de los datos de secuenciación de la descendencia del trío [Figura 4.3 (panel medio)]. Existen tres tipos de valores de este índice: 1) El RCI del CS anterior a la bajada en cobertura (prevCS), el cual debe de ser cercano a -1 para las regiones homocigotas no referencia o cercano a -0.5 para las regiones heterocigotas; 2) el RCI del último CS antes de la subida (postCS), el cual debe ser cercano a 1 para las regiones homocigotas no referencia o cercano a 0.5 para las regiones heterocigotas; 3) los valores de RCI entre los dos CSs descritos anteriormente (interCSs), los cuales se espera que sean cercanos a cero, los valores para el RCI de esta zona son un reflejo del ruido intrínseco en la cobertura que existe en cualquier experimento de secuenciación .

Los resultados se muestran en la Figura 4.3 (panel inferior). Podemos observar que en la inmensa mayoría de los casos (más del 99 % de los valores analizados) el RCI para el prevCS (rojo) y el postCS (azul) se comporta de la manera esperada. Además, podemos observar que el 99.9 % de los valores del RCI para los CS intermedios (región verde) son muy cercanos a cero. Con esto se concluye que el RCI es capaz de distinguir entre fluctuaciones en la cobertura producidas por causas no genéticas y aquéllos cambios drásticos que se deben a la existencia de variantes genéticas.

4.3. Firmas complejas producidas por eventos de variación cercanos

En el Proyecto de los 1000 Genomas se cuantificó que la cantidad de variantes esperadas en cualquier genoma individual con respecto al genoma de referencia es de aproximadamente

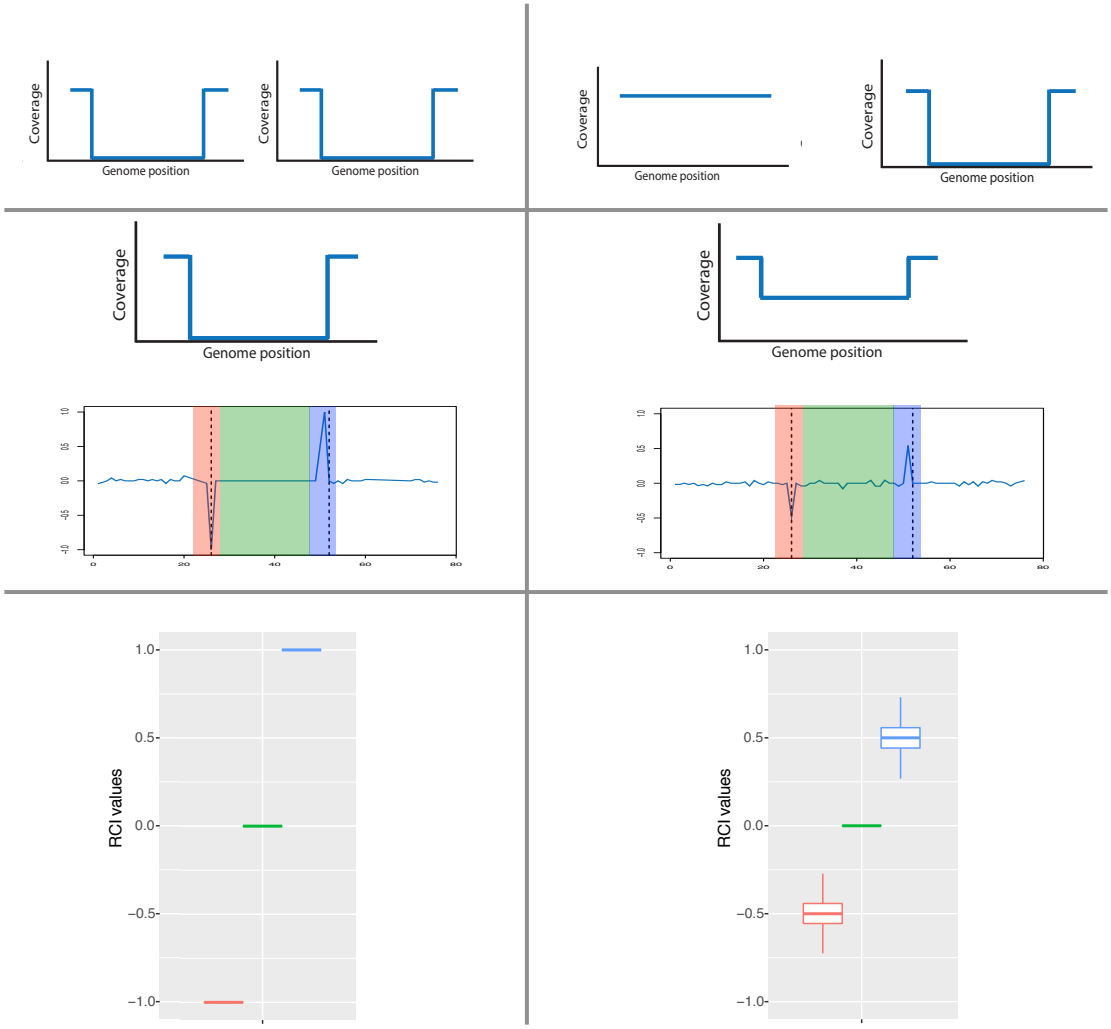


FIGURA 4.3: Comportamiento del índice RCI.

3 millones de variantes, contando únicamente las variantes de un sólo nucleótido (SNVs) [4]. Además, en años posteriores, se descubrió que la frecuencia de mutación no es constante a lo largo del genoma y que existen algunos sitios altamente variables, en los cuales las variantes genéticas tienden a acumularse [68]. Es claro que estos eventos podrían afectar la cobertura de ciertas regiones genómicas y el comportamiento esperado del RCI, por lo tanto, se hicieron simulaciones para entender el efecto de la existencia de variantes genéticas aledañas. Se simuló la secuenciación de un genoma problema diploide en el cual se incluyeron mutaciones bajo distintos escenarios.

Los resultados se muestran en la Figura 4.4. La primera conclusión de estos experimentos es que variantes aledañas tendrán un efecto en las firmas de variación, sólo sí la distancia entre dichas variantes es menor a k nucleótidos; donde k corresponde al tamaño de palabra utilizado para localizar a las subcadenas del genoma que serán consideradas como CSs. En

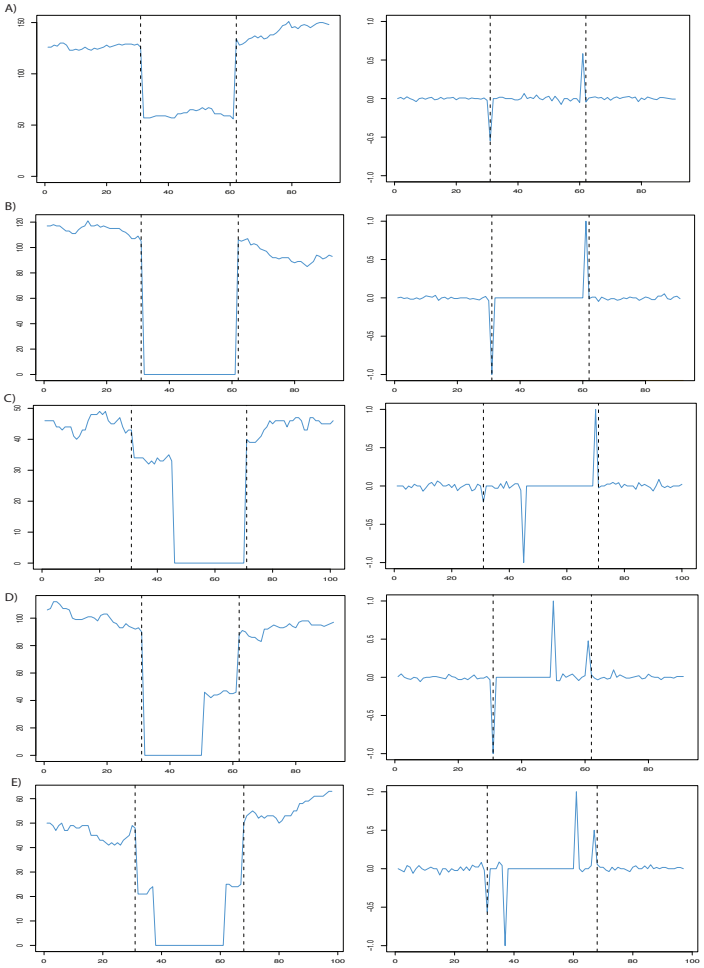


FIGURA 4.4: Gráficas de cobertura y RCI obtenidas a partir de simulaciones con múltiples variantes genéticas aledañas.

el panel A se ilustra una variante clásica heterocigota; cuando existen múltiples variantes heterocigotas próximas, para las cuales el alelo alternativo se encuentra en el mismo cromosoma, la firma clásica se alarga hasta la última de las variantes, es decir, el valle de la firma (la zona entre la bajada y la subida) tiene una mayor longitud y la firma no presenta un patrón de escalera en ninguno de sus bordes. En el panel B se muestra una variante clásica homocigota; cuando existen múltiples variantes homocigotas próximas, el valle de la firma tiene una mayor longitud y la firma no presenta un patrón de escalera en ninguno de sus bordes, un comportamiento similar al que se ha descrito para el caso anterior. Los tres paneles inferiores muestran firmas que presentan características adicionales. En el panel C se ilustra el caso en el que una variante heterocigota va seguida de una variante homocigota río abajo; en este caso, la firma presenta un patrón de escalera al inicio. En el panel D, se observa el patrón producido cuando después de una variante homocigota existe una variante heterocigota; en este caso, se observa un patrón de escalera al final de la firma. Por último, en

Cuadro 4.2: Número de firmas complejas obtenidas a partir de la secuenciación de un individuo.

Tipo de Firma	Número de firmas	Porcentaje
Sin escalera	3,783,316	0.871
Escalera al inicio	237,532	0.054
Escalera al final	202,973	0.046
Doble escalera	119,212	0.027

el panel D se muestra el caso en el que existen dos variantes heterocigotas para las cuales el alelo alternativo se ubica en distintos cromosomas; en este caso, la firma presenta un patrón de escalera doble, es decir, existe una escalera tanto al inicio como al final de la firma.

La importancia de conocer que este tipo de regiones existe subyace en que cualquier CS inmerso en esta región sólo se encuentra en una fracción de las lecturas que pertenecen a esta zona. El método desarrollado debe ser capaz de identificar correctamente los bordes de una firma de variación, es decir, debe identificar regiones con múltiples bajadas y subidas como una única firma. Para cuantificar qué tan comunes son las firmas de variación complejas descritas en el párrafo anterior se utilizaron los datos de secuenciación de la descendencia del trío analizado durante este proyecto. La profundidad promedio a la que se secuenció dicho individuo fue de 100x. A partir de estos datos, se obtuvieron todas las firmas de variación de la manera descrita en la sección de Métodos del Apéndice A y se contó el número de firmas de cada tipo. Los resultados obtenidos se muestran en la Tabla 4.2. Se puede observar que las firmas con patrones de escalera corresponden a menos del 13 % de todas las firmas identificadas.

4.4. Determinación del sexo para cada individuo del trío como una prueba de concepto

COBASI está basado en la identificación de secuencias únicas del genoma de referencia en un genoma blanco. Para que el método sea exitoso estas secuencias únicas deben de existir en el genoma blanco y deben de ser únicas en el genoma blanco. Si estos requisitos se cumplen, entonces cualquier CS que presente alguna mutación en el genoma blanco presentará los patrones de cobertura descritos anteriormente. Como una prueba de concepto se utilizó la región genómica del gene SRY, ubicada en la región no pseudoautosomal del cromosoma Y, lo cual implica que este gene existe únicamente en hombres. Por lo tanto, esta región debe de existir en el padre, no debe de existir en la madre, y su comportamiento en la descendencia depende del sexo de la misma.

Por lo tanto, como un experimento de prueba de concepto, se determinó el sexo de cada individuo del trío utilizando los panoramas genómicos producidos en la primera mitad del método COBASI. Se obtuvieron todos los CSs para esta región genómica (subcadenas únicas

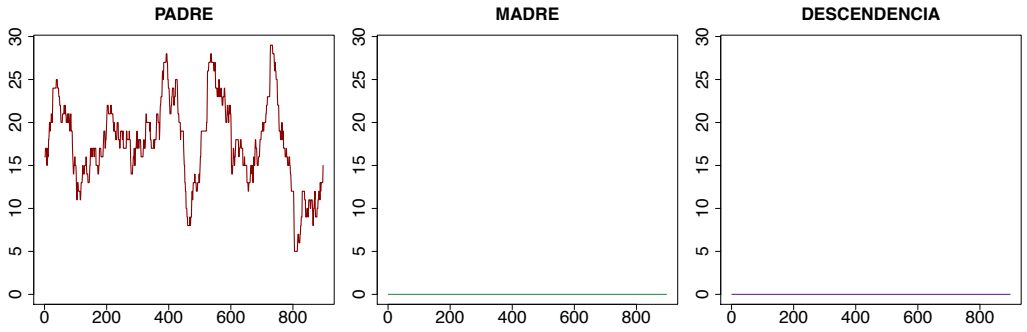


FIGURA 4.5: Gráficas de cobertura para el gene SRY. En el eje X se observa la posición de inicio de cada CS de la región (posición relativa al inicio del gen SRY). En el eje Y se observa la cobertura para cada uno de los CSs de esta región.

a nivel de genoma completo). A partir de los datos de secuenciación de cada individuo se obtuvo la cobertura para cada CS en esta región. La profundidad promedio de secuenciación de los genomas parentales fue de 35x y de 100x para la descendencia. Los resultados se observan en la Figura 4.5. Para el padre y la madre los resultados concuerdan con lo esperado, para el primero podemos observar que todos los CSs de esta región existen en al menos 5 lecturas, lo cual contrasta con la cobertura promedio de 0 que se observa en la madre. Además, con estas gráficas podemos determinar con exactitud el sexo de la descendencia: un individuo de sexo femenino. De este experimento se deriva una aplicación importante para el método COBASI: este método puede ser utilizado para estudiar regiones específicas del genoma sin la necesidad de mapear todas las lecturas producidas durante un proyecto de secuenciación.

4.5. Modificación del método para la asignación de genotipo

La asignación de genotipo es el proceso durante el cual se determina el genotipo de cierta posición del genoma con base en el número de lecturas en las cuales existe cada uno de los distintos alelos para esa posición. Esta tarea podría considerarse trivial si el proceso de secuenciación fuera absolutamente perfecto, sin embargo, no lo es; existen varias etapas durante este proceso en las cuales se puede imponer un sesgo sobre la representación de alguno de los alelos existentes, algunos ejemplos incluyen la amplificación diferencial de algún alelo durante la preparación de la librería por PCR o la adición de múltiples errores durante la corrida de secuenciación. Por lo tanto, la asignación de genotipo hace uso de técnicas estadísticas para tomar la mejor decisión acerca del genotipo más probable.

Uno de los métodos más ampliamente utilizados para asignar genotipo es el propuesto por Li [41]. Este método consiste en calcular tres probabilidades para cada sitio en el genoma: que el genotipo del sitio en cuestión sea homocigoto referencia, homocigoto no referencia o heterocigoto. Es evidente a partir de este planteamiento que se asume que todas las variantes

son bialélicas; en el caso de la población humana esta suposición no es tan grave debido a que se sabe que la fracción de SNPs trialélicos es de 0.2 % [69]. Además, Li propone una métrica, el *LogRatio*, para determinar el grado de confianza que se tiene al momento de asignar un genotipo. El *LogRatio* está definido como $\log(P(\mathrm{G1})/P(\mathrm{G2}))$, donde $P(\mathrm{G1})$ es igual a la probabilidad del genotipo más probable y $P(\mathrm{G2})$ es igual a la probabilidad del segundo genotipo más probable. De esta forma, el *LogRatio* nos dirá por cuantos órdenes de magnitud es más probable el genotipo que estamos eligiendo respecto al siguiente genotipo más probable, es decir, $LogRatio = 1$ cuando $G1$ es 10 veces más probable que $G2$; $LogRatio = 2$ cuando $G1$ es 100 veces más probable que $G2$, y así sucesivamente.

Sin embargo existe un problema con este método: alelos producidos por errores de secuenciación caen en la categoría de alelos no referencia . Supongamos un sitio en el genoma para el cual el alelo de referencia es A, ahora asumamos que nuestros resultados de secuenciación nos indican que para este sitio hay 6 lecturas con A, 4 con C y 3 con T. En el método de Li esta información se debe de clasificar en referencia (6 lecturas en nuestro caso) y no referencia (7 lecturas). Por lo tanto, de acuerdo al método de Li, este sitio sería heterocigoto. Sin embargo, la asignación de genotipo para este sitio es ambigua porque, incluso cuando se clasifique como heterocigoto, no es posible definir cuál es el alelo no referencia adecuado.

Una ampliación del método anterior se basa en obtener la probabilidad de los 10 posibles genotipos para cada sitio en el genoma (AA, AT, AC, AG, TT, TC, TG, CC, CG, GG). En general, el rendimiento de este método no es mejor al método de Li debido a que, para algunos genotipos, se pierde información al reducir la cantidad de lecturas utilizadas para asignar el genotipo. Existen algunos métodos más sofisticados, como el implementado por la última versión de GATK [70]. El algoritmo de GATK ensambla localmente los haplotipos que existen en una región específica a partir de los datos de secuenciación, este enfoque ha demostrado tener un mejor rendimiento; sin embargo, para hacer el ensamble se requiere acceder a una gran cantidad de muestras secuenciadas para utilizar la información contenida en todas las muestras durante el proceso.

En este trabajo se desarrolló un nuevo algoritmo para la asignación de genotipo. Nuestro método representa una ampliación del método de Li, la cual se enfoca en descartar sitios para los que existen más de dos alelos conviviendo en las lecturas en tal frecuencia que resultan en un sitio genómico para el que no hay suficiente información para asignar un genotipo sin ambigüedad, como en el caso del ejemplo descrito anteriormente. Los detalles de este método pueden ser consultados en el Apéndice B, el cual corresponde al material suplementario del artículo "*Precise detection of de novo Single Nucleotide Variants in human genomes*".

Para probar el rendimiento de los distintos métodos se utilizaron simulaciones. En cada simulación se generaron dos cromosomas homólogos sintéticos con variantes conocidas y se simuló su secuenciación por Illumina. Por lo tanto, en cada simulación se conocen los dos alelos verdaderos para cada sitio del cromosoma, cada sitio puede ser homocigoto referencia, heterocigoto referencia/alternativo u homocigoto alternativo. En cada simulación se aplicó el método COBASI y para realizar la asignación de genotipo se utilizó el algoritmo desarrollado por nosotros (el cual llamaremos COBASI en las siguientes secciones) y el algoritmo de Li,

CUADRO 4.3: Los resultados para dos métodos diferentes utilizados para la asignación de genotipo se muestran a continuación. Las filas LR corresponden a los resultados obtenidos con el método de Li.

	35x			100X		
Método	**FN**	**FP**	**TP**	**FN**	**FP**	**TP**
COBASI	6,620	29	110,258	4,451	5	112,398
LR=2	6,874	40	109,554	4,455	13	112,386
LR=3	7,265	23	108,773	4,457	13	112,384
LR=4	8,066	17	108,773	4,462	12	112,380
LR=5	9,958	13	106,883	4,468	12	112,374
LR=6	12,273	11	104,570	4,487	11	112,356

independientemente. Existen dos tipos de sitios que son útiles para medir el rendimiento de cada método 1) aquéllos sitios variables reales, es decir, aquéllos que fueron designados como variables desde el inicio de la simulación, para los cuales se puede identificar si el genotipo asignado por cada método fue correcto o no y 2) aquéllos sitios que son clasificados como variables por cada uno de los métodos, los cuales incluyen los sitios del inciso 1) y algunos más, los cuales son errores del método. Para cada método se calcula el número de Verdaderos Positivos (TP), Falsos Negativos (FN) y Falsos Positivos (FP). El número de sitios TP corresponde a aquéllos sitios para los que el genotipo asignado no fue correcto; el número de sitios FN son aquéllos para los que incorrectamente se asignó un genotipo homocigoto referencia; y el número de sitios FP corresponde a aquéllos para los que el método identificó una posición variable pero asignó un genotipo incorrecto o para los que incorrectamente se asignó un genotipo variable.

Para el método de Li se utilizaron varios valores de corte para el estadístico *LogRatio* (LR). Los resultados, para dos profundidades promedio de secuenciación diferentes (35x y 100x), se muestran en la Tabla 4.3. En esta tabla se puede observar que la asignación correcta de genotipo es altamente dependiente de la cobertura del experimento de secuenciación sin importar el método utilizado para asignar genotipo. Para el caso de una profundidad promedio de 35x el número de TPs obtenidos por COBASI es comparable con el número de TPs obtenidos por el método de Li utilizando un LR de 2; aunque, notablemente, el número de FPs obtenidos por COBASI es menor que el número de FPs obtenidos por el método de Li para este mismo LR. Para esta cobertura COBASI presenta el mayor número de TPs, y un número intermedio de FPs. Por otro lado, en el caso de un experimento con una profundidad promedio de 100x podemos observar que el rendimiento de ambos métodos mejora. Para esta cobertura, COBASI presenta el mayor número de TPs y, al mismo tiempo, el menor número de FPs. Tomando en cuenta estos resultados se decidió utilizar nuestro método modificado para la asignación de genotipo en el resto de este trabajo.

4.6. Definición de genoma accesible

La definición de "genoma accesible" por COBASI puede ser consultada en el Apéndice B en la sección de Métodos Suplementarios. La tabla comparativa indicando la fracción del genoma que puede ser interrogada únicamente por COBASI, únicamente por el proyecto 1000HGP, por ambos o por ninguno, se puede encontrar en el Apéndice B, en la Tabla S1. El Apéndice B corresponde al material suplementario del artículo "*Precise detection of de novo Single Nucleotide Variants in human genomes*".

En resumen, la diferencia entre las dos definiciones de accesibilidad se centra en que COBASI necesita cierta fracción de secuencias únicas para poder interrogar una región específica del genoma, mientras que en el proyecto 1000HGP filtran regiones genómicas a las que mapean demasiadas o muy pocas lecturas, o lecturas con valores de calidad particularmente bajos, de manera reproducible en muchos experimentos. 90 % del genoma puede ser interrogado por el 1000HGP, mientras que 84 % del genoma puede se interrogado si se consideran las regiones con una densidad de CSs mayor a 0.5, para este umbral de densidad 2 % del genoma puede ser interrogado sólo por COBASI, 8 % sólo por 1000HGP y 9 % por ninguno de los dos.

La Figura 4.6 muestra una comparación de las regiones clasificadas como "accesibles" dependiendo de la definición utilizada para el cromosoma 17 completo (panel izquierdo) o para una región de 50,000 nucleótidos (panel derecho). El genoma de referencia se muestra como una línea roja en la parte inferior de la figura. La línea verde representa las regiones accesibles de acuerdo a la definición del Proyecto 1000HGP. Las líneas azules representan las regiones accesibles por COBASI; la línea azul inferior corresponde a regiones accesibles cuando se pide una densidad mayor a 0 CSs (todo el genoma), la siguiente línea azul corresponde a una densidad mayor a 0.1, la siguiente línea a una densidad de 0.2 y así sucesivamente. Los valores de densidad utilizados como umbral se indican en el eje Y.

Podemos observar que la clasificación (accesibles, no accesibles) de la gran mayoría de las regiones se conserva independientemente de la definición utilizada. Sin embargo, existen unas pocas excepciones, es decir, algunas regiones que pueden ser analizadas únicamente utilizando alguna definición específica. En el caso de este trabajo se definieron como "regiones accesibles" aquéllas zonas del genoma con una densidad de CSs mayor a 0.5. Todos los análisis subsecuentes de este trabajo, se centran exclusivamente en estas regiones.

4.7. Rendimiento del método COBASI

Para medir el rendimiento de COBASI se realizaron experimentos de simulación. Brevemente, se introdujeron SNVs al azar en el cromosoma 21 humano con una tasa del 0.001 generando un genoma diploide problema, se simuló la secuenciación con Illumina de este genoma mutado y se utilizó COBASI para realizar el proceso de identificación de variantes. Las simulaciones se encuentran descritas con mayor detalle en el Apéndice B de este trabajo.

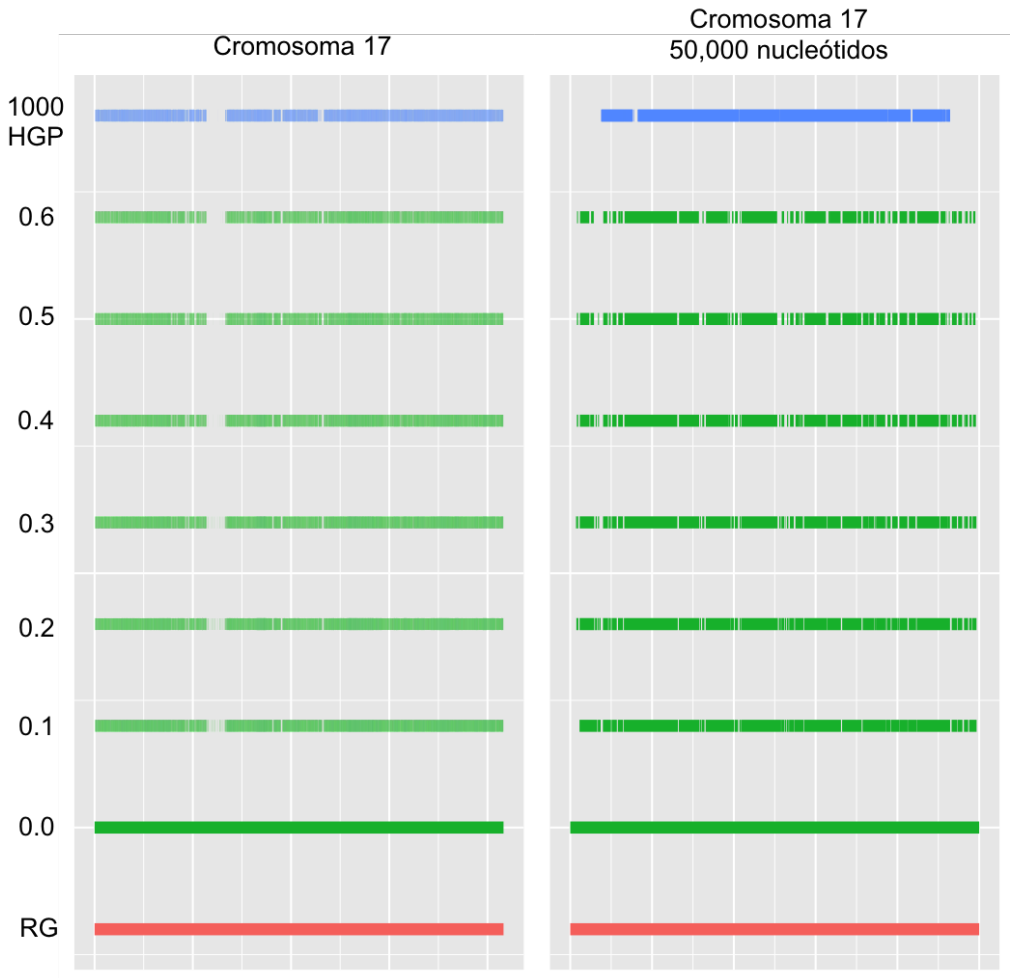


FIGURA 4.6: Las regiones definidas como "accesibles" varían dependiendo de la definición utilizada

En cada simulación se realizó el proceso de identificación de variantes probando distintos valores para los parámetros clave de COBASI: la cobertura de secuenciación, elegida al momento de diseñar el experimento; el tamaño de la subcadena o kmer (k), el cual definirá el tamaño de las subcadenas únicas, CSs, utilizadas en este estudio; el umbral para el Índice de Cobertura Relativo (RCI), con el cual se elegirán las bajadas y subidas significativas que bordearán a las firmas de variación; y el mínimo número de alineamientos que deben contener a la variante, al PrevCS y al PostCS (Total), este tipo de alineamientos dan confianza en la veracidad de la variante identificada.

Para cada una de las simulaciones se obtuvieron el número de sitios TPs, FPs y FNs, definidos como se ha explicado anteriormente. Estos números se utilizaron para obtener métricas de rendimiento: la Tasa de Verdaderos Positivos o sensibilidad (TPR o recall), el Valor Predictivo Positivo o precisión (PPV o precision), y la Tasa de Descubrimientos Falsos (FDR).

Los resultados para distintas coberturas de secuenciaciónn (35x, 50x y 100x) se pueden observar en las Tablas: Tabla 4.4, Tabla 4.5 y Tabla 4.6, respectivamente.

El principal objetivo de COBASI es disminuir lo más posible el número de Falsos Positivos. La precisión, definida como TP/(TP+FP), mide cuántos de los SNVs identificados se identificaron correctamente, por lo tanto, la precisión aumenta cuando el número de FPs disminuye. En las tablas de resultados podemos observar que, con cualquier combinación de parámetros, la precisión es sumamente alta. Los parámetros críticos en la disminución de FPs son el aumento de la cobertura de 35X a 50X, seguido por el incremento en el número mínimo de alineamientos con ambos CSs.

Por otro lado, la sensibilidad, definida como TP/(TP+FN), mide cuántos del total de SNVs que existían en el genoma problema fueron identificados correctamente. En este caso, una vez más, la cobertura de secuenciación es decisiva. Se puede observar que a una baja cobertura de secuenciación cualquier combinación de parámetros se comporta de manera similar. Sin embargo, cuando la cobertura de secuenciación aumenta, la sensibilidad varía desde 0.63 hasta 0.95 dependiendo de los parámetros utilizados; a partir de estos datos se puede concluir que el número de alineamientos con ambos CSs es el parámetro más importante seguido del tamaño del kmer, además de que el valor de corte para el RCI no es un parámetro crítico.

A partir de los resultados de estas simulaciones se obtuvieron los parámetros óptimos para cada una de las posibles coberturas de secuenciación. Los parámetros elegidos fueron utilizados en la identificación de SNVs *de novo*, tanto en las simulaciones como en el análisis de los datos del trío secuenciado en este trabajo.

4.8. Comparación del rendimiento de COBASI contra un método basado en alineamientos

Cuando se ha secuenciado el genoma de un individuo problema y se quieren identificar las variantes del mismo, generalmente se utiliza un método basado en alineamientos. La serie de pasos utilizados se ilustran en la Figura 4.7. Primero se mapean las lecturas del individuo problema hacia el genoma de referencia, eliminando aquéllas que mapean con una baja calidad; opcionalmente se eliminan los duplicados de PCR; después se realinean las zonas problemáticas localizadas generalmente alrededor de indeles y se hacen algunas correcciones de las métricas de calidad por base. Para finalizar, se asigna el genotipo tomando en cuenta la calidad conjunta del mapeo de la lectura y de cada nucleótido particular. Los algoritmos utilizados más frecuentemente son: BWA para la primer parte del proceso y GATK para la segunda. Para identificar variantes *de novo* al final del proceso se agrega un paso de comparación de genotipos entre los padres y el hijo(a), identificando aquéllas variantes del hijo(a) con una buena calidad que no se encuentran en los padres.

Por medio de simulaciones se comparó el rendimiento de COBASI con el rendimiento de la combinación BWA-GATK. Se hicieron dos clases de experimentos, la identificación de

Cuadro 4.4: Número de FPs, TPs y FNs junto con los distintos estadísticos de resumen utilizados. Cobertura de secuenciación: 35x

Parámetros	**TP**	**FP**	**FN**	**TPR**	**PPV**	**FDR**
			35X - k25			
RCI(0.2)-Total(1)	95,029	65	21,811	0.81	0.99	0.0007
RCI(0.2)-Total(2)	108,935	19	7,904	0.93	0.99	0.0002
RCI(0.2)-Total(3)	108,796	15	8,044	0.93	0.99	0.0001
RCI(0.2)-Total(4)	107,831	12	9,011	0.92	0.99	0.0001
RCI(0.2)-Total(6)	103,672	12	13,170	0.88	0.99	0.0001
RCI(0.3)-Total(1)	94,217	47	22,623	0.80	0.99	0.0005
RCI(0.3)-Total(2)	108,128	16	8,712	0.92	0.99	0.0001
RCI(0.3)-Total(3)	108,057	15	8,783	0.92	0.99	0.0001
RCI(0.3)-Total(4)	107,198	13	9,643	0.91	0.99	0.0001
RCI(0.3)-Total(6)	103,367	13	13,474	0.88	0.99	0.0001
			35X - k30			
RCI(0.2)-Total(1)	100,466	44	16,362	0.86	0.99	0.0004
RCI(0.2)-Total(2)	110,024	19	6,802	0.94	0.99	0.0002
RCI(0.2)-Total(3)	108,448	26	8,380	0.92	0.99	0.0002
RCI(0.2)-Total(4)	105,168	21	11,655	0.90	0.99	0.0002
RCI(0.2)-Total(6)	92,367	17	24,470	0.79	0.99	0.0002
RCI(0.3)-Total(1)	99,391	40	17,433	0.85	0.99	0.0004
RCI(0.3)-Total(2)	109,022	30	7,802	0.93	0.99	0.0003
RCI(0.3)-Total(3)	107,636	28	9,190	0.92	0.99	0.0003
RCI(0.3)-Total(4)	104,631	22	12,201	0.89	0.99	0.0002
RCI(0.3)-Total(6)	92,355	18	24,481	0.79	0.99	0.0002

CUADRO 4.5: Número de FPs, TPs y FNs junto con los distintos estadísticos de resumen utilizados. Cobertura de secuenciación: 50x

Parámetros	**TP**	**FP**	**FN**	**TPR**	**PPV**	**FDR**
			50x - k25			
RCI(0.2)-Total(1)	89,845	21	27,005	0.76	0.99	0.0002
RCI(0.2)-Total(2)	109,070	7	7,779	0.93	0.99	0.0001
RCI(0.2)-Total(3)	109,734	5	7,115	0.93	0.99	0.0000
RCI(0.2)-Total(4)	109,380	4	7,470	0.93	0.99	0.0000
RCI(0.2)-Total(6)	108,296	4	8,554	0.02	0.99	0.0000
RCI(0.3)-Total(1)	89,676	13	27,175	0.76	0.99	0.0001
RCI(0.3)-Total(2)	108,893	6	7,957	0.93	0.99	0.0001
RCI(0.3)-Total(3)	109,570	4	7,280	0.93	0.99	0.0000
RCI(0.3)-Total(4)	109,236	3	7,615	0.93	0.99	0.0000
RCI(0.3)-Total(6)	108,221	3	8,630	0.92	0.99	0.0000
			50x - k30			
RCI(0.2)-Total(1)	97,115	13	19,731	0.83	0.99	0.0001
RCI(0.2)-Total(2)	111,209	11	5,634	0.95	0.99	0.0001
RCI(0.2)-Total(3)	111,092	10	5,752	0.95	0.99	0.0001
RCI(0.2)-Total(4)	110,076	9	6,769	0.94	0.99	0.0001
RCI(0.2)-Total(6)	106,354	9	10,491	0.91	0.99	0.0001
RCI(0.3)-Total(1)	96,830	11	20,015	0.82	0.99	0.0001
RCI(0.3)-Total(2)	110,913	11	5,930	0.94	0.99	0.0001
RCI(0.3)-Total(3)	110,831	10	6,013	0.94	0.99	0.0001
RCI(0.3)-Total(4)	109,864	8	6,982	0.94	0.99	0.0001
RCI(0.3)-Total(6)	106,275	8	10,571	0.90	0.99	0.0001

Cuadro 4.6: Número de FPs, TPs y FNs junto con los distintos estadísticos de resumen utilizados. Cobertura de secuenciación: 100x

Parámetros	**TP**	**FP**	**FN**	**TPR**	**PPV**	**FDR**
			100x - k25			
RCI(0.2)-Total(1)	73,938	11	42,913	0.63	0.99	0.0001
RCI(0.2)-Total(2)	106,984	6	9,865	0.91	0.99	0.0001
RCI(0.2)-Total(3)	110,199	5	6,650	0.94	0.99	0.0000
RCI(0.2)-Total(4)	110,279	5	6,570	0.94	0.99	0.0000
RCI(0.2)-Total(6)	110,008	5	6,841	0.94	0.99	0.0000
RCI(0.3)-Total(1)	73,938	11	42,913	0.63	0.99	0.0001
RCI(0.3)-Total(2)	106,983	6	9,866	0.91	0.99	0.0001
RCI(0.3)-Total(3)	110,200	5	6,649	0.94	0.99	0.0000
RCI(0.3)-Total(4)	110,281	5	6,568	0.94	0.99	0.0000
RCI(0.3)-Total(6)	110,017	5	6,831	0.94	0.99	0.0000
			100x - k30			
RCI(0.2)-Total(1)	84,615	11	32,236	0.72	0.99	0.0001
RCI(0.2)-Total(2)	110,601	5	6,248	0.94	0.99	0.0000
RCI(0.2)-Total(3)	112,317	5	4,532	0.96	0.99	0.0000
RCI(0.2)-Total(4)	112,165	5	4,684	0.96	0.99	0.0000
RCI(0.2)-Total(6)	111,564	4	5,286	0.95	0.99	0.0000
RCI(0.3)-Total(1)	84,617	8	32,234	0.72	0.99	0.0001
RCI(0.3)-Total(2)	110,608	3	6,243	0.94	0.99	0.0000
RCI(0.3)-Total(3)	112,328	3	4,523	0.96	0.99	0.0000
RCI(0.3)-Total(4)	112,181	3	4,670	0.96	0.99	0.0000
RCI(0.3)-Total(6)	111,583	2	5,269	0.95	0.99	0.0000

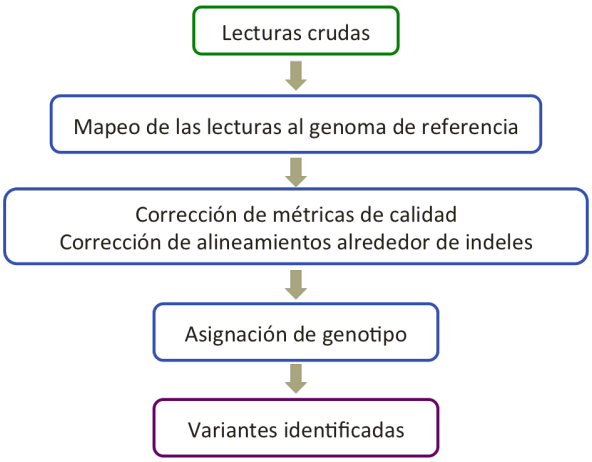


FIGURA 4.7: Los métodos basados en alineamientos normalmente siguen estos pasos durante el proceso de identificación de variantes

SNVs en un individuo y la identificación de SNVs *de novo*. Para la identificación de SNVs en un individuo se creó un genoma diploide sintético; se introdujeron SNVs al azar en el cromosoma 21 humano con una tasa del 0.001, se simuló la secuenciación con Illumina de este genoma sintético y a partir de las lecturas producidas se realizó el proceso de identificación de variantes utilizando COBASI. Para la identificación de SNVs *de novo* se utilizó el genoma previamente generado como uno de los genomas parentales, a partir del cual se crearon el genoma diploide de la madre y el genoma diploide de la hija; en el genoma de la hija se introdujeron mutaciones *de novo* con una tasa de 3e-7 (39 SNVs), se simuló la secuenciación por Illumina de los tres individuos. Cada experimento se repitió 5 veces. En cada experimento se obtuvo el número de SNVs TPs, FPs y FNs, definidas como se describe en la sección "Rendimiento del método COBASI". Además, para cada experimento se obtuvo el número de TPs, FPs y FNs para distintos umbrales de cobertura, a partir de estos datos se calculó el área bajo la curva de la curva de Precisión-Sensibilidad (AUC). Para leer con mayor detalle como se hicieron las simulaciones consulte el Apéndice B de este trabajo.

Los resultados de la identificación de SNVs para un individuo se encuentran en la Tabla 4.7. Cuando los datos son analizados por COBASI, el promedio del estadístico AUC es igual a 0.96; en cambio, cuando se utiliza la combinación de software BWA-GATK el AUC promedio es de 0.99. Los resultados de la identificación de SNVs *de novo* se encuentran en la Tabla 4.8. En este caso, cuando se identifican los SNVs utilizando COBASI se observa una precisión promedio de 1 y una sensibilidad promedio de 0.91, lo cual corresponde a un AUC promedio de 0.86; por otro lado cuando se utiliza BWA-GATK se observa una precisión de 0.89 y una sensibilidad de 1, lo cual corresponde a una AUC de 0.91. En ambos experimentos se observa que COBASI disminuye la cantidad de FPs, aumentando, al mismo tiempo, la cantidad de FNs. Es importante destacar que en el caso de la identificación de SNVs *de novo*, los FPs se logran eliminar completamente, aumentando de manera muy sutil el número de FNs. Lo anterior sugiere que COBASI puede ser una herramienta útil en los protocolos de

Cuadro 4.7: Comparación del rendimiento de COBASI contra BWA-GATK en la identificación de SNVs para un individuo

COBASI				BWA-GATK			
TP	**FP**	**FN**	**AUC**	**TP**	**FP**	**FN**	**AUC**
93,522	8	3,737	0.96	97,250	71	17	0.99
93,629	10	3,758	0.96	97,382	69	15	0.99
93,611	6	3,763	0.96	97,368	62	12	0.99
93,606	8	3,751	0.96	97,345	78	20	0.99
93,677	6	3,772	0.96	97,442	82	13	0.99

Cuadro 4.8: Comparación del rendimiento de COBASI contra BWA-GATK para la identificación de SNVs *de novo*

COBASI				BWA-GATK			
TP	**FP**	**FN**	**AUC**	**TP**	**FP**	**FN**	**AUC**
32	0	2	0.91	34	5	0	0.81
34	0	3	0.86	37	1	0	0.96
36	0	1	0.94	37	2	0	0.91
31	0	4	0.85	35	4	0	0.97
26	0	6	0.78	32	6	0	0.91

investigación en los cuales uno de los objetivos principales es disminuir el número de validaciones experimentales necesarias.

Por otro lado, durante los experimentos de simulación para identificar SNVs en un individuo se midió el tiempo que cada algoritmo utiliza para generar una lista de variantes a partir de lecturas crudas. Recordemos que durante este experimento se utilizó como base el cromosoma 12 humano y se simuló su secuenciación con una profundidad 100x. En el caso de COBASI se requieren seis horas para obtener una lista de SNVs a partir de las lecturas crudas. Estas seis horas se utilizan de la siguiente manera: la generación de un mapa de cobertura toma una hora y media; la identificación de firmas de variación a partir de este mapa se logra en 15 minutos; por último, hacer los alineamientos de las regiones de interés y asignar los genotipos es el paso más tardado tomando alrededor de 4 horas. En el caso de la combinación BWA-GATK se necesitan más de 60 horas para obtener una lista de SNVs a partir de las lecturas crudas: el mapeo de las lecturas hacia el genoma de referencia por BWA es el paso más tardado, tomando alrededor 40 horas; hacer la corrección de alineamiento alrededor de indeles y las correcciones estadísticas de los valores de calidad para cada base para cada lectura utilizando GATK, toma 27 horas; por último se necesitan 10 horas para realizar la asignación de genotipo utilizando GATK. A partir de estos datos podemos concluir que COBASI es mucho más rápido que los algoritmos basados en alineamientos comúnmente utilizados.

4.9. Análisis de calidad de las lecturas del trío

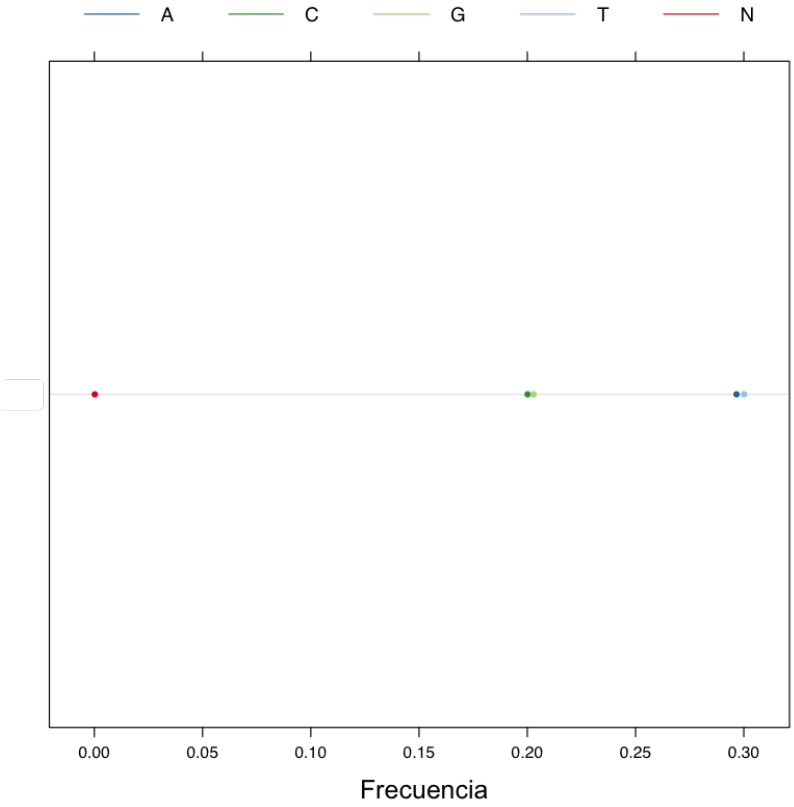


FIGURA 4.8: Se muestra en contenido de G, C, A y T en las lecturas analizadas

Durante el proceso de secuenciación se generan valores de calidad para cada nucleótido para cada lectura generada. Los valores de calidad por base reflejan la probabilidad de que esa base sea un error. En general, los métodos para llamar variantes toman en cuenta estos valores de calidad para darle una confianza al proceso de mapeo de cada lectura, y para calcular una probabilidad de que el genotipo asignado a cierta posición en el genoma sea correcto. En contraste, COBASI no utiliza los valores de calidad asignados por el secuenciador en ningún paso del proceso; más bien, se basa en alineamientos de alta calidad, en los cuales es altamente probable que la secuencia variable no sea un error, debido a que en estos alineamientos la secuencia variable está flanqueada por secuencias idénticas entre las lecturas y el genoma de referencia. Aún así, realizar un análisis de calidad de las lecturas en cualquier proyecto de secuenciación es un paso necesario para eliminar aquéllas muestras problemáticas o, incluso, contaminadas, que podrían oscurecer o equivocar los resultados y conclusiones de cualquier análisis posterior.

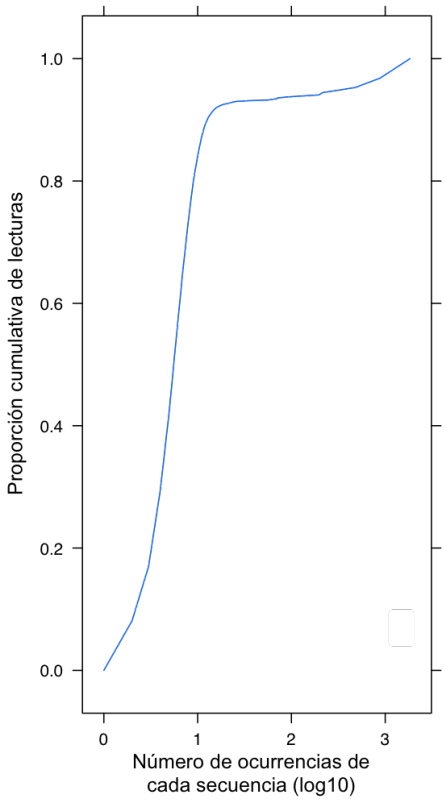


FIGURA 4.9: Se muestra el número de ocurrencias de cada secuencia (en log10) contra la proporción acumulativa de lecturas

Los análisis de calidad se deben realizar para cada librería por separado. En nuestro caso, para cada uno de los padres se preparó únicamente 1 librería y en el caso de la hija se prepararon 3 librerías diferentes. Para cada librería se tomó una muestra de 2 millones de lecturas. Se utilizó el paquete ShortRead de R para analizar la calidad de las lecturas. Sólo se muestran las gráficas para una de las librerías analizadas, estas gráficas reflejan la calidad obtenida para el resto de las librerías analizadas.

Uno de los análisis de calidad necesarios consiste en determinar sí existe contaminación en la muestra de interés. La contaminación se puede traducir en un contenido de GC que no corresponde al contenido de GC del organismo secuenciado o en una muestra de baja complejidad, es decir, una muestra formada por un número pequeño de secuencias repetidas muchas veces. La Figura 4.8 muestra el contenido de G, C, A y T en las lecturas analizadas, podemos observar que el contenido de GC de nuestra muestra es de alrededor del 40 %, el cual corresponde al contenido de GC de una muestra humana, dado que el contenido de GC de genoma humano es de alrededor del 40 % [71]; lo cual indica que no existe contaminación, o al menos, no existe contaminación de una especie con un contenido de GC diferente. Por otro lado, la Figura 4.9 muestra el número de ocurrencias de cada secuencia (en $log10$) contra la proporción acumulativa de lecturas; si se traza una línea imaginaria en el 1 del eje X, se puede

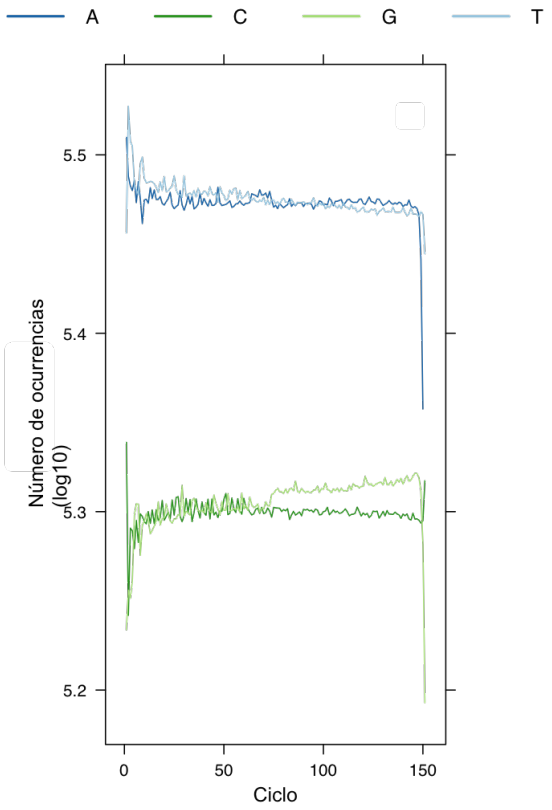


FIGURA 4.10: Se observa el número de GCAT añadido en cada ciclo del proceso de secuenciación

observar que casi 90 % de las lecturas pertenecen a secuencias que están representadas menos de 10 veces [$log10(10) = 1$], lo cual implica que nuestra muestra no es de baja complejidad.

El siguiente paso en el análisis de calidad, es cuantificar la calidad del proceso por ciclo y la calidad del proceso de una forma global. Una baja calidad por ciclo se puede traducir en patrones diferentes de adición de alguna base particular en algún ciclo específico o en valores bajos de calidad en ciclos intermedios de la corrida de secuenciación. Una baja calidad global se puede traducir en valores bajos de calidad promedio para un número grande de lecturas. En la Figura 4.10 se observa el número de G, C, A y T añadido en cada ciclo, como es de esperarse el número de CG y AT añadido a las lecturas en cada ciclo corresponde al contenido de GC del genoma problema, en ciclos intermedios no se observan picos inesperados, y al principio y al final de la corrida se observan subidas y bajadas extremas que concuerdan con las zonas esperadas de baja calidad de secuenciación. En la Figura 4.11 se ilustra la calidad promedio para cada ciclo del proceso de secuenciación, se observa que las primeras 5 bases y las últimas 30 tienen una calidad menor a las bases intermedias, aunque no dramáticamente baja; este comportamiento es el esperado en las corridas de secuenciación por Illumina. La Figura 4.12 corresponde a la distribución de calidad promedio, es claro que

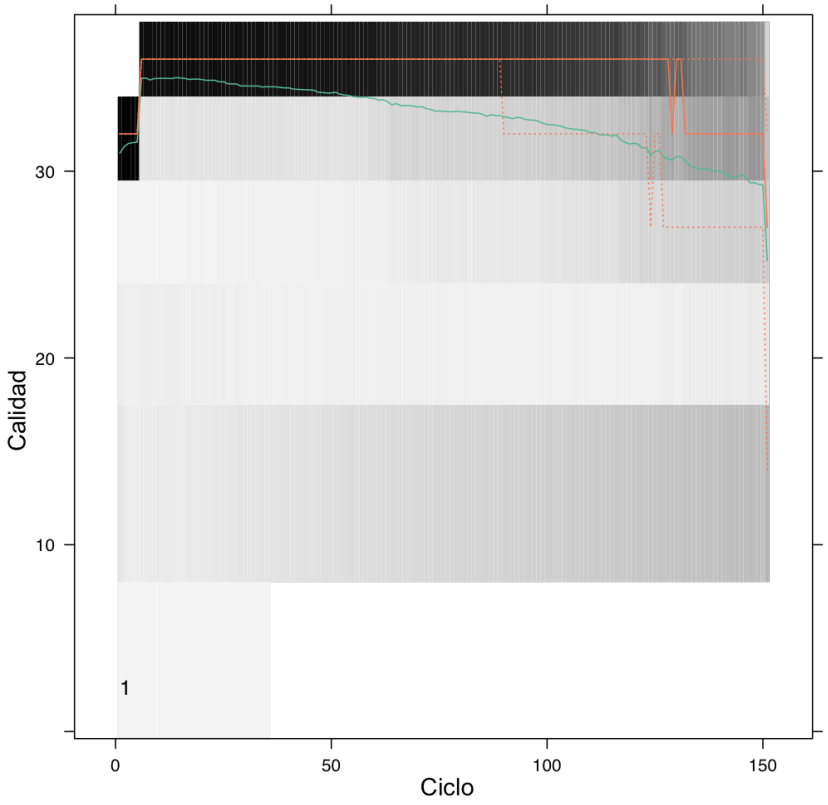


FIGURA 4.11: Se observa la calidad promedio de todas las lecturas para cada ciclo del proceso de secuenciación

la mayoría de las lecturas tienen una calidad promedio mayor a 30. Estos tres resultados reflejan una buena calidad, tanto por ciclo como global, para nuestras corridas de secuenciación.

4.10. Análisis de cobertura.

En general, la cobertura de una posición genómica particular en un experimento de secuenciación se define como el número de veces que se encuentra representada esa posición particular en las lecturas generadas por el secuenciador. Sin embargo, la definición de cobertura utilizada en este trabajo es ligeramente diferente; la cobertura de una posición específica en el genoma se obtiene contando el número de lecturas que contienen al CS que empieza en esa posición particular. Recordemos que los CSs son cadenas únicas de tamaño k (30 nucleótidos, en el caso de este trabajo). Aunque se sabe que no todas las posiciones genómicas serán secuenciadas en la misma proporción, no se espera una desviación extrema para la mayoría de las posiciones a lo largo del genoma. Además, utilizando nuestra definición de cobertura, todas las posiciones con variación tendrán una cobertura menor. En los siguientes párrafos

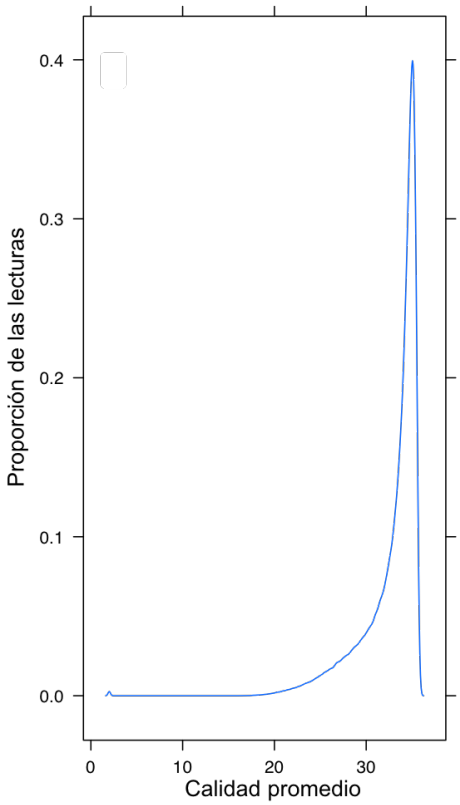


FIGURA 4.12: Se observa el número de GCAT añadido en cada ciclo del proceso de secuenciación

nos referimos a la "cobertura" en el sentido en el que se ha definido en este trabajo.

Al analizar la cobertura obtenida para cada uno de los proyectos de secuenciación de los tres individuos del trío, se observa un patrón interesante. La Figura 4.13 muestra intervalos de cobertura en el eje X y la fracción del genoma que se encuentra en ese intervalo determinado de cobertura en el eje Y. Se observa que aproximadamente 40 % de las posiciones genómicas en el caso de la muestra del padre y de la madre se secuenciaron con una cobertura de entre 31 y 40x, y entre 101 y 150X en el caso de la muestra de la hija. También se observa que la gran mayoría de las posiciones genómicas tienen una cobertura cercana a este valor, lo cual corresponde con el comportamiento esperado.

Sin embargo, en los experimentos se observa una distribución con una cola larga hacia la derecha, lo cual implica que hay algunos CSs que están en un número elevado de lecturas. Una proporción muy pequeña de CSs (en el orden de $10e-6$) se encuentran en más de 5,000 lecturas en el caso de la hija y en más de 1,000 lecturas en el caso de los padres. Aunque es una proporción pequeña, $10e-6$ corresponde a más de 20,000 CSs, los cuales, en su mayoría, son los mismos para cualquiera de los tres individuos y se encuentran dispersos a lo largo de todo el genoma. Recordando la suposición de que los CSs son secuencias únicas en el genoma

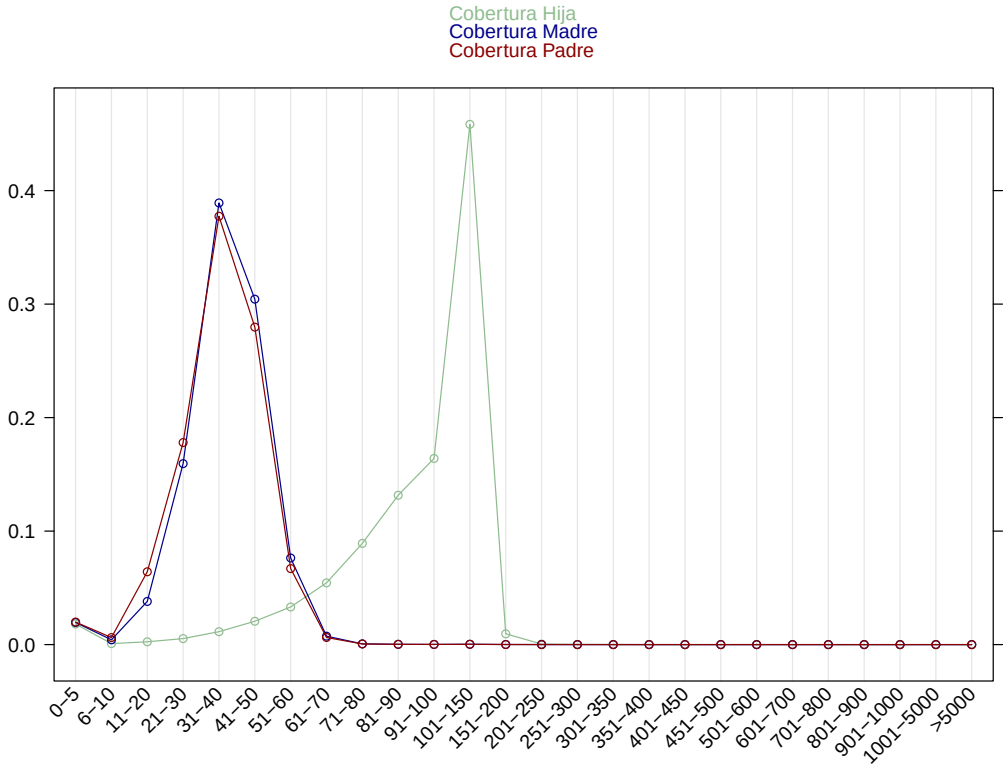


FIGURA 4.13: Se observa fracción del genoma secuenciada en un intervalo determinado de cobertura.

de referencia, estos resultados nos podrían indicar que existen regiones con problemas de ensamble o con un alto número de copias que se han colapsado en el genoma de referencia. Para evitar estas regiones COBASI incluye un filtro que elimina CSs con una cobertura mayor a cierto umbral.

Capítulo 5

MÉTODOS

5.1. Descripción del método COBASI

Una descripción general y detallada del método se proporciona en el cuerpo principal y en la sección de métodos el artículo que se adjunta en la sección de Resultados. Además, información adicional puede ser consultada en el material suplementario que se adjunta en el Apéndice A "Precise detection of de novo Single Nucleotide Variants in human genomes, SI".

5.2. Descripción de los experimentos de simulación

Una descripción general y detallada del método se proporciona en el cuerpo principal y en la sección de métodos el artículo que se adjunta en la sección de Resultados. Además, información adicional puede ser consultada en el material suplementario que se adjunta en el Apéndice A "Precise detection of de novo Single Nucleotide Variants in human genomes, SI".

5.3. Manual de usuario, método COBASI

El manual de usuario del método se encuentra en el Apéndice B "Manual de Usuario" del presente documento. Datos de ejemplo y el bash para realizar una corrida completa del software con estos datos pueden ser consultados en: https://github.com/Laura-Gomez/COBASI.

Capítulo 6

DISCUSIÓN Y CONCLUSIONES

Actualmente existen dos tipos de tecnologías de secuenciación; el primer tipo genera lecturas cortas de buena calidad mientras que el segundo produce lecturas largas con calidades bajas por base. Con el paso del tiempo, todas las tecnologías se han ido mejorando, sin embargo, para cada tipo se han mejorado distintos aspectos técnicos; para el caso de las tecnologías de lecturas cortas se ha logrado aumentar, de manera considerable, el rendimiento de los equipos, desarrollando máquinas con un rendimiento de hasta 100 gigabases de secuencia de alta calidad por carril en cada corrida de secuenciación, al mismo tiempo se ha logrado disminuir el precio y se espera que esta tendencia se mantenga en los años próximos; por otro lado, para el caso de las tecnologías de secuenciación que producen lecturas largas el enfoque ha sido extender el tamaño de las lecturas, desarrollando máquinas que pueden generar lecturas de hasta 100,000 nucleótidos con una baja calidad global.

Como es de esperarse, las aplicaciones para cada tipo de tecnología son diferentes. En el caso de la identificación de mutaciones puntuales, las tecnologías de alta calidad son las preferidas. Dado su bajo costo y alto rendimiento, la secuenciación se ha convertido en el recurso de elección por muchos laboratorios, lo cual se traduce en una generación masiva de datos con una tendencia a incrementar con el paso del tiempo. Dado lo anterior, la determinación precisa de variantes genómicas es de vital importancia, con el fin de reducir al mínimo la cantidad de tiempo necesario para realizar las validaciones experimentales.

COBASI representa una solución rápida y precisa al problema de identificación de variantes. Nuestro método está basado en el concepto de que la variación genética puede ser identificada con alta precisión al utilizar únicamente alineamientos perfectos de cadenas únicas. En este estudio, utilizamos cadenas únicas de 30 nucléotidos, con las cuales podemos interrogar cerca del 84 % del genoma de referencia, incluyendo secuencias repetitivas, como regiones de baja complejidad y duplicaciones segmentales de alta identidad. Para acceder a una proporción más alta del genoma sería necesario utilizar cadenas más grandes, sin embargo disminuiría la robustez del método ante la presencia de errores de secuenciación en las lecturas.

El "panorama de variación", construido en las primeras etapas de nuestro método, representa una herramienta poderosa para identificar con alta precisión regiones de polimorfismo al reconocer cambios bruscos en la cobertura local. Durante este trabajo, se probó que estas diferencias abruptas son robustas a las fluctuaciones intrínsecas en la cobertura de cualquier proyecto de secuenciación. Además, el proceso encargado de generar dicho "panorama de

variación" es un proceso rápido y computacionalmente eficiente. Adicionalmente, este panorama representa una descripción completa de la cobertura a lo largo del genoma con una resolución a nivel de un nucleótido.

La tarea de identificar las variantes genéticas que existen en un genoma problema sería trivial si el proceso de secuenciación fuera totalmente perfecto e infalible y sí, además, la longitud de las lecturas permitiera identificar su posición de origen sin ambigüedad. Sin embargo, en varios pasos durante este proceso existen factores, tanto técnicos como biológicos, que pueden confundir los resultados. El mapeo de las lecturas hacia el genoma de referencia se dificulta por factores como la existencia de secuencias de baja complejidad, regiones altamente repetidas o hipervariables en algunos genomas, así como por la introducción de errores de secuenciación, los cuales pueden presentar cierto sesgo hacia algún contexto genómico particular o alguna posición específica en las lecturas. Por otro lado, el proceso de asignación de genotipo puede complicarse debido a los errores de secuenciación (una vez más) y a los sesgos en la proporción en la que se secuencian distintas regiones genómicas o los distintos alelos para la misma región.

La identificación de variantes *de novo* es una tarea particularmente difícil debido a que cualquier variante falso positivo en el hijo(a) o cualquier variante falso negativo en los padres podría resultar en una variante incorrectamente identificada como *de novo*. Para resolver este problema se han desarrollado varios algoritmos especializados que analizan los datos de secuencia de todos los individuos de la familia al mismo tiempo. Estos algoritmos están basados en la existencia de una probabilidad anterior que describe el proceso de generación de las mutaciones *de novo*, la cual se utiliza para calcular la probabilidad de que cada mutación *de novo* haya sido correctamente identificada (11, 25). Dichos algoritmos deben ser entrenados utilizando un conjunto de variantes previamente clasificadas como positivas y negativas, de las cuales se derivan varias métricas de calidad [10]. Además, en reportes previos se han utilizado una gran cantidad de muestras poblacionales para eliminar artefactos producidos por el proceso de secuenciación, junto con filtros de calidad rigurosos para identificar variantes *de novo* reales [7][12][11][72].

La estrategia presentada en este trabajo se basa en el tipo de alineamientos más confiables: alineamientos perfectos entre cadenas única del genoma seguido de un análisis de la gráfica de cobertura. Otros algoritmos se basan en alineamientos menos confiables, los cuales pueden abarcar regiones repetitivas, y los cuales necesitan establecer límites de probabilidad para cuantificar la calidad de los descubrimientos.

El rendimiento de COBASI se midió utilizando experimentos de simulación. Para el llamado de SNVs en un individuo se obtuvo una AUPR de 0.94 y 0.96 en un experimento con una profundidad de secuenciación de 35x y 100x, respectivamente. En la mayoría de los casos, la AUPR reportada para COBASI fue similar a AUPRs reportadas en estudios previos [66], incluso cuando en reportes previos se han utilizado únicamente datos de exoma, los cuales representan el 2 % del genoma. Para el llamado de SNVs *de novo* utilizando COBASI se obtuvo una precisión de 1.0 y una sensibilidad de 0.91; por otro lado se obtuvo una precisión de 0.89 y una sensibilidad de 1 utilizando métodos comúnmente utilizado por la comunidad científica

los cuales están basados en alineamientos. Por lo tanto, se puede concluir que COBASI logra un buen balance entre el incremento de la precisión a expensas de una pequeña disminución en la sensibilidad. Además, es importante hacer notar que COBASI se probó en el genoma accesible, el cual constituye el 84 % del genoma. Por último, se demostró que COBASI es más rápido que algunos métodos basados en alineamientos.

La identificación precisa de variantes utilizando COBASI se basa en alineamientos globales que incluyen el sitio variable y dos cadenas únicas, cada una bordeando el sitio variante. Debido al tamaño reducido de las lecturas, únicamente pequeñas inserciones o deleciones producirán estos alineamiento de alta calidad. Para lograr una resolución adecuada en estos casos se requieren alineadores y algoritmos de detección especializados. En este momento, estos algoritmos no están incluidos en COBASI; su inclusión podría ser una mejora futura al método actual.

Los recursos computacionales y tiempo requeridos por COBASI permiten su aplicación rutinaria. Generar una lista de SNVs, a nivel de genoma completo, a partir de los datos crudos de secuenciación para un experimento con una cobertura 35x requiere alrededor de 40 horas en un servidor con 12 núcleos y 64 Gb de memoria RAM. Además, el "panorama de variación" del genoma humano completo se puede obtener en tan sólo 8 horas. Por último, si el análisis se restringe a sólo algunas regiones de interés, el tiempo requerido para generar un lista de SNVs puede reducirse considerablemente.

En este trabajo se analizó un proyecto de secuenciación de genoma completo de un trío padres-descendencia, el cual fue secuenciado a una profundidad promedio de 35x para cada uno de los padres y 100x para la hija. Durante el desarrollo del proyecto no se asumió ninguna probabilidad anterior para la generación de mutaciones *de novo*. Aunque en este trabajo sí se aplicaron filtros de cobertura en ningún momento se aplicaron filtros de calidad sobre las lecturas. Aún así, no se encontraron falsos positivos en las predicciones de SNVs *de novo* o en los SNVs mendelianos seleccionados al azar para validación experimental. Como resultado de este trabajo, se identificaron 58 SNVs *de novo*, lo cual es consistente con el número de SNVs *de novo* esperados de acuerdo a trabajos previos, los cuales reportan una tasa de mutación germinal de entre 1.0-1.8x10-8 [73] [7] por nucleótido por generación, lo cual se traduce en 44-82 SNVs *de novo* por individuo.

Esto se logró porque nuestro método combina un descubrimiento de SNVs sensible en el genoma de la hija con una validación exhaustiva en ambos padres. Sin embargo, el número de variantes identificadas podría ser un subestimado debido a que COBASI puede interrogar únicamente el 84 % del genoma. No obstante, con una capacidad de secuenciación mundial que tiende hacia los cientos de miles de genomas cada año [23], nuestro mayor interés es en maximizar la precisión de las variantes identificadas logrando disminuir lo más posible la cantidad de validación experimental requerida.

Los algoritmos comúnmente utilizados para identificar variantes han ocupado varias estrategias para eliminar artefactos en los distintos pasos del proceso. Una de ellas consiste en

realinear las lecturas alrededor de los indeles, recalibrar los puntajes de calidad por base y calcular la probabilidad de los distintos genotipos tomando en cuenta frecuencias poblacionales; otra estrategia consiste en generar haplotipos por cada muestra para facilitar la asignación de genotipo. Durante ambas estrategias se requiere información adicional a la muestra que está siendo procesada, en el primer caso se requiere una base de datos de variantes conocidas y frecuencias poblacionales, y en el segundo caso se necesita que muchas muestras sean analizadas al mismo tiempo. COBASI no requiere información adicional durante el proceso de identificación de variantes, es decir, no requiere frecuencias poblaciones y puede ser aplicado en una muestra aislada.

Recientemente, algunas publicaciones se han enfocado en resolver el problema de identificación de variantes utilizando estrategias sin alineamientos. SNVs conocidos se han identificado a partir de las lecturas de un proyecto de secuenciación si se detecta la presencia de cadenas únicas que contienen el alelo alternativo [54]. La transformada de Burrows–Wheeler de las lecturas se ha utilizado para identificar SNVs utilizando diferencias en la frecuencia de cadenas con ciertas características[53]. Además, cambios en la frecuencia de cadenas se han utilizado para reconstruir haplotipos de regiones que contienen variantes largas, concentrándose en regiones específicas del genoma [55]. Un trabajo publicado recientemente por nuestro grupo usa la frecuencia de cadenas para identificar variantes en genomas naturales y cromosomas sintéticos de cepas haploides de levadura [74]. Sin embargo, ningún trabajo anterior se ha concentrado en encontrar SNVs *de novo* en genomas humanos completos.

COBASI podría ser utilizado para identificar SNVs en distintos organismos dado que la aplicación exitosa de este método está limitada únicamente por la ploidía del organismo problema y por la fracción de su genoma que puede ser representado por cadenas únicas. Además, este método podría ser utilizado para analizar CSs derivados de regiones de interés, tal como un panel de genes relacionados con cáncer, acelerando el tiempo de análisis.

Por último la documentación de COBASI es extensiva permitiendo a cualquier investigador replicar el algoritmo descrito en este trabajo. Además, el método está dividido en módulos de forma que mejoras futuras puedan ser fácilmente incorporadas. Todo el software desarrollado se encuentra disponible en: https://github.com/Laura-Gomez/COBASI.

Capítulo 7

PERSPECTIVAS

La identificación de variantes es sumamente importante. Este proceso se puede volver complicado cuando se conjuntan diversos factores tanto técnicos como biológicos. Por un lado, las tecnologías de secuenciación distan de ser perfectas, y el desarrollo tecnológico en esa área cada vez se inclina más hacia producir lecturas de mayor tamaño a costa de un declive en la calidad de las mismas. Por otro lado, existen muchos organismos poliploides para los que la asignación de genotipo no es directa. El desarrollo de tecnologías de bajo costo con una calidad mayor ayudaría a facilitar el proceso de identificación de variantes.

Por otro lado, la aplicación de varias tecnologías de secuenciación en el mismo proyecto se ha utilizado para reducir los sesgos existentes. De igual manera, realizar el proceso de identificación de variantes utilizando varios algoritmos y comparando sus resultados puede ser utilizado para darle mayor peso y confiabilidad a los resultados encontrados. Con respecto a lo anterior, COBASI representa una solución ortogonal, que utiliza una premisa totalmente diferente a los algoritmos existentes, por lo que podría utilizarse como una técnica de comprobación independiente.

Un reto en el área de identificación de variantes es la veracidad y calidad de los genomas de referencia. Durante este trabajo se encontraron algunas regiones que consistentemente presentaban una cobertura altísima en los tres individuos secuenciados. Alrededor de una proporción de $10e-6$ CSs (más de 20,000 CSs dispersos a lo largo de todo el genoma) se encuentran en más de 5,000 lecturas en la hija y en más de 1,000 lecturas en los padres. Esto podría deberse a que estas regiones pertenecen a secuencias idénticas que han sido concatenadas en el genoma de referencia. En este trabajo estos CSs se filtraron. Sin embargo, sería interesante analizar las causas de este fenómeno e investigar sí esto reproducible en una muestra más grande de individuos con distintas ancestrías. .

Por otro lado, la detección de indeles es un problema de investigación abierto. Programas especializadas como Scalpel utilizan técnicas de micro ensamble para detectar indeles con alta precisión. COBASI utiliza un alineador global no especializado en resolver indeles por lo que con esta versión del algoritmo es imposible detectar indeles mayores a una decena de bases. Una posible extensión de COBASI podría consistir en ensamblar las lecturas que presentan un porcentaje definido de CSs de una región blanco particular.

Durante el análisis de los panoramas genómicos se observó un patrón particular: firmas de variación con tres niveles en el patrón de escalera al inicio o al final de la misma. Este

comportamiento indica la existencia de tres alelos para esa posición determinada. En el caso de un organismo diploide, como humano, se esperan dos alelos por cada posición genómica. Eventos raros como mosaicismo o mutaciones somáticas presentes en un alto porcentaje de la población celular podrían ser los causantes de este tipo de patrones. Por lo tanto, este tipo de firmas abre la posibilidad de aplicar el método COBASI, y el principio en el que subyace, para identificar algunos tipos de mosaicismo o mutaciones somáticas. Por otro lado, esta clase de patrón también se produce cuando existen diferencias puntuales en segmentos genómicos que se han duplicado en el individuo secuenciado.

El área de identificación de variantes es un área en continuo cambio; distintas estrategias se han adoptado para resolver el problema desde una gran variedad de enfoques. Esto ha resultado en un número considerable de métodos disponibles para identificar variantes. La gran mayoría de estos algoritmos están basados en alineamientos; sin embargo, recientemente se han desarrollado varios métodos que explotan los cambios en frecuencia de subcadenas del genoma de referencia. En el caso de COBASI se probó que su rendimiento es tan bueno como el de los métodos basados en alineamientos, incluso con una reducción en el tiempo del proceso. Esperamos que esta área siga en desarrollo, aumentando la sensibilidad de estos métodos, sin disminuir su precisión.

El principio subyacente de COBASI puede ser aplicado en distintas áreas de investigación. En el área de diagnóstico clínico, se podrían identificar aquéllas lecturas que pertenecen a ciertos genes o regiones genómicas de interés, acelerando el análisis de los datos. En el campo de la transcriptómica se podrían generar subconjuntos de cadenas que identifiquen, sin ambigüedad, un gene particular o un exon específico. En el análisis de splicing alternativo, se podrían identificar CSs que contengan las uniones exon-exon, y que sean únicos para cada isoforma. En el área de la metagenómica se podrían identificar CSs únicos para cada una de las especies microbianas de interés, a partir de los cuales se podrían hacer estudios de diversidad. Esto nos habla de la gran versatilidad del principio fundador de COBASI. Esperamos que este principio se utilice en la creación de distintos algoritmos que aprovechen sus características únicas.

Bibliografía

1. *Initial sequencing and analysis of the human genome* **409.** 6822. International Human Genome Sequencing Consortium (Nature, 2001), 860.
2. *Finishing the euchromatic sequence of the human genome* **431.** 7011. International Human Genome Sequencing Consortium (Nature, 2004), 931.
3. *A global reference for human genetic variation* **526.** 7571. 1000 Genomes Project Consortium (Nature, 2015), 68.
4. *A map of human genome variation from population-scale sequencing* **467.** 7319. 1000 Genomes Project Consortium (Nature, 2010), 1061.
5. *An integrated map of genetic variation from 1,092 human genomes* **491.** 7422. 1000 Genomes Project Consortium (Nature, 2012), 56.
6. Acuna-Hidalgo, R., Veltman, J. A. y Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome biology* **17,** 241 (2016).
7. Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A. y Jonasdottir, A. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471 (2012).
8. Wong, W. S., Solomon, B. D., Bodian, D. L., Kothiyal, P., Eley, G., Huddleston, K. C., Baker, R., Thach, D. C. e Iyer, R. New observations on maternal age effect on germline de novo mutations. *Nature communications* **7,** 10486 (2016).
9. Goldmann, J. M., Wong, W. S., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A. B., Glusman, G., Vissers, L. E., Hoischen, A., Roach, J. C. *y col.* Parent-of-origin-specific signatures of de novo mutations. *Nature genetics* **48,** 935 (2016).
10. Michaelson, J. J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A. *y col.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151,** 1431-1442 (2012).
11. Girard, S. L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., Dionne-Laporte, A., Spiegelman, D., Henrion, E., Diallo, O. *y col.* Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature genetics* **43,** 860 (2011).
12. Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., Ercan-Sencicek, A. G., DiLullo, N. M., Parikshak, N. N., Stein, J. L. *y col.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237 (2012).
13. Jin, S. C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S. R., Zeng, X., Qi, H., Chang, W., Sierant, M. C. *y col.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nature genetics* **49,** 1593 (2017).

14. Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A. y Waterston, R. H. DNA sequencing at 40: past, present and future. *Nature* **550,** 345 (2017).

15. Zhang, J., Chiodini, R., Badr, A. y Zhang, G. The impact of next-generation sequencing on genomics. *Journal of genetics and genomics* **38,** 95-109 (2011).

16. Goodwin, S., McPherson, J. D. y McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17,** 333 (2016).

17. Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G. *y col.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78-81 (2010).

18. Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J. A., Costa, G., McKernan, K. *y col.* A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome research* **18,** 1051-1063 (2008).

19. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z. *y col.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376 (2005).

20. Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M. *y col.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475,** 348 (2011).

21. Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S. y Bayley, H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature nanotechnology* **4,** 265 (2009).

22. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. *y col.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133-138 (2009).

23. Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S. y Robinson, G. E. Big data: astronomical or genomical? *PLoS biology* **13,** e1002195 (2015).

24. Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C. y Jaffe, D. B. Characterizing and measuring bias in sequence data. *Genome biology* **14,** R51 (2013).

25. Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., Topol, E. J., Levy, S. *y col.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome biology* **10,** R32 (2009).

26. Nielsen, R., Paul, J. S., Albrechtsen, A. y Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **12,** 443 (2011).

27. DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M. *y col.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43,** 491 (2011).

28. Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V. y Gibrat, J.-F. Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *Journal of Computational Biology* **19,** 796-813 (2012).

29. Langmead, B., Trapnell, C., Pop, M. y Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10,** R25 (2009).

30. Langmead, B. y Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9,** 357 (2012).

31. Lunter, G. y Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research* **21,** 936-939 (2011).

32. Ning, Z., Cox, A. J. y Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome research* **11,** 1725-1729 (2001).

33. Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K. y Wang, J. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966-1967 (2009).

34. Jiang, H. y Wong, W. H. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24,** 2395-2396 (2008).

35. David, M., Dzamba, M., Lister, D., Ilie, L. y Brudno, M. SHRiMP2: sensitive yet practical short read mapping. *Bioinformatics* **27,** 1011-1012 (2011).

36. Rizk, G. y Lavenier, D. GASSST: global alignment short sequence search tool. *Bioinformatics* **26,** 2534-2540 (2010).

37. Rivals, E., Salmela, L., Kiiskinen, P., Kalsi, P. y Tarhio, J. *MPSCAN: fast localisation of multiple reads in genomes* en *International Workshop on Algorithms in Bioinformatics* (2009), 246-260.

38. Li, H. y Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26,** 589-595 (2010).

39. Simola, D. F. y Kim, J. Sniper: improved SNP discovery by multiply mapping deep sequenced reads. *Genome biology* **12,** R55 (2011).

40. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. y Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078-2079 (2009).

41. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987-2993 (2011).

42. Ebbert, M. T., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., Duce, J., Kauwe, J. S. y Ridge, P. G. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC bioinformatics* **17,** 239 (2016).

43. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. *y col.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20,** 1297-1303 (2010).

44. Narzisi, G., O'rawe, J. A., Iossifov, I., Fang, H., Lee, Y.-h., Wang, Z., Wu, Y., Lyon, G. J., Wigler, M. y Schatz, M. C. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nature methods* **11,** 1033 (2014).

45. Ewing, B., Hillier, L., Wendl, M. C. y Green, P. Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. *Genome research* **8,** 175-185 (1998).

46. Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J. *y col.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60 (2008).

47. Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K. y Wang, J. SNP detection for massively parallel whole-genome resequencing. *Genome research* **19,** 1124-1132 (2009).

48. Martin, E. R., Kinnamon, D., Schmidt, M. A., Powell, E., Zuchner, S. y Morris, R. SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics* **26,** 2803-2810 (2010).

49. Browning, S. R. y Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* **81,** 1084-1097 (2007).

50. Howie, B. N., Donnelly, P. y Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5,** e1000529 (2009).

51. Compeau, P. E., Pevzner, P. A. y Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nature biotechnology* **29,** 987 (2011).

52. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. y McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics* **44,** 226 (2012).

53. Kimura, K. y Koike, A. Ultrafast SNP analysis using the Burrows–Wheeler transform of short-read data. *Bioinformatics* **31,** 1577-1583 (2015).

54. Pajuste, F.-D., Kaplinski, L., Möls, M., Puurand, T., Lepamets, M. y Remm, M. FastGT: an alignment-free method for calling common SNVs directly from raw sequencing reads. *Scientific reports* **7,** 2537 (2017).

55. Audano, P., Ravishankar, S. y Vannberg, F. Mapping-free variant calling using haplotype reconstruction from k-mer frequencies. *Bioinformatics* **10,** 1659-1665 (2017).

56. Reyes, J., Gómez-Romero, L., Ibarra-Soria, X., Palacios-Flores, K., Arriola, L. R., Wences, A., Garcıa, D., Boege, M., Dávila, G., Flores, M. *y col.* Context-dependent individualization of nucleotides and virtual genomic hybridization allow the precise location of human SNPs. *PNAS* **108,** 15294-15299 (2011).

57. Wolf, B., Kuonen, P. y Dandekar, T. *GNATY: Optimized NGS Variant Calling and Coverage Analysis* en *International Conference on Bioinformatics and Biomedical Engineering* (2016), 446-454.

58. Kelly, B. J., Fitch, J. R., Hu, Y., Corsmeier, D. J., Zhong, H., Wetzel, A. N., Nordquist, R. D., Newsom, D. L. y White, P. Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. *Genome biology* **16,** 6 (2015).

59. Francioli, L. C., Cretu-Stancu, M., Garimella, K. V., Fromer, M., Kloosterman, W. P., Wijmenga, C., Swertz, M. A., van Duijn, C. M., Boomsma, D. I., Slagboom, P. *y col.* A framework for the detection of de novo mutations in family-based sequencing data. *European Journal of Human Genetics* **25,** 227 (2017).

60. Liu, Y., Loewer, M., Aluru, S. y Schmidt, B. SNVSniffer: an integrated caller for germline and somatic single-nucleotide and indel mutations. *BMC systems biology* **10,** 47 (2016).

61. Salzberg, S. L., Pertea, M., Fahrner, J. A. y Sobreira, N. DIAMUND: Direct comparison of genomes to detect mutations. *Human mutation* **35,** 283-288 (2014).

62. O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W. E. *y col.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome medicine* **5,** 28 (2013).

63. Zook, J. M. y Salit, M. Genomes in a bottle: creating standard reference materials for genomic variation-why, what and how? *Genome biology* **12,** P31 (2011).

64. Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W. y Salit, M. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature biotechnology* **32,** 246 (2014).

65. Cornish, A. y Guda, C. A comparison of variant calling pipelines using genome in a bottle as a reference. *BioMed research international* **2015,** 1-11 (2015).

66. Hwang, S., Kim, E., Lee, I. y Marcotte, E. M. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific reports* **5,** 17875 (2015).

67. Hu, X., Yuan, J., Shi, Y., Lu, J., Liu, B., Li, Z., Chen, Y., Mu, D., Zhang, H., Li, N. *y col.* pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics* **28,** 1533-1535 (2012).

68. Rogozin, I. B. y Pavlov, Y. I. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Mutation Research/Reviews in Mutation Research* **544,** 65-85 (2003).

69. Hodgkinson, A. y Eyre-Walker, A. Human triallelic sites: evidence for a new mutational mechanism? *Genetics* **184,** 233-241 (2010).

70. Detailed description of HaplotypeCaller; best reference for germline joint calling. *bioRxiv* (2017).

71. Li, W. G+ C content evolution in the human genome. *eLS* (2013).

72. Besenbacher, S., Liu, S., Izarzugaza, J. M., Grove, J., Belling, K., Bork-Jensen, J., Huang, S., Als, T. D., Li, S., Yadav, R. *y col.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nature communications* **6,** 5969 (2015).

73. Campbell, C. D. y Eichler, E. E. Properties and rates of germline mutations in humans. *Trends in Genetics* **29,** 575-584 (2013).

74. Palacios-Flores, K., Garcıa-Sotelo, J., Castillo, A., Uribe, C., Aguilar, L., Morales, L., Gómez-Romero, L., Reyes, J., Garciarubio, A., Boege, M. *y col.* A perfect match genomic landscape provides a unified framework for the precise detection of variation in natural and synthetic haploid genomes. *Genetics,* genetics-300589 (2018).

Apéndice A

Precise detection of *de novo* SNVs in human genomes, SI

En este apéndice se adjunta el material suplementario de la publicación relacionada con el software COBASI. Este material suplementario está enfocado hacia la comunidad científica nacional e internacional por lo que se presenta en inglés, tal como fue publicado en el artículo original.

Los cromatogramas incluidos en el material suplementario del artículo original no se incluyeron en este escrito. Dichos cromatogramas pueden ser consultados en el material suplementario en la página oficial de PNAS.

SUPPLEMENTARY METHODS

Definition of CS regions from the RG. The RG hg38 (GRCh38) was downloaded from the UCSC Golden Path website. All kmers (with a sliding window of 1 bp, k=30 nt) from GRCh38 were obtained, and the unique kmers (CSs) were retrieved using Bowtie (1). A list containing the start and end positions of regions composed only by CSs was obtained.

Criteria to define inconsistent reads. The reads with: i) a different strand assigned to the PrevCS and PostCS alignment, ii) multiple alignments of the same CS, iii) and no PrevCS alignment were saved to an inconsistencies file and were not used in subsequent analyses

Genotype assignment. The probability of three possible genotypes was computed: homozygous reference (homo-R), heterozygous reference/no-reference (hete-R/NR), and homozygous non-reference (homo-NR). If the assigned genotype was different than homo-R, the allele with the highest frequency other than the reference was obtained and it was considered to be the major allele. Additionally, three probabilities were computed: homozygous major (homo-M), heterozygous major/no major, and homozygous no major (homo-NM). If the genotypes with the highest probability were either homo-NR and homo-NM or hete-R/NR and hete-M/NM or hete-R/NR and homo-NM, there were at least three probable alleles that could be assigned with high probability as the genotype at that particular site. This resulted in an ambiguous genotype calling. Ambiguous SNVs were written to an ambiguous genotype file, and no further analysis was done with these variant sites. For genotype assignment, the reads with the allele N were not taken into account.

Criteria to define a *bona fide de novo* variant. The criteria for defining a possible variant, such as a *bona fide de novo* variant, were: 1) at least 10 reads spanning that site in the child and at least 10 reads in each parent, at least 2 total alignments in the child and at least 2 in each parent; 2) the variant allele should not be contained in more than one high-quality alignment (total alignments) in any parent; 3) the variant allele of the child should be in more than one-fourth of all the reads spanning that site or the variant region could be duplicated in the child genome (15); and 4) the candidate *de novo* SNV must be absent from public SNV databases, such as dbSNP.

Definition of *accessible genome*. All 100-nt windows (with a sliding window of 1 bp) were obtained. For each window, the number of CSs was computed. All consecutive windows with a CS density higher than 0.5 were concatenated, creating a CS-accessible-region. The CS-accessible-regions constituted the callable genome, except for k nucleotides at the start and end of each region. For all simulations and real sequencing data results, only SNVs in the accessible genome were reported.

Simulation experiments. SNVs were introduced into chromosome 12 with a mutation rate of 0.001. The position for every variant site was chosen at random, in addition to the phase and the alternative allele. We used the ART Simulator to generate sequencing reads using the HiSeq Illumina error profile (100 bp paired-end reads) (2), and we applied the COBASI pipeline to call the SNVs. We varied several key parameters, such as sequencing depth, kmer size, minimum coverage for the Signature CSs, absolute value for the RCI, maximum difference in coverage between the Signature CSs, minimum number of whole-VSR alignments, and optimal extension for the partial alignments. To compute Precision-Recall curves, we obtained the number of False-Negative (FN), False-Positive (FP), and True-Positive (TP) calls at different coverage thresholds for each set of parameters. We calculated the Area Under the Curve (AUPR) as a performance score.

In the case of the parent-offspring simulation, SNVs were introduced into chromosome 12 with a mutation rate of 0.001 to create the father diploid chromosome. The position for every variant site was randomly chosen, as well as the phase and the alternative allele. To create the mother diploid chromosome, for every variant site for the father, the phase for the mother was chosen at random. One father chromosome and one mother chromosome were chosen to create the child's pair of chromosomes. *De novo* mutations were introduced in positions not previously mutated in the child with a mutation rate of 3e-7 (39 SNVs). The *de novo* mutation rate was artificially increased to yield a considerable amount of *de novo* SNVs. For all three individuals, we used the ART Simulator to generate sequencing reads using the HiSeq Illumina error profile (100 bp paired-end reads). The coverage for each individual was chosen to resemble our real sequencing experiments, 35x coverage for each parent and 100x coverage for the child. We applied the COBASI pipeline to discover the *de novo* SNVs

with the set of parameters that maximizes the APR for each sequencing depth (obtained from one individual simulation). For the child: sequencing depth = 100x, kmer size = 30, minimum coverage for the Signature CSs = 10, absolute value for the RCI = 0.2, maximum difference in coverage between the Signature CSs = 2.0, minimum number of whole-VSR alignments = 3, optimal extension for the partial alignments = 10. For the parents: sequencing depth = 35x, kmer size = 30, minimum coverage for the Signature CSs = 5, absolute value for the RCI = 0.2, maximum difference in coverage between the Signature CSs = 2.0, minimum number of whole-VSR alignments = 2, optimal extension for the partial alignments = 5. We repeated this simulation experiment 20 times and obtained the median values for the FP, FN, and TP calls.

Variant calling using alignment-based pipelines. The best practices guideline (3) was followed to call SNV from 5 (chosen at random) out of the 20 simulations: reads were mapped using BWA, duplicate reads were removed using Picard, local realignment around indels was done, base quality score was recalibrated, genotypes were assigned using GATK HaplotypeCaller and variants were filtered using a hard filter. Finally, *de novo* variants were identified using GATK VariantAnnotator.

TRIO sequencing and COBASI application. DNA from whole blood was extracted using the QIAmp DNA Blood Mini Kit as described by the manufacturer. Three libraries were prepared for the child and one for each parent. The CODIS STRs were determined for each individual. The DNA libraries were sequenced by paired-end Illumina HiSeq 2000 with a read length of 100 bp. The COBASI pipeline was used to discover *de novo* SNVs from the TRIO sequencing data using the same parameters as in *de novo* simulations.

Experimental validation of *de novo* SNVs. PCR primers were designed using the Oligo7 software and manual inspection. PCR was performed using the Accuprime Pfx kit according to the manufacturer's instructions. PCR products were sequenced by Sanger sequencing at Macrogen, Inc. To determine the specific position corresponding to the nucleotide of interest, each Sanger sequence was aligned to the RG using BLAST (4). The genotypes of the sites of interest were determined by manual inspection of the chromatograms.

Probability of one mutation occurring independently at the same site in two unrelated genomes. . There is some disagreement about the human mutation rate (5). However, for the sake of argument, we will assume the worst-case scenario. This means the highest mutation rate, which means 80 new mutations per haploid genome. (6, 7). There are $\binom{3.2\times10^9}{80}$ ways to select a set of 80 mutated base pairs in a genome of length nucleotides. Of these, $\binom{3.2\times10^9-1}{79}$ contain a fixed base pair. Therefore, the probability of a fixed base pair being contained in the set of 80 mutations of a genome is 2.5×10^{-8} (we recover the mutation rate). The probability of any fixed base pair being contained in the set of mutations of two independent genomes is $(2.5\times10^{-8})^2 = 6.5\times10^{-16}$, which is very low. Because of this, any *de novo* SNV is not expected to be found in any population SNV database.

SUPPLEMENTARY TABLES

TABLE S1. Comparison of genomic regions defined as accessible by COBASI and the 1000 Human Genomes Project

DENSITY CUTOFF	BOTH	ONLY COBASI	ONLY 1000HGP	NEITHER	TOTAL COBASI	TOTAL 1000HGP
0	90	10	0	0	100	90
10	88	1	2	8	90	90
20	87	1	3	8	88	90
30	86	1	4	9	87	90
40	83	2	7	9	85	90
50	82	2	8	9	84	90

The callable genome by COBASI was defined in the Methods. In the 1000 Genomes Project, the "accessible genome" was defined based on coverage and mapping quality criteria. Regions with very high or low coverage, as well as many low-quality mapped reads, were defined as unaccessible regions. In the table, several CSs density cutoffs are shown, and the percentage of the genome that is defined as callable by 1) both projects, 2) only COBASI, 3) only 1000HGP, 4) neither project, 5) COBASI, or 6)1000HGP is shown.

TABLE S2. The Area Under the Curve for the Precision-Recall curves (APR) for the COBASI simulation in one individual, part I.

Parameters	35x	50x	75x	100x
k=25,RCIV=0.2,total-aln=2	0.932	0.933	0.927	0.915
k=30,RCIV=0.2,total-aln=2	**0.943**	0.952	0.951	0.946
k=25,RCIV=0.2,total-aln=3	0.931	0.939	0.943	0.943
k=30,RCIV=0.2,total-aln=3	0.928	0.951	0.959	**0.961**
k=25,RCIV=0.2,total-aln=4	0.923	0.936	0.942	0.944
k=30,RCIV=0.2,total-aln=4	0.900	0.942	0.956	0.960
k=25,RCIV=0.2,total-aln=6	0.887	0.927	0.938	0.941
k=30,RCIV=0.2,total-aln=6	0.790	0.910	0.947	0.955
k=30,RCIV=0.3,total-aln=2	0.933	0.949	0.951	0.946
k=25,RCIV=0.3,total-aln=2	0.925	0.932	0.926	0.915
k=30,RCIV=0.3,total-aln=3	0.921	0.948	0.959	0.961
k=25,RCIV=0.3,total-aln=3	0.925	0.938	0.942	0.943
k=30,RCIV=0.3,total-aln=4	0.895	0.940	0.956	0.960
k=25,RCIV=0.3,total-aln=4	0.917	0.935	0.942	0.944
k=30,RCIV=0.3,total-aln=6	0.790	0.909	0.947	0.955
k=25,RCIV=0.3,total-aln=6	0.884	0.926	0.938	0.941

One human chromosome (chromosome 12) was mutated (mutation rate = 0.001), and simulated reads were produced for this mutant chromosome. SNVs were called using COBASI by varying three parameters: the kmer size (k), the minimum relative change in coverage to identify a VSR (RCIV), and the minimum number of reads that should contain both SignatureCSs (total-aln). The Area under the curve for the Precision-Recall (APR) curves are shown. To compute the plots, the precision and recall were calculated for different coverage thresholds in a particular simulation. Invariant parameters over these simulations: the extension for alignments of reads containing only the PrevCS (n = 5) for all sequencing depths, the minimum coverage for any Signature CS (rmin = 5 for sequencing depths of 35 and 50´ and rmin = 10 for sequecning depths of 75 and 100´). The set of parameters chosen to perform the parent-offspring simulation are highlighted as bold numbers

TABLE S3. The Area Under the Curve for the Precision-Recall curves (APR) for the COBASI simulation in one individual, part II.

Parameters	35x	100x
ratio=1.5	0.919	0.961
ratio= 2.0	0.943	0.961
ratio=2.5	0.943	0.961
ratio=2.0, n=10	0.941	0.961
ratio=2.0, rmin=5	--	0.960
ratio=2.0, rmin=10	0.943	--

Parameters	35x	100x
ratio=1.5	9411 20	4502 3
ratio=2.0	**6620** **29**	**4451** **5**
ratio=2.5	6561 36	4421 15
ratio=2.0, n=10	6780 46	**4463** **5**
ratio=2.0, rmin=5	--	4591 5
ratio=2.0, rmin=10	6581 32	--

One human chromosome (chromosome 12) was mutated (mutation rate = 0.001) and simulated reads were produced for this mutant chromosome. SNVs were called using COBASI by varying three parameters: the minimum coverage for any Signature CS (rmin), a maximum ratio between the coverage of the Signature CSs (ratio), and the extension for alignments of reads containing only the PreCS (n). The Area under the curve for the Precision-Recall (APR) curves are shown. To compute the plots, the precision and recall were calculated for different coverage thresholds in a particular simulation. Invariant parameters over these simulations: 35x: k =30, RCIV = 0.2, total-aln = 2; 100x: k = 30, RCIV = 0.2, total-aln =3. Default parameters (otherwise mentioned): for all coverage thresholds: n= 5; 35x: rmin = 5 and 100x: rmin = 10. The left table contains the APR score for every simulation, and the right table contains the FN and FP for every simulation. The set of parameters chosen to perform the parent-offspring simulation are highlighted as bolded numbers

TABLE S4. Experimental validation of each predicted *de novo* SNVs.

CHR	POS	REF	FATHER	MOTHER	CHILD	STATUS
chr1	24862021	G	G/G	G/G	G/T	OK
chr1	90547932	G	G/G	G/G	GA	OK
chr1	167295816	A	A/A	A/A	A/G	OK
chr1	172427805	G	G/G	G/G	T/G	OK
chr1	207061328	G	G/G	G/G	G/T	NoPCR
chr1	233278131	A	A/A	A/A	A/G	OK
chr2	7834800	G	G/G	G/G	G/C	OK
chr2	24287324	T	T/T	T/T	T/C	OK
chr2	64935802	G	G/G	G/G	G/T	OK
chr2	117515206	A	A/A	A/A	A/G	OK
chr2	159087258	C	C/C	C/C	C/T	BQ
chr2	166134730	G	G/G	G/G	G/A	OK
chr2	174144299	C	C/C	C/C	C/T	OK
chr3	13257366	C	C/C	C/C	C/T	OK
chr3	35344598	T	T/T	T/T	T/A	OK
chr3	84019551	C	C/C	C/C	C/T	OK
chr3	85475191	G	G/G	G/G	G/C	OK
chr3	130405591	G	G/G	G/G	G/T	OK
chr3	154730842	A	A/A	A/A	A/G	OK
chr3	177039650	C	C/C	C/C	C/T	OK
chr3	193814289	G	G/G	G/G	G/T	OK
chr4	12050118	T	T/T	T/T	T/G	OK
chr4	122532439	C	C/C	C/C	C/T	OK
chr4	165308533	C	C/C	C/C	C/T	OK
chr4	183179287	C	C/C	C/C	C/T	OK
chr5	42087606	T	T/T	T/T	T/C	OK
chr6	54488698	A	A/A	A/A	A/T	OK
chr6	110925590	T	T/T	T/T	T/C	OK
chr6	145688494	A	A/A	A/A	A/G	OK
chr6	149023483	C	C/C	C/C	C/T	OK
chr7	8845957	C	C/C	C/C	C/A	OK
chr7	18840247	A	A/A	A/A	A/T	OK
chr7	131254278	G	G/G	G/G	G/T	OK
chr7	148217676	A	A/A	A/A	A/G	OK
chr8	38433070	G	G/G	G/G	G/A	OK
chr8	68845327	T	T/T	T/T	T/A	OK
chr9	74292655	A	A/A	A/A	A/T	OK
chr9	135134043	C	C/C	C/C	C/A	OK
chr10	967661	T	T/T	T/T	T/C	OK
chr10	69932637	A	A/A	A/A	A/C	OK
chr10	124545656	T	T/T	T/T	T/G	NoPCR*
chr11	46199782	A	A/A	A/A	A/C	NoPCR*
chr11	9834859	C	C/C	C/C	C/T	OK
chr11	22218005	G	G/G	G/G	G/T	OK
chr11	57031949	C	C/C	C/C	C/T	OK
chr11	66915741	A	A/A	A/A	A/G	OK
chr11	98890913	G	G/G	G/G	G/A	OK
chr11	120059843	A	A/A	A/A	A/C	OK

chr12	7422099	C	C/C	C/C	C/T	OK
chr13	78641958	C	C/C	C/C	C/T	OK
chr15	81812391	T	T/T	T/T	T/C	OK
chr16	76704617	C	C/C	C/C	C/T	PCRInesp
chr17	61212465	A	A/A	A/A	A/G	OK
chr19	7406505	A	A/A	A/A	A/G	OK
chr20	59356016	A	A/A	A/A	A/C	PrimInes
chrX	87169908	T	T/T	T/T	T/G	OK
chrX	125321179	A	A/A	A/A	A/G	OK

The table contains all the predicted *de novo* SNVs and the results of their experimental validation. Each row shows the chromosome, the geomic position, and the genotype predicted for each individual for each SNV. In the column "experimental status:" OK means that the Sanger sequencing results and the COBASI prediction are consistent for all the individuals. PrimInesp means that no specific primers could be designed because of the presence of a highly repetitive region surrounding the SNV. PCRInesp means that no unique PCR product could be obtained even when specific primers were designed. NoPCR means that no PCR product could be obtained. BQ means that no quality sequence could be obtained even when the sequencing was repeated several times, likely the result of the presence of low-complexity regions (long stretches of poli-dT) found in that specific region.

TABLE S5. Experimental validation for a subset of Mendelian SNVs.

ID	CHR	POS	REF	FATHER	MOTHER	CHILD	STATUS
1	1	108095723	A	G/G	G/G	G/G	OK
2	1	147610227	G	G/G	A/A	G/A	OK
3	2	19463414	G	G/G	G/A	G/A	OK
4	2	161057267	G	G/T	T/T	G/T	OK
5	3	4085479	T	T/T	T/C	T/C	OK
6	3	157221449	G	G/A	G/A	A/A	OK
7	4	107667929	C	C/G	C/C	C/G	OK
8	4	146842576	C	C/C	C/A	C/A	OK
9	5	44277000	C	C/T	C/C	C/T	OK
10	5	80058324	C	C/T	C/T	C/T	OK
11	6	67230929	G	G/G	G/A	G/A	OK
12	6	147785976	C	A/A	C/C	C/A	OK
13	7	77752055	C	G/G	G/G	G/G	OK
14	7	109782464	A	T/T	T/T	T/T	OK
15	8	21514268	T	T/C	T/C	C/C	OK
16	8	27092236	T	T/C	T/T	T/C	OK
17	9	4516070	C	C/T	C/T	C/T	OK
18	9	117229359	C	C/C	C/T	C/T	OK
19	10	8766364	C	C/A	C/A	A/A	OK
20	10	79241132	T	C/C	T/T	T/C	OK
21	11	7937566	C	C/G	C/C	C/G	OK
22	11	6722659	C	C/T	C/C	C/T	OK
23	12	17811890	G	G/T	G/T	G/T	OK
24	12	53864157	A	G/G	G/G	G/G	OK
25	13	74721621	A	G/G	G/G	G/G	OK
26	13	85333647	T	G/G	G/G	G/G	OK
27	14	20550811	G	A/A	G/A	A/A	OK
28	14	55442549	G	A/A	A/A	A/A	OK
29	15	78319135	G	A/A	A/A	A/A	OK
30	15	80916473	G	C/C	C/C	C/C	OK
31	16	5978486	C	C/C	C/G	C/G	OK
32	16	7924503	A	A/A	C/C	A/C	OK
33	17	8758708	A	A/A	A/G	A/G	OK
34	17	72131667	G	A/A	G/A	G/A	OK
35	18	5901790	C	C/G	C/G	G/G	OK
36	18	72951051	T	C/C	T/C	C/C	OK
37	19	19774267	C	C/A	A/A	C/A	OK
38	19	28737356	C	C/T	T/T	T/T	OK
39	20	10778727	T	C/C	T/C	C/C	OK
40	20	64264757	C	C/T	C/C	C/T	OK
41	21	28205983	A	G/G	G/G	G/G	OK
42	21	41715615	G	G/C	G/G	G/C	OK
43	22	23908608	C	G/G	C/G	C/G	OK
44	22	36265520	G	G/A	G/A	A/A	OK
45	X	8928469	T	T/T	C/C	T/C	OK
46	X	22643455	T	C/C	T/T	T/C	OK

The table contains a subset of Mendelian SNVs and the results of their experimental validation. Each row shows the chromosome, the genomic position, and the genotype predicted for each individual for each SNV. In the column, experimental status “OK” means that the Sanger sequencing results are consistent for all individuals.

TABLE S6. Computing time, core number, and RAM required for every stage for the COBASI approach.

	12 N[1] 64Gb RAM	12 N[1] 128Gb RAM	24 N[1] 64Gb RAM	24 N 128(RAI
1. ONE TIME PROCESS. CS DATABASE CREATION				
OBTAIN CS DATABASE				
Cut reference genome	00:32	00:32	00:17	00:1
Obtain unique kmers	01:26	01:26	00:58	00:5
Obtain non-overlapping kmers	00:13	00:13	00:10	00:1
TOTAL 1	**02:11**	**02:11**	**01:25**	**01:2**
2a. GENOME-WIDE SNV DISCOVERY				
OBTAIN LANDSCAPE				
Count kmers	02:04	02:36	01:52	01:2
Obtain whole-genome coverage	06:00	03:20	06:00	03:2
Obtain landscape	00:18	00:18	00:12	00:1
SUBTOTAL	**8:22**	**6:14**	**8:04**	**5:0**
GET SIGNATURE REGIONS AND SIGNATURE READS				
Get Variant Signature Regions	00:40	00:40	00:25	00:25
Obtain Signature CSs sequence	00:09	00:09	00:09	00:09
Get Signature Reads	06:00	06:00	06:00	06:00
FILTER READS AND GET SNVs				
Get SNVs	28:40	28:40	23:00	20:00
SUBTOTAL	**35:29**	**35:29**	**29:24**	**26:34**
TOTAL 2a	**43:51**	**41:43**	**37:28**	**31:35[3]**
2b. *DE NOVO*-ORIENTED SNV DISCOVERY (PARENTAL GENOMES)				
GET SIGNATURE REGIONS AND SIGNATURE READS				
Get Variant Signature Regions	00:40	00:40	00:25	00:25
Obtain Signature CSs sequence	01:09	01:09	01:09	01:09
Get Signature Reads	06:00	06:00	06:00	06:00
FILTER READS AND GET SNVs				
Get SNVs	01:51	01:51	01:25	01:10
TOTAL 2b	**9:40**	**9:40**	**8:59**	**8:44**

The first process of the COBASI pipeline is the CS database creation. This process must be done once per reference genome. To discover the *de novo* SNV, all SNVs must be called in the child (Stage 2a), and these positions must be interrogated in the parents (Stage 2b). For every step, the computation time required for different hardware specifications is shown.
[1]N denotes the number of processors.

[2]The whole-genome Variation Landscape can be generated in only 5 hours.
[3]A SNV list from the raw whole-genome sequencing data is generated in less than 36 hours.
[4]If only some regions of interest are chosen for further investigation, the COBASI approach can generate a list of resulting SNVs from the whole-genome sequencing raw data in less than 9 hours

SUPPLEMENTARY FIGURES

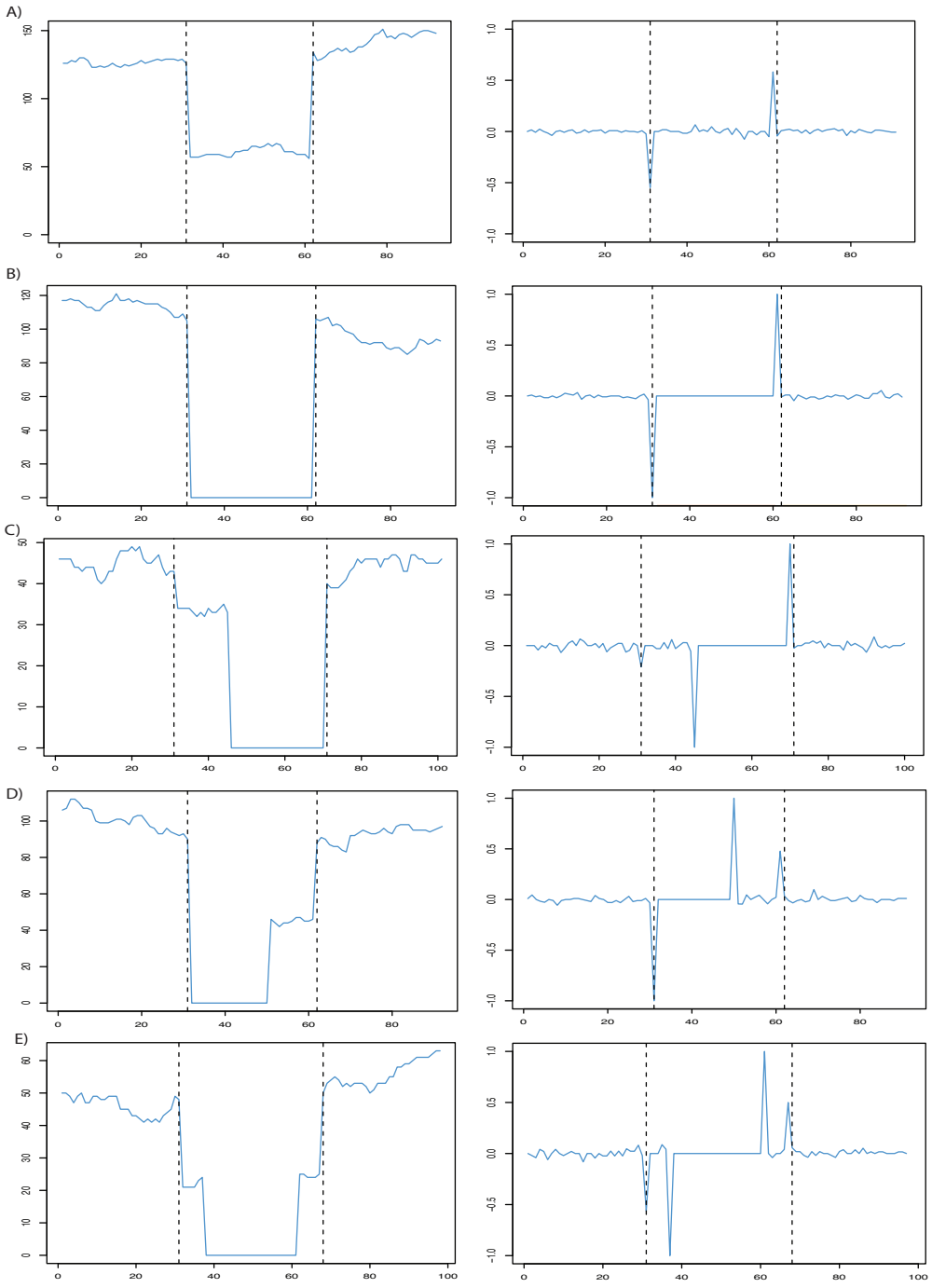


Fig. S1. DIFFERENT TYPES OF VARIANT SIGNATURE REGIONS (VSR). Several variants can be close enough to be concatenated on the same VSR. Depending on their zygosity and chromosomal localization, four different VSR patterns are found. A) A classic VSR formed by only one heterozygous SNV is shown. If there is more than one heterozygous SNV localized on the same chromosome, the SVR is extended. Position 1 on the X axis corresponds to chr7:9,449,337. B) A classic VSR formed by only one homozygous SNV is shown. If there is more than one homozygous SNV localized on the same chromosome, the SVR is extended. Position 1 on the X axis corresponds to chr21:21,616,422. C) A VSR formed when a heterozygous variant is followed by a homozygous SNV is shown. Position 1 on the X axis corresponds to chr17:83,187,030. D) A SNV formed when a homozygous SNV is followed by a heterozygous SNV is shown. Position 1 on the X axis corresponds to chr12:133,163,159. E) A VSR formed when two heterozygous SNVs localized on different chromosomes are found. Position 1 on the X axis corresponds to chr14:104,928,266. Left, the VL for a specific genomic region is shown. Every plot shows the start position of each CS (X axis) and the coverage for each CS (Y axis). Right, the RVL for the same regions is shown. Every plot shows the start position of each CS (X axis) and the RCI values associated with each CS (Y axis). The start positions for the PrevCS and PostCS are shown as dashed vertical lines. The VL and RVL depicted correspond to the child's genome.

SUPPORTING REFERENCES

1. Langmead B, Trapnell C, Pop M Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 485(7397):237-241.
2. Huang W, Li L, Myers JR, Marth GT (2012) ART: A next-generation sequencing read simulator. *Bioinformatics* 28(4):593-594.
3. Van der Auwera GA, Carneiro MO, Hartl C, *et al.* (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics.11(1110):11.10.1-11.10.33.
4. Altschul, SF, Gish, W, Miller, W, Myers, EW, Lipman, DJ. (1990) Basic local alignment search tool. *J Mol Biol.* 215:403-410.
5. Li B, et al. (2012) A likelihood-based framework for variant calling and *de novo* mutation detection in families. *PLoS Genet* 8(10):e1002944.
6. Michaelson JJ, et al. (2012). Whole genome sequencing in autism identifies hotspots for de novo germline mutation. Cell 151(7):1431-1442.
7. Sanders SJ, et al. (2012) *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485(7397):237-241.

Apéndice B

Manual de usuario

En este apéndice se adjunta el manual de usuario del software COBASI. Este manual fue creado para que la comunidad científica nacional e internacional pueda hacer uso del software. Por lo tanto, se presenta en inglés. Tal como se encuentra en la página GitHub del proyecto: `https://github.com/Laura-Gomez/COBASI`.

COBASI puede ser utilizado en dos modalidades: para identificar todos los SNVs de un individuo, o para identificar los SNVs *de novo* a partir de los datos de secuenciación de un trio familiar.

Los programas utilizados en ambos escenarios se describen en el manual de usuario adjunto, asi como cada uno de los parámetros utilizados.

User Guide
COBASI Pipeline
Universidad Nacional Autónoma de México

COBASI
V5.1

USER GUIDE

DEVELOPED BY
LAURA GOMEZ-ROMERO

JANUARY 2018

User Guide
COBASI Pipeline
Universidad Nacional Autónoma de México

CONTENT

Centro de Ciencias Genómicas

LIIGH-UNAM

User Guide
COBASI Pipeline
Universidad Nacional Autónoma de México

INTRODUCTION

The COBASI approach is a unique solution to the variant calling problem. This approach is used to generate a list of SNVs from raw whole genome sequencing data from one individual (One individual framework). Besides, it can be extended to be used in a familiy-based framework.

Get CS Regions

The VGH pipeline requires a list of all unique regions in the genome. These unique regions are defined as those regions in which every kmer inside that region (of a defined size) is a COIN-String (unique string). The input and output parameters are indicated in the next figure.

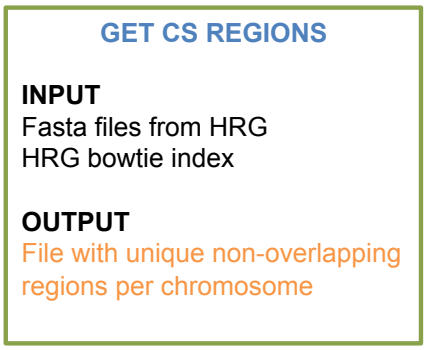


FIG 1. The input and output file for the stage: “Get CS Regions” are shown

One individual framework.

The VGH approach can be divided in four sequential stages. The four stages are illustrated in the next figure. For every stage the input and output parameters are indicated. It can be noted that the output files from one stage are used as input for the next one. Every stage is composed of several processes. In the first stage the Variation Landscape is computed (VL); in the second stage the Variation Siganture Regions are identified (VSR’s); in the third stage the Signature Reads are retrieved; and finallly, in the fourth stage a list of Single Nucleotide Variants (SNV’s) is generated.

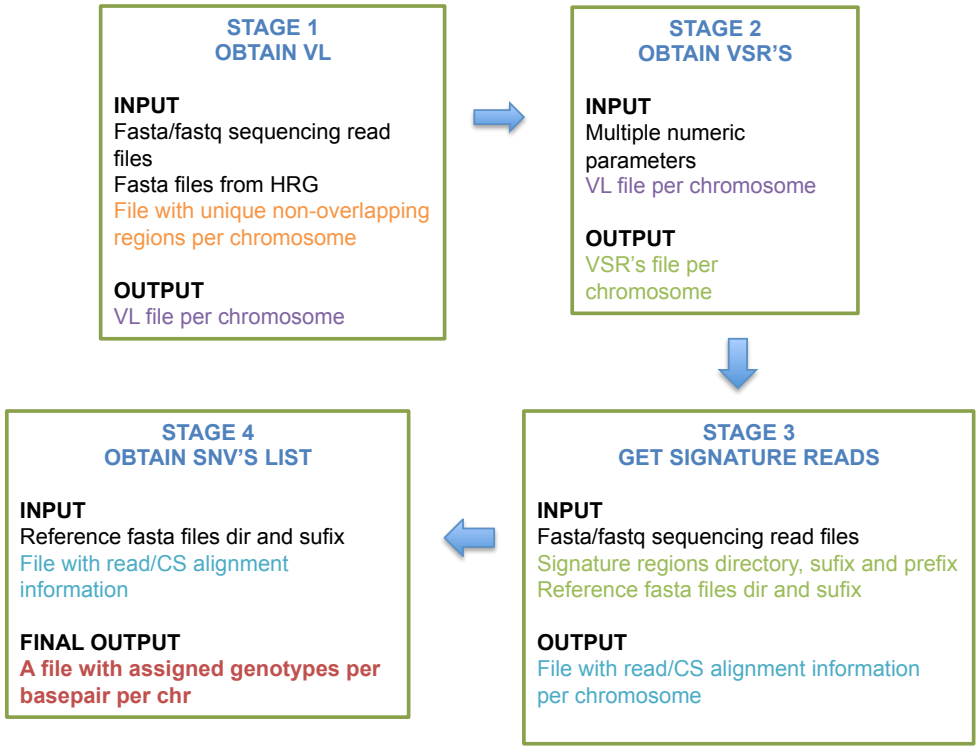


FIG 2. The whole one individual framework is shown.

Family-based framework

The family–based frameworw that is described in this manual is an extension of the one-individudal framework VGH approach. For the child all the processes from the one-individual framework must be completed. In the case of the parents only a list of VSR's must be generated (Stop at Stage 3 from One Individual Framework). It is important to note that different RCI thresholds must be set for each individual (See Stage 3- Obtain VSR's list, One-individual framework). The family-based framework uses as input such data.

In the Stage 1 the Signature Reads are obtained for each parent, in the Stage 2 a SNV is be generated for each parent. The processes described in Stage 1 and 2 must be completed for each parent independently. In the Stage 3, the data for the three individual of the family trio is analyzed and the *de novo* SNV are obtained. The whole pipeline is illustrated in Figure2.

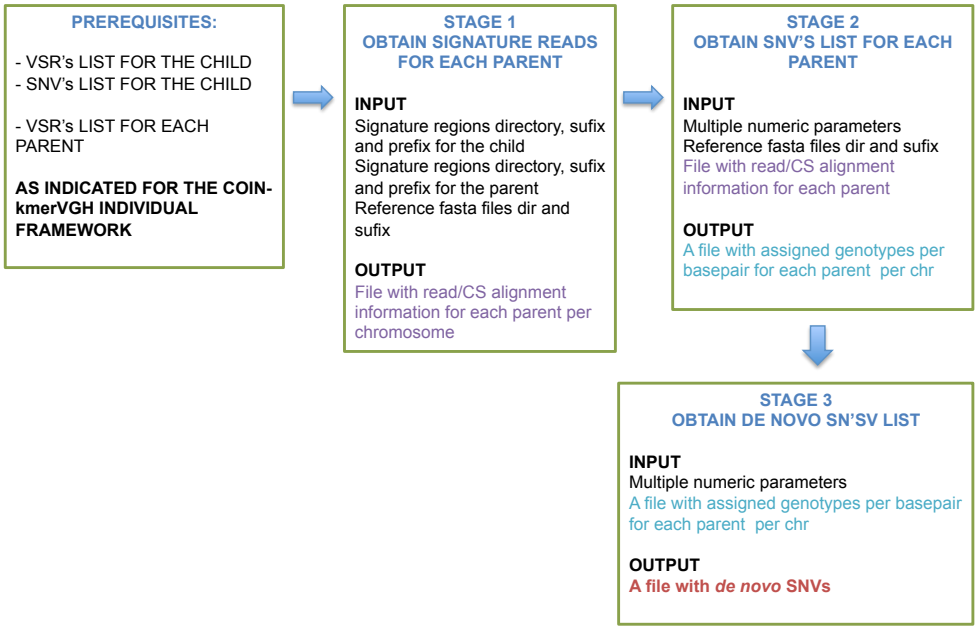


FIG 3. The family-based framework is illustrated.

User Guide
COBASI Pipeline
Universidad Nacional Autónoma de México

PREREQUISITES

Software required. The pipeline was developed using the versions specified for each software.

- Bowtie [version supported 1.1]
- Jellyfish [version supported 1.1.6]
- AMOS [version supported 3.1.0]

  BOOST, Jellyfish and qt4 are required to install AMOS

  Execute these command lines for AMOS installaton:

  git clone git://amos.git.sourceforge.net/gitroot/amos/amos
  ./bootstrap
  ./configure --with-Boost-dir=/bin/BOOST/ --with-jellyfish=/bin/jellyfish/ --with-qmake-qt4=/bin/qt4/bin/qmake --prefix /bin/AMOS/
  make
  make install
- Python [version supported 2.7.2]

  Biopythhon is required
- Perl [version supported v5.14.2]

  Module Switch.pm is required
- c++ [version supported 4.4.6]

Hardware recommended for the analysis of human WGS (30-40X)

- 128GB RAM
- 12 cores

REQUIRED FILES

- Bowtie RG index
- Fasta files from RG
- Fasta/fastq sequencing reads files

Centro de Ciencias Genómicas

LIIGH-UNAM

User Guide
COBASI Pipeline
Universidad Nacional Autónoma de México

USE DESCRIPTION: GET CS REGIONS

This process is required to be run only once for each RG. In this stage a Reference Genome (RG) COIN-String (CS) Regions database is generated. This is composed of three sequential steps.

FIG 4. The three sequential steps required in this stage are listed.

CUT REFERENCE GENOME

This script cuts one chromosome from the reference genome getting all kmers by sliding-windows of one base pair .

COMMAND LINE.

perl Cut_RG.pl -fna FASTA -k K –out OUT.STR

PARAMETERS.

- fna A fasta file with the sequence of one RG chromosome
- k kmer size (set to K=30)
- out Output file

OUTPUT.

A multi-fasta file with the sequence for every kmer along each chromosome of the RG The ID for every multi-fasta sequence will correspond to the start position of each kmer

NOTES.

This process must be run per each chromosome to be analyzed
IMPORTANT: The ID in the fasta file should correpond to the file name:
chr1.fasta:
>chr1
NNNNNNNNNNNNNN

OBTAIN UNIQUE KMERS

This process gets all kmers from a chromosome found only once in the whole RG

COMMAND LINE.

bowtie -v V -m M –f INDEX FASTA OUTPUT &

PARAMETERS.

- v — Alignments (end-to-end) may have no more than V mismatches (SET to V=0)
- m — Suppress all alignments for a particular read or pair if more than <int> reportable alignments exist for it (SET to M=1)
- f — Query input files are (multi-)FASTA
- index — RG bowtie index (ebwt files path and prefix)
- fasta — Multi-fasta file with kmers from RG
- output — Output file [recommended extension .bowtie.out]

OUTPUT.

A bowtie output file per chromosome.
To see file explanation see Bowtie documentation
[http://bowtie-bio.sourceforge.net/index.shtml]

NOTES.

This process must be run per each chromosome.

OBTAIN NON-OVERLAPPING UNIQUE REGIONS

This script merges all positions from adjacent unique kmers (found by bowtie). It generates a list of unique non-overlapping regions.

COMMAND LINE.

perl Compute_CSDB_FromBowtie.pl –dir DIR_BOWTIE/ -sufix .BOWTIE.OUT –out DIR_OUT/

PARAMETERS.

- dir — Directory that contains the output from bowtie
- sufix — Bowtie output files sufix
- out — Output directory

OUTPUT.
One file per chr with a list of start and end positions of unique non-overlapping regions.
The ouput files will have the extension "_cs_uniq_regions.tab"

NOTES.
This script should be run only once

Centro de Ciencias Genómicas

LIIGH-UNAM

User Guide
COBASI Pipeline
Universidad Nacional Autónoma de México

USE DESCRIPTION:
ONE INDIVIDUAL FRAMEWORK.

STAGE 1 – OBTAIN VL

In Stage 1 the Variation Landscape is computed (VL). This stage is composed of four sequential steps. However, process 2 (MERGE KMER COUNT) is only required if multiple databases are generated in process 1.

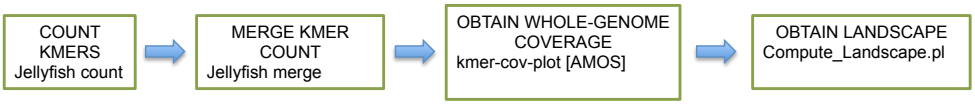


FIG 5. The processes from Stage 1 are listed.

COUNT KMERS

Jellyfish is a parallel-processing software that counts the ocurrences of all kmers (of k size) in a list of fasta or fastq files. Read the parameters used to understand how.

COMMAND LINE.

jellyfish count –o DB_OUT -m K -s S –t T --both-strands -L L [fastq1 fastq2]

PARAMETERS.

- Output — Output prefix
- m — Kmer size (SET to K=30)
- s — Hash size (RECOMMENDED S=10G)
- t — Number of threads (RECOMMENDED T=24 cores)
- both-strands — For any k-mer m, its canonical representation is m itself or its reverse-complement, whichever comes first lexicographically
- L — Don't output k-mer with count lower than L (SET to L=2)
- [files] — List of fasta or fastq files (sequencing reads files, separated by spaces)

OUTPUT.

One or more jellyfish databases storing the kmer counts for the sequencing reads.

NOTES.

Include all the fasta or fastq files from a single sequencing project in the same jellyfish command. For more documentation see:
http://www.cbcb.umd.edu/software/jellyfish/

MERGE KMER COUNT [OPTIONAL]

This command is required only if several jellyfish databases are generated in the previous step. It merges the count for every kmer in any of the listed databases.

COMMAND LINE.
jellyfish merge –o DB_MERGED [db_out_0 db_out_1]

PARAMETERS.
- o Output database
- databases List of jellyfish databases that will be merged

OUTPUT.
An unique merged jellyfish database per sequencing project

NOTES.
For more documentation see:
http://www.cbcb.umd.edu/software/jellyfish/

OBTAIN WHOLE-GENOME COVERAGE

kmer-cov-plot is a software included in the AMOS package. It returns the count (found in a kmer-count jellyfish database) for every kmer along a provided reference genome. It outputs the start position for every kmer with its count. This kmer count is taken as the read coverage for every kmer.

COMMAND LINE.
kmer-cov-plot --jellyfish –s DB < FASTA > OUTPUT.COVERAGE

PARAMETERS.
- jellyfish Use k-mer counts from a Jellyfish hash table.
- s Display only the combined count of the forward and reverse complement k-mers
- db Jellyfish database
- fasta RG fasta file
- output Output file

OUTPUT.
A file with the start position for every kmer along the fasta file sequence (RG chromosome) in column 1 and its count (retrieved from the jellyfish database) in column 2 (coverage files)

NOTES.
This process must be run per each chromosome.
For more documentation see:
[http://sourceforge.net/projects/amos/]

OBTAIN LANDSCAPE

This script filters out the positions for every non-unique kmers. It outputs a list of start positions for every unique kmer (CSs) asociated wit its count.

COMMAND LINE.
perl Compute_Landscape.pl -cov OUTPUT.COVERAGE -unique DIR_CSs/ -suf_unique SUFIX -out_dir DIR_OUT /

PARAMETERS.

- cov — Coverage file
- unique — The directory containing the files with the start and end positions of the non-overlapping unique region
- suf_unqiue — Sufix of the unique regions files (SET to SUFIX =_cs_uniq_regions.tab)
- out_dir — Output directory. The output files will be named {RG chromosome}{.land}

OUTPUT.
The Variation Landscape (VL) file per chromosome. It contains how many times each CS is found in the sequencing reads. The VL file contains two columns:

- column 1, RG position
- column2, the number of ocurrences per each CS along the reads.

NOTES.
This process must be run per each chromosome.

STAGE 2 – OBTAIN VSR's

In Stage 2 the Variation Signature Regions are identified (VSR's). Remember that every VSR can be composed of at most two drops and two rises at the start or at the end of the VSR, respectively. This stage is composed of two sequential steps.

NOTE1: The RCI threshold used for the VSR's discovery should be conservative (recommended 0.3)
NOTE2: To choose the different coverage thresholds an additional script has been developed. See Additional Scripts.

FIG 6. The processes from Stage 2 are listed.

GET ONE-DROP REGIONS (INTERNAL PART OF VSR'S)

This script obtain a list of partial VSR's. It only looks for the internal drop and rise in coverage characteristics of any VSR.

COMMAND LINE.

perl Get_Onedrop.pl -land FILE.LAND -max M -fac RCI -rmin R -fst FST -nt NT –density DEN –out FILE.ONEDROP

PARAMETERS.

- land — Input file. VL file [FILE.LAND]
- max — CS's with a coverage count higher than M are skipped in the VSR identification process
- fac — A Relative Coverage Index absolute value higher than RCI is considered as *bona fide* variation signal (SET RCI=0.3)
- rmin — The PrevCS and PostCS coverage must be at least R (SET R=10)
- fst — The median of the coverage values for InterCS's must be lower than FST (SET FST=third quartile coverage+10(IQR))
- nt — The length of the VSR must be longer than NT
- density — DEN is the mínimum density of CSs required inside the VSR
- out — Output file [FILE.ONEDROP]

OUTPUT.
The output file contains either whole VSRs or partial VSRs (only the internal drop and rise). In this step, the PrevCS and PostCS will be the ones before and after this (possibly incomplete) VSR signal. These regions will be extended, if possible, in the next process. This output file is composed of 13 columns:

1. PrevCS position
2. PrevCS count
3. Next CS after PrevCS position
4. Next CS after PrevCS count
5. Number of nucleotides in the internal region of the VSR
6. Number of interCSs (a VSR with a CS density of 1 will have the same value in columns 5 and 6)
7. Last CS before PostCS position
8. Last CS before PostCS count
9. PostCS position
10. PostCS count
11. RCI value for the VSR start
12. RCI value for the VSR end
13. Median of the coverage values for the InterCS's

NOTES.
This process must be run per each chromosome.

GET VARIATION SIGNATURE REGIONS

In the previous step possibly incomplete VSRs are detected. In this step, additional drops will be searched at the start of the previously identitfied partial VSRs and adittional rises will be looked at the end of such VSRs. Finally adjacent partial VSR's will be concatenated. Intermediate rises between the drops or intermediate drops between the rises are not allowed.

COMMAND LINE.
perl Get_SR.pl -land FILE.LAND –var FILE.ONEDROP -fac RCI -rmin R -max M -cs_size K –ratio RAT –sr_max SRM –out FILE.SR

PARAMETERS.

- land — Input file. VL file [FILA.LAND]
- var — File obtained in the Get One-Drop Regions process [FILE.ONEDROP]
- fac — A Relative Coverage Index absolute value higher than RCI is considered as *bona fide* variation signal (set to RCI=0.3)

- rmin — The PrevCS and PostCS coverage must be at least R (set to R=10)
- max — CS's with a count higher than M are skipped in the VSR process
- cs_size — An additional drop or rise are searched in the K nucleotides before and after the partial VSR that has been detected in the var file
- ratio — The ratio of the coverage between the PrevCS and PostCS should be < than RAT (recommended set RAT= 1.8)
- sr_max — The distance between PrevCS and PostCS should be < than SRM (recommenden set SRM=100000)
- out — Output file with final VSR information [FILE.SR]

OUTPUT.

The output file contains final VSRs information (VSR files). This output file is composed of 14 columns:

1. PrevCS position
2. PrevCS count
3. Next CS after PrevCS position
4. Next CS after PrevCS count
5. Number of nucleotides in the internal region of the VSR
6. Number of inter CSs (a VSR with a CS density of 1 will have the same value in columns 5 and 6)
7. Last CS before PostCS position
8. Last CS before PostCS count
9. PostCS position
10. PostCS count
11. RCI value for the VSR start
12. RCI value for the VSR end
13. Median of the coverage values for InterCS's
14. Flag indicating in which end a ladder is found (UP, DOWN, BOTH, NA)

NOTES.
This process must be run per each chromosome.

STAGE 3 - GET SIGNATURE READS

In Stage 3 the Signature Reads that contain either the PreCS or both SignatureCSs are retrieved. This stage is composed of three sequential steps.

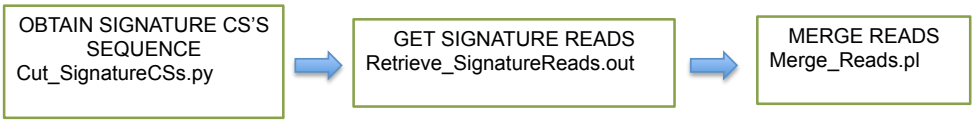


FIG 7. The processes from Stage 3 are listed.

OBTAIN SIGNATURE CS'S SEQUENCE

This script will obtain the sequences for the PreCS and PostCS for every VSR. All the VSR file must be located on the same directory and they must have the same extension. The header of the multi-FASTA file that is generated will change for the CHILD and for the PARENTS in the family-based framework

COMMAND LINE.

python Cut_SignatureCSs.py --VSR=SR/ --REFDIR=REFDIR/ --sufixREF=sufixREF --sufixVSR=sufixVSR --prefixVSR=prefixVSR --kSIZE=kmer_size --output=SignatureCS.fa

python Cut_SignatureCSs.py -v SR/ -r REFDIR/ -x sufixREF -y sufixVSR -z prefixVSR -k kmer_size -o SignatureCS.fa

PARAMETERS.

- Script — Cut_SignatureCSs.py
- VSR, v — VSR directory path
- REFDIR , r — RG directory path
- sufixREF, x — Sufix of the fasta reference files
- sufixVSR , y — Sufix of VSR files
- prefixVSR , z — Prefix of VSR files (if the VSR files have no prefix, set to NA)
- kSIZE, k — CS size (SET to 30)
- output, o — Output file (multi-fasta file)

The name of the files must follow the next rules. Example:
RG chr1 file name: chr1.{sufixREF}
VSR chr1 file name: {prefixSR}chr1{sufixSR}

OUTPUT.
A multi-fasta file with the sequences of all Signature CSs

NOTES.
This step requires only one process

GET SIGNATURE READS

This script will obtain the sequences and some useful information from the reads that contain either the PreCS or both SignatureCSs. This script will process one FASTA or FASTQ file (with the read sequences) per run. The header of the input multi-FASTA is slightly different for the CHILD and for the PARENTS in the family-based framework

COMMAND LINE.

Retrieve_SignatureReads -c SignatureCS –t ReadType –f ReadFileX –k kmer_size –i Ind > outX

PARAMETERS.

- c [SignatureCS] A multi-fasta file with the sequences of all Signature CSs
- t [ReadType] FASTA or FASTQ formats are supported [FASTA|FASTQ]
- f [ReadFileX] A fasta or fastq file with the reads of the sequencing project
- k [kmer_size] CSs size = Jellyfish database kmer size
- i [Ind] Set to CHILD for SNV discovery in one individual
- outX Output file

OUTPUT.

The ouput file contains information about the alignment between the read and the respective Signature CS's (READALN-UNORDER files). This output file is composed of 10 columns:

1. RG chromosome
2. VSR PrevCS start position
3. VSR PostCS start position
4. SignatureCS aligned to the read (either the PrevCS or the PostCS)
5. CS sequence
6. Downstream read position of the alignment
7. Orientation of the aligment
8. Read ID
9. Read Sequence
10. Read Quality (NA, if the initial format file is FASTA)

NOTES.

This process must be run per each read file. Each process will produce one output file.

MERGE READS

If the sequencing experiments generates multiple FASTQ read files, the script that gets the Signature Reads will be run multiple times, and the hits for every SignatureCS will be distributed over multiple files. This script will concatenate such information.

COMMAND LINE.

perl Merge_Reads.pl -dir_read ALNDIR/ -sufix_read sufixREAD -dir_var SR/ -prefix_var prefixSR -sufix_var sufixSR -ind CHILD -dir_out OUTDIR/ -prefix_out GenomePrefix -sufix_out sufixREADALN &

PARAMETERS.

- dir_read — The directory where the READALN-UNORDER files were written
- sufix_read — Sufix of the READALN-UNORDER files
- dir_var — The directory where the VSR files were written.
- prefix_var — VSR file name must be: {prefixSR}chrName{sufixSR}. If there is no prefix set to NA
- sufix_var — Sufix of the VSR files. All the files from SR/ ending in sufixSR will be analyzed.
- ind — Set to CHILD for one individual SNV discovery
- dir_out — Output directory.
- prefix_out — Prefix for the output files. Set to NA if no prefix is desired.
- sufix_out — Sufix for the output files [READALN files]

OUTPUT.

One output file per chromosome.
Each output file contains ordered information about the alignment between the read and the respective Signature CS's (READALN files).
The information contained in these files is the same unordered information contained in the READALN-UNORDER files.

NOTES.

VSR file name must be: {prefixSR}chrName{sufixSR}.
This step requires only one process.
One output file is produced per each chromosome, the chromsome names are indicated in the names of the VSR files.
Output file name: {VSR file name}{ .reads}

STAGE 4 – OBTAIN SNV LIST

In Stage 4 a list of Single Nucleotide Variants (SNV's) is generated from information recorded from the Signature CSs.

FIG 8. The processes from Stage 4 are listed.

ALIGN READ-REFERENCE GENOME

In this step every read is cut based on the positions recorded in the READALN files. This partial sequence read is aligned to the corresponding region in the RG. There are two different kind of reads, 1) the ones containing only the PrevCS from which only partial alignments of the VSR can be generated; and 2) the ones containing both Signature CSs from which global alignments (total aln) of the VSR can be generated . In this script reads with incongruency in the alignments with the Signature CS's are discarded, the PCR duplicates are eliminated, the total alignments are computed and variable regions are obtained. These variable regions must be in more than ReadRegion number of total alignments to be considered, only one allele per region is allowed, low complexity alignments from which multiple alignment positions can be obtained are discarded. . From these total alignments the minimum length of partial alignments is obtained. Partial alignments are computed. Alleles per base (reference

or alternative) are obtained, no gaps are allowed and only alleles found in ReadPerbase alignments are recorded.

COMMAND LINE.

Align_read_RG –a READALN-file –r REFDIR/ -s sufixREF –k k –t ReadRegion –p ReadPerbase –l LengthPartial –c ChildSNV –d RemoveDupFlag –i Ind –o out.total –q out.perbase

PARAMETERS.

- a [READALN-FILE] File per chromosome that contains information about the alignment between the read and the respective Signature CS's
- r [REFDIR] RG directory path
- s [sufixREF] Sufix of the fasta reference files
- k [k] CSs size = Jellyfish database kmer size
- t [ReadRegion] Each mismatch region between the reads and the RG must be supported by at least ReadRegion different reads
- p [ReadPerbase] Each mismatch nucleotide between the reads and the RG must be supported by at least ReadPerbase different reads
- l [LengthPartial] Length of extensión of partial alignment (set to 10)
- c [ChildSNV] Set to NA (Parameter used in family-framework)
- d [RemoveDupFlag] Set to TRUE to remove PCR duplicates (FALSE otherwise)
- i [Ind] For one individual SNV discovery set to CHILD
- o [out.total] Output file with regions of polymorphism data
- q [out.perbase] Output file with single nucleotide polymorphism data

OUTPUT.
A file with genotypes per region per chromosome was obtained with the following information:

1. RG chromosome
2. PrevCS start position
3. PostCS start position
4. Polymorphism region RG start
5. Polymorphism region RG end
6. RG nucleotide
7. Reads variant allele
8. Total alignments supporting variant allele/Number of total alignments

A file with genotypes per basepair per chromosome was obtained with the following information:

1. RG chromosome
2. PrevCS start position
3. PostCS start position
4. SNV RG position
5. RG nucleotide
6. Read alleles (allele1/allele2)
7. Total alignments supporting every allele (allele1/allele2/total)
8. Partial alignments supporting every allele (allele1/allele2/total)
9. Total number of alignments supporting every allele (allele1/allele2/total)

NOTES.
Nomenclature: RG chr1 file name: chr1{sufixREF}.
One process must be run per chromosome.

To detect the low complexity alignments: start and end positions for regions that contain consecutive polymorphisms were obtained, if multiple regions spanning the same nucleotide intervals were obtained, all the alignments for those low complexity regions were discarded.

To detect PCR duplicates: multiple reads for which the PrevCS was aligned to the same position were considered as PCR duplicates and one of them (randomly chosen) was assigned to be the representative read.

COMPUTE SNVs GENOTYPE

In this script, the likelihood for each genotype is computed and the most probable genotype (given the observed data) is reported as the final genotype. The probability for four possible genotypes is computed: Heterozygous reference (R/NR), Homozygous

reference (R/R), Heterozygous non-reference (NR1/NR2), Homozygous non-reference (NR/NR). For the child, this script does not print the homozygous reference-sites.

COMMAND LINE.

Compute_Genotype–p out.perbase –t FilePerbaseType –i IndividualType –u undetermined.out > genotype.out

PARAMETERS.

- p [out.perbase] File with single nucleotide polymorphism data
- t [FilePerbaseType] P if total number of alignments per allele is in column 9
- i [IndividualType] For individual SNV discovery set to CHILD
- u [undetermined.out] Output file with undetermined genotypes
- genotype.out File with genotype information per SNV

OUTPUT.

A file with assigned genotypes per basepair per chromosome with 11 columns:

- 1-9 will contain the information from the file out.perbase
- A numeric classificarion for the most probable genotype:
  - 1 - Heterozygous reference (Ref/NoRef)
  - 2 - Homozygous reference (Ref/Ref)
  - 3 - Heterozygous non-reference (NoRef1/NoRef2)
  - 5 - Homozygous non-reference (NoRef1/NoRef1)
- The different alleles for the assigned genotype

NOTES.

One process must be run per chromosome.

USE DESCRIPTION
FAMILY-BASED FRAMEWORK

STAGE 1 – OBTAIN SIGNATURE READS

In Stage 1 the Signature Reads for each parent are retrieved. This stage must be completed for each parent independently. In this stage only the regions that contain a SNV in the child are analyzed in the parents. For every SNV and its associated VSR in the child, the reads that contain the child Signature CS's for the associated VSR are retrieved from the parent fastq files.

FIG 9. The processes from Stage 1 are listed.

OBTAIN SIGNATURE CS'S SEQUENCE

This script will copy the file that contains the SignatureCSs identified in the child to the working directory for the analysis of every parent. The identifier of every Signature CSs will now reflect that is being used in the parent SNV discovery

COMMAND LINE.

python SignatureCS_Parent.py -c SignatureCSs_child.fa -o SignatureCSs_parent.fa

python SignatureCS_Parent.py --SignatureCSs SignatureCSs_child.fa
--OUTPUT SignatureCSs_parent.fa

PARAMETERS.

- Script — SignatureCS_Parent.py
- SignatureCSs, c — Multi-fasta file with all Signature CSs for the child
- OUTPUT, o — Output file (multi-fasta file), the same Child SignatureCS with IDs for the parents

OUTPUT.

A multi-fasta file with the sequences of all retrieved CSs sequence.
The identifier if every fasta sequence contains:

- CS ID for the sequence retrieved
- PrevCS for the child VSR
- Post CS for the child VSR

The last two positions will ve repeated twice

NOTES.
This step requires only one process

GET SIGNATURE READS

This script will obtain the sequences and some useful information from the reads that contain either the PreCS or both SignatureCSs. This script will process one FASTA or FASTQ file (with the read sequences) per run. The header of the input multi-FASTA is slightly different for the CHILD and for the PARENTS in the family-based framework. The software used at this stage is the same software used in One Individual – Stage3.

COMMAND LINE.

Retrieve_SignatureReads -c SignatureCS –t ReadType –f ReadFileX –k kmer_size –i Ind > outX

PARAMETERS.

- c [SignatureCS] A multi-fasta file with the sequences of all Signature CSs
- t [ReadType] FASTA or FASTQ formats are supported [FASTA|FASTQ]
- f [ReadFileX] A fasta or fastq file with the reads of the sequencing project
- k [kmer_size] CSs size = Jellyfish database kmer size
- i [Ind] Set to PARENT for SNV discovery in the parents
- outX Output file

OUTPUT.
The ouput file contains information about the alignment between the read and the respective Signature CS's (READALN-UNORDER files). This output file is composed of 12 columns:

1. RG chromosome
2. SignatureCS aligned to the read (either the PrevCS or the PostCS)
3. Parent VSR PrevCS start position
4. Parent VSR PostCS start position
5. Child VSR PrevCS start position
6. Child VSR PostCS start position
7. CS sequence
8. Downstream read position of the alignment
9. Orientation of the aligment

10. Read ID
11. Read Sequence
12. Read Quality (NA, if the initial format file is FASTA)

NOTES.
This process must be run per each read file
One output file is produced per each read file

MERGE READS

If the sequencing experiments generates multiple FASTQ read files, the script that gets the Signature Reads will be run multiple times, and the hits for every SignatureCS will be distributed over multiple files. This script will concatenate such information. The software used at this stage is the same software used in One Individual – Stage3.

COMMAND LINE.

perl Merge_Reads.pl -dir_read ALNDIR/ -sufix_read sufixREAD -dir_var SR/ -prefix_var prefixSR -sufix_var sufixSR -ind PARENT -dir_out OUTDIR/ -prefix_out GenomePrefix -sufix_out sufixREADALN &

PARAMETERS.

- dir_read — The directory where the READALN-UNORDER files were written
- sufix_read — Sufix of the READALN-UNORDER files
- dir_var — The directory where the VSR files were written.
- prefix_var — VSR file name must be: {prefixSR}chrName{sufixSR}. If there is no prefix set to NA
- sufix_var — Sufix of the VSR files. All the files from SR/ ending in sufixSR will be analyzed.
- ind — Set to PARENT for SNV discovery in the parents
- dir_out — Output directory.
- prefix_out — Prefix for the output files. Set to NA if no prefix is desired.
- sufix_out — Sufix for the output files [READALN files]

OUTPUT.
One output file per chromosome.
Each output file contains ordered information about the alignment between the read and the respective Signature CS's (READALN files).
The information contained in these files is the same unordered information contained in the READALN-UNORDER files.

NOTES.
VSR file name must be: {prefixSR}chrName{sufixSR}.
This step requires only one process.

One output file is produced per each chromosome, the chromsome names are indicated in the names of the VSR files.
Output file name: {VSR file name}{ .reads}

STAGE 2 – OBTAIN SNV LIST

In Stage 2 a list of Single Nucleotide Variants (SNV's) is generated for each parent. This Stage is almost identical to Stage 5 in the One-Indivual framework, the only different parameter is the information contained in the input files.

FIG 10. The processes from Stage 2 are listed.

ALIGN READ-REFERENCE GENOME

In this step every read is cut based on the positions recorded in the READALN files. This partial sequence read is aligned to the corresponding region in the RG. There are two different kind of reads, 1) the ones containing only the PrevCS from which only partial alignments of the VSR can be generated; and 2) the ones containing both Signature CSs from which global alignments (total aln) of the VSR can be generated . In this script reads with incongruency in the alignments with the Signature CS's are discarded, the PCR duplicates are eliminated**. Only the alignments which contain regions that has been categorized as variable in the child are analyzed**, no gaps are allowed and only alleles found in ReadPerbase alignments are recorded. The software used at this stage is the same software used in One Individual – Stage4.

COMMAND LINE.

Align_read_RG –a READALN-file –r REFDIR/ -s sufixREF -k k –t ReadRegion -p ReadPerbase –l LengthPartial -c ChildSNV –d RemoveDupFlag –i Ind –o out.total -q out.perbase

PARAMETERS.

- a [READALN-FILE] File per chromosome that contains information about the alignment between the read and the respective Signature CS's
- r [REFDIR] RG directory path
- a [sufixREF] Sufix of the fasta reference files
- k [k] CSs size = Jellyfish database kmer size
- t [ReadRegion] Parameter used in One Individual SNV discovery(set to 1)
- p [ReadPerbase] Each mismatch nucleotide between the reads and the RG must be supported by at least ReadPerbase different reads
- l [LengthPartial] Length of extensión of partial alignment (set to 10)
- c [ChildSNV] Path to variable regions file for child individual (out.total)
- d [RemoveDupFlag] Set to TRUE to remove PCR duplicates (FALSE otherwise)
- i [Ind] Set to PARENT for SNV discovery in the parents
- o [out.total] Set to NA for SNV discovery in the parents
- q [out.perbase] Output file with single nucleotide polymorphism data

OUTPUT.

A file with genotypes per basepair per chromosome was obtained with the following information:

1. RG chromosome

2. PrevCS start position
3. PostCS start position
4. SNV RG position
5. RG nucleotide
6. Read alleles (allele1/allele2)
7. Total alignments supporting every allele (allele1/allele2/total)
8. Partial alignments supporting every allele (allele1/allele2/total)
9. Total number of alignments supporting every allele (allele1/allele2/total)

NOTES.
Nomenclature: RG chr1 file name: chr1{sufixREF}.
One process must be run per chromosome.

To detect the low complexity alignments: start and end positions for regions that contain consecutive polymorphisms were obtained, if multiple regions spanning the same nucleotide intervals were obtained, all the alignments for those low complexity regions were discarded.

To detect PCR duplicates: multiple reads for which the PrevCS was aligned to the same position were considered as PCR duplicates and one of them (randomly chosen) was assigned to be the representative read.

GET SNVs GENOTYPE

In this script, the likelihood for each genotype is computed and the most probable genotype (given the observed data) is reported as the final genotype. The probability for four possible genotypes is computed: Heterozygous reference (R/NR), Homozygous reference (R/R), Heterozygous non-reference (NR1/NR2), Homozygous non-reference (NR/NR). For the child, this script does not print the homozygous reference-sites. The software used at this stage is the same software used in One Individual – Stage4.

COMMAND LINE.

Compute_Genotype –p out.perbase –t FilePerbaseType –i IndividualType
-u undetermined.out > genotype.out

PARAMETERS.

- p [out.perbase] File with single nucleotide polymorphism data
- t [FilePerbaseType] P if total number of aln per allele is in column 9
- i [IndividualType] Set to PARENT for SNV discovery in the parents
- u [undetermined.out] Output file with undetermined genotypes
- genotype.out File with genotype information per SNV

OUTPUT.
A file with assigned genotypes per basepair per chromosome with 11 columns:

- 1-9 will contain the information from the file out.perbase
- A numeric classificarion for the most probable genotype:
  - 1 - Heterozygous reference (Ref/NoRef)
  - 2 - Homozygous reference (Ref/Ref)
  - 3 - Heterozygous non-reference (NoRef1/NoRef2)
  - 5 - Homozygous non-reference (NoRef1/NoRef1)
- The different alleles for the assigned genotype

NOTES.
One process must be run per chromosome.

STAGE 3 – OBTAIN *DE NOVO* SNV LIST

In this Stage the SNV data for all individuals from the TRIO are analyzed jointly. In the first process the inheritance mode for each SNV in the child is obtained (either congruent with mendelian inheritance or incongruent). To interrogate any position, genotypes must be succesfuly assigned for any individual and total alignments should exist in the region. In the second process a minimum coverage is required for the variant region and for the variant allele, besides the difference in coverage between the different alleles must be lower than a required threshold

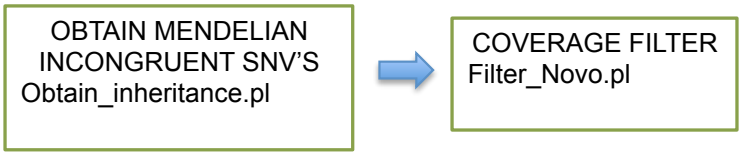


FIG11. The processes from Stage 3 are listed.

OBTAIN MENDELIAN INCONGRUENT SNV'S

In this scrip the inheritance mode for each SNV in the child is obtained (either congruent with mendelian inheritance or incongruent). To interrogate any genomic position, genotypes must be succesfuly assigned for any individual and total alignments should exist in the region.

COMMAND LINE.

User Guide
COBASI Pipeline
Universidad Nacional Autónoma de México

Obtain_Inheritance.pl –hg3 childGenotype –hg1 fatherGenotype -hg2 motherGenotype –chr chr –out OUTDIR

PARAMETERS.

- dir_hg3 File with genotype information for the child
- dir_hg1 File with genotype information for the father
- dir_hg2 File with genotype information for the mother
- chr Chromosome
- out Output diectory

OUTPUT.

- A file with all SNVs with their inheritance mode: either mendelian congruent or incongruent. The information contained in this file is the same information of the genotype file for the child. In addition the last column contained the inheritance mode [CONGRUENT | INCONGRUENT]
- A SUMMARY file with mendelian congruent genotypes per chromosome [mend.congruent]
- A SUMMARY file with mendelian incongruent genotypes per chromosome [mend.incongruent]
- Error files with all genomic positions that:
  - Are not present in either the father or the mother SNV list
  - Have failled to succesfully assign a genotype in any individual
  - Don't have total alignments in any individual

The SUMMARY files contain:

1. RG chromosome
2. Child PrevCS start position
3. Child PostCS start position
4. SNV RG position
5. RG nucleotide
6. Read alleles for the child (allele1/allele2)
7. Total alignments supporting every allele for the child (allele1/allele2/total)
8. Partial alignments supporting every allele for the child (allele1/allele2/total)
9. Total number of alignments supporting every allele for the child (allele1/allele2/total)
10. A numeric classification for the most probable genotype for the child:
    a. 1 - Heterozygous reference (Ref/NoRef)
    b. 2 - Homozygous reference (Ref/Ref)
    c. 3 - Heterozygous non-reference (NoRef1/NoRef2)
    d. 5 - Homozygous non-reference (NoRef1/NoRef1)
11. The different alleles for the assigned genotype for the child
12. Read alleles for the father
13. Total alignments supporting every allele for the father
14. Partial alignments supporting every allele for the father
15. Total number of alignments supporting every allele for the father

16. A numeric classification for the most probable genotype for the father.
17. The different alleles for the assigned genotype for the father
18. Read alleles for the mother
19. Total alignments supporting every allele for the mother
20. Partial alignments supporting every allele for the mother
21. Total number of alignments supporting every allele for the mother
22. A numeric classification for the most probable genotype for the mother.
23. The different alleles for the assigned genotype for the mother.

These files will be written to:

- OUTDIR /{chr}.mend
- OUTDIR /{chr}.mend.congruent
- OUTDIR /{chr}.mend.incongruent

NOTES.
This step require one process per chromsome
If there is none total alignments in at least one of the individual, that event is nos reported in subsequent output files.

COVERAGE AND PURITY FILTER

Several characteristics are required for a mendelian incongruent SNV to be classified as *de novo*. 1) A minimum coverage is required for the variant region and for the variant allele for every individual; 2) The difference in coverage between the different alleles(in the child) must be lower than a required threshold; 3)No parent should habe both child alleles contained in more than one total alignment.

COMMAND LINE.

Filter_Novo.pl -dir child-GENOTYPE/ -sufix mend.incongruent –cov_child 5 –total_child 5 –cov_parent 5 –total_parent –min MIN -stat Stat.out > result.out

PARAMETERS.

- dir — Directory with mendelian incongruent genotype
- sufix — Sufix for the mendelian incongruent genotype files
- cov_child — Minimum number of reads [either partial or total] in the child
- total_child — Minimum number of total alignments in the child
- cov_parent — Minimum number of reads [either partial or total] in either parent
- total_parent — Minimum number of total alignments in either parent
- min — Events with a ratio higher than MIN between the read counts for the different alleles are filtered out
- stat — Output file with statistics of filtered SNVs

- result.out Output with the SNVs that PASSED the filtering criteria

OUTPUT.
The same information as the SUMMARY file with mendelian incongruent genotypes per chromosome [mend.incongruent]

NOTES
In the VGH pipeline, all the SNVs found in dbSNP are not considered as real de novo SNVs

ADDITIONAL SCRIPTS

COVERAGE STATISTICS

COMMAND LINE.
perl Calculate_Coverage_Statistics.pl –dir LANDDIR –pat sufixLAND -out out.stat

PARAMETERS
- dir VL directory path
- pat VL files sufix
- out Output file

OUTPUT
A file with coverage statistics:
- Total number of CS's analyzed.
- The coverage mean
- The coverage standard deviation
- The maximum coverage
- The first quartile, the median, the third quartile and the IQR of the coverage

NOTES
Some other debug values are printed

POST-PROCESSING

Once the candidate de novo SNVs have been identified, this script will analyze the regions corresponding to the child VSR for the three family individuals to identify undesired patterns. 1) regions with low CS density; 2) regions in which any CS has a

coverage higher than expected; 3) for any individual regions with low coverage for the CSs corresponding to the child Signature CSs, 4) regions with additional peaks inside the region corresponding to the child VSR : in the case of the child if there is any additional drop or rise it should correspond to a region with almost no coverage; in the case of the parents there should not exist any drop or rise that indicates a posible heterozigosity for the child SNV position or there should not exist a drop and rise that correspond to the exact same child's VSR boundaries; and 5) for the child, regions with unequal coverage in both sides of the VSR

COMMAND LINE.

perl Postprocessing.pl -novo Novo.out –land LAND-DIR/ -chr CHR –parent P -density DEN –max MAX -higher HIGH -cov COV -rci RCI –rmin RMIN -low LOW -peaks PEAK -out Novo.chr.genome.tp

PARAMETERS

- novo — File with de candidate *de novo* SNVs
- land — Directory with the VLs (Variation landscapes)
- chr — The chromosome to be analyzed [CHR]
- parent — If the landscape corresponds to the child set [P] to 0
If the landscape corresponds to the parent set [P] to 1
- density — Regions with a density of at least [DEN] will be kept
- max/higher — Regions with at most [HIGH] CSs with a coverage higher than [MAX] will be kept
- cov — Regions for which the CSs corresponding to the child's Sginature CSs have a coverage of at least COV will be kept
- rci — A change higher than RCI in the RCI index will be considered significative
- rmin — A change in coverage will be considered significative only if the coverage for any CS is no lower than RMIN
- low — For the child, a CS with a coverage lower than LOW will be considered as a region of "almost no coverage"
- peaks — Regions with less than PEAK significative changes in coverage inside the VSR will be kept
- out — Output file. This will contain all TP de novo SNVs

OUTPUT

Same columns as the input file de de novo SNVs (Novo.out)

NOTES

This script should be run for every chromosome for every individual landscape.
The TP for all individuals for every chromosome will be merged with the script in the following section.

MERGE POST-PROCCESING

The SNVs for which its VSR is classified as a TP region for all three individuals are identified.

COMMAND LINE.

python Compare_TP.py -f HG1_TP.tab -m HG2_TP.tab -c HG3_TP.tab -o Novo_TP.tab –p Novo_TP.info

python Compare_TP.py --fatherTP HG1_TP.tab --motherTP HG2_TP.tab --childTP HG3_TP.tab --output Novo_TP.tab --outputALL Novo_TP.info

PARAMETERS

- fatherTP, f — File with the whole-genome TP *de novo* SNVs for the father
- motherTP, m — File with the whole-genome TP *de novo* SNVs for the mother
- childTP, c — File with the whole-genome TP *de novo* SNVs for the child
- output, o — Output file. TP in all three individuals
- outputALL, p — More information about TP in all three individuals

OUTPUT

Same columns as the input file de de novo SNVs (Novo.out)

NOTES

This script should be run only once

ACRONIMS

CS	Coin String
RG	Reference Genome
SP	Sequencing Project
VL	Variation Landscape
VSR	Variation Signature Region
SNV	Single Nucleotide Variant
PrevCS	PreviousCS
PostCS	PosteriorCS

REFERENCES

- Bowtie version supported 1.1.0 [http://bowtie-bio.sourceforge.net/index.shtml]
- Jellyfish version supported 1.1.6 [http://www.cbcb.umd.edu/software/jellyfish/]
- Python version supported 2.7.2